Sankhyā : The Indian Journal of Statistics
2017, Volume 79-B, Part 2, pp. 247-277

Minimum Risk Point Estimation of Gini Index

Shyamal K. De
National Institute of Science Education and Research, HBNI, Jatni, India
Bhargab Chattopadhyay
Indian Institute of Information Technology Vadodara, Gujarat, India
University of Texas at Dallas, Richardson, USA

Abstract

This paper develops a theory and methodology for estimation of Gini index such that both cost of sampling and estimation error are minimum. Methods in which sample size is fixed in advance, cannot minimize estimation error and sampling cost at the same time. In this article, a purely sequential procedure is proposed which provides an estimate of the sample size required to achieve a sufficiently smaller estimation error and lower sampling cost. Characteristics of the purely sequential procedure are examined and asymptotic optimality properties are proved without assuming any specific distribution of the data. Performance of our method is examined through extensive simulation study.

AMS (2000) subject classification. Primary: 62L12, 62G05; Secondary: 60G46, 60G40, 91B82.
Keywords and phrases. Asymptotic efficiency, Ratio regret, Reverse submartingale, Sequential point estimation, Simple random sampling, U-statistics

1 Introduction

Economic inequality exists in all societies or regions because of the existence of gap in income and wealth among individuals. In order to reduce the gap between the income levels of individuals, the government of a country devise several economic policies. Periodic evaluation of the effect of economic policies in reducing the income gap between rich and poor is important. There are several inequality indexes in the economic literature. Allison (1978) mentioned that among those indices, Gini inequality index is the most widely used measure. The Gini index satisfies four basic desirable criteria such as (i) anonymity, (ii) scale independence, (iii) population independence, and (iv) Pigou-Dalton transfer principle. Moreover, Gini index

has an easy interpretation and a relation to Lorenz curve. For more details on Gini index, please refer to Yitzhaki and Schechtman (2013, pp. 11–31).

The most celebrated Gini index, as given in Xu (2007), is

$$G_F(X) = \frac{\Delta}{2\mu}, \text{ where } \Delta = E\left|X_1 - X_2\right|, \mu = E(X) \tag{1.1}$$

and X_1 & X_2 are two i.i.d. copies of non-negative random variable X with an unknown distribution function $F(.)$. If there are n randomly selected individuals with incomes given by $X_1, X_2, \ldots, X_n$, then an estimator of Eq. 1.1 is given by

$$\widehat{G}_n = \frac{\widehat{\Delta}_n}{2\overline{X}_n}, \tag{1.2}$$

where $\overline{X}_n$ is the sample mean and $\widehat{\Delta}_n$ is the sample Gini's mean difference (GMD) defined as

$$\widehat{\Delta}_n = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} \left|X_{i_1} - X_{i_2}\right|. \tag{1.3}$$

For continuous evaluation of economic policies of a country, periodic computation of Gini index for the country is very important. One source from which Gini index of a region or a country can be calculated is census data which is typically collected every 10 years. But for estimating the Gini index in intermediate years, data from annual household survey conducted by government agencies can be used. For instance, National Sample Survey (NSS) in India, European Statistics on Income and Living Conditions in European Union and other agencies conduct household surveys annually or biennially in respective regions or countries. However, many countries, for example, Burundi, Chad, Mozambique (as per world bank website), can not afford or do not collect data from households annually or biennially on a relatively large scale.

If household survey data is not available, one has to draw a relatively small sample to estimate the Gini index for that region using appropriate sampling technique. The sampling technique should be chosen depending on the size and socio-economic diversity of the country. For a brief review of several sampling techniques, we refer to Cochran (1977). In order to compute Gini index for regions or smaller countries, with lesser social diversity, simple random sampling technique can be used to collect income or expenditure data. There exists literature on statistical inference for inequality indices which is computed from household income or expenditure by means of simple

random sampling from the population of interest (e.g., Beach and Davidson 1983; Davidson 2009; Davidson and Duclos 2000; Gastwirth 1972; Xu 2007). In this paper, we will use simple random sampling technique to collect income or expenditure data in order to estimate Gini index accurately.

It is well known that error in estimation decreases or in other words accuracy increases when the sample size increases. This in turn increases the overall cost of sampling. To minimize the cost of sampling, one has to reduce the sample size which in turn may lead to higher error in estimation. Thus, a method of estimation should be developed such that both the cost of sampling and the error in estimation are kept as low as possible. In other words, a procedure is required which can act as a trade-off between the estimation error and the sampling cost. To achieve this trade-off, fixed-sample methodologies cannot be used, i.e., the sample size should not be fixed in advance. This problem falls in the domain of sequential analysis where this is known as minimum risk point estimation problem. For more details on the literature of sequential analysis, we refer to Ghosh and Sen (1991), Ghosh et al. (1997), Mukhopadhyay and de Silva (2009), and others.

Unlike fixed-sample procedures, sequential procedures do not require sample size to be fixed in advance. Instead, in a sequential procedure, statistical analysis is continued as the observations are collected. Sampling is terminated according to a pre-defined criterion, also known as stopping rule. Sequential sampling allows the estimation process to finish early requiring small sample size. We are certainly not the first one to suggest sequential methods in econometrics. Several articles published in economics and econometrics journals pursued the idea of using sequential or multi-stage inference procedures. Examples include Aguirregabiria and Mira (2007), Arcidiacono and Jones (2003), Greene (1998), Kanninen (1993), etc. Recently, Chattopadhyay and De (2016) developed a sequential procedure for constructing a confidence interval for Gini index under the same i.i.d. set up, but the procedure cannot be used for solving the minimum risk point problem for Gini index.

Below, we provide a brief literature review of some relevant concepts and also our contribution to the literature of statistical inference and economics.

1.1. Literature Review and Our Contributions The estimator of Gini index in Eq. 1.2 involves sample mean and the estimator of Gini's mean difference which belong to a class of unbiased estimators known as U-statistics. Below, we briefly discuss the literature on U-statistics.

1.1.1. Literature on U-statistics. The theory and practice of U-statistics began with the pioneering papers of Hoeffding (1948, 1961). Hoeffding (1948) derived a general method for obtaining unbiased estimators for a

parameter θ associated with an unknown distribution function $F(.)$. Suppose that $X_1, \ldots, X_n$ are *independent and identically distributed* (i.i.d.) random variables from a population with a common distribution function $F(.)$ with an associated parameter $\theta \equiv \theta(F)$, $\theta \in \Theta \subseteq \mathcal{R}$. Then the U-statistic associated with θ is written as follows

$$U \equiv U_n^{(m)} = \binom{n}{m}^{-1} \sum_{(n,m)} g^{(m)}(X_{i_1}, ..., X_{i_m}),$$

where $\sum_{(n,m)}$ denotes the summation over all possible combinations of indices $(i_1, \ldots, i_m)$ such that $1 \leq i_1 < i_2 < \cdots < i_m \leq n$, and $m < n$. Here, $g^{(m)}(.)$ is a symmetric kernel of degree m such that $E_F\left[g^{(m)}(X_1, ..., X_m)\right] = \theta(F)$ for all $F(.)$. Thus, both GMD and the sample mean are U-Statistics with kernels of degree 2 and 1 respectively. Also, one may note that for a large class of probability distributions, U-statistics can be used to derive the minimum-variance unbiased estimator of the related parameter. Detailed literature on U-statistics can be found in standard textbooks such as Hollander and Wolfe (1999), Lee (1990), and others.

Apart from being unbiased estimators, U-statistics are also strongly consistent and reverse martingales with respect to some non-increasing filtration as proven in Lee (p. 119, 1990). We have exploited maximal inequalities for reverse martingales and also the strong consistency property of U-statistics in proving several lemmas in Section 7 and asymptotic efficiency and asymptotic risk-efficiency properties of our method in Section 3. For more literature on reverse martingales, we refer to classical textbooks on probability theory and stochastic processes such as Doob (1953), Loève (1963), and others.

As discussed before, we estimate the Gini index by a sequential method known as minimum risk point estimation (MRPE). Below, we briefly discuss the developments on minimum risk point estimation.

1.1.2. Literature on MRPE. Minimum risk point estimation was first introduced by Robbins (1959). He suggested a purely sequential procedure for estimating mean of a normal distribution. Ghosh and Mukhopadhyay (1979) generalized this idea to a distribution free scenario and developed a purely sequential procedure for minimum risk point estimation of a population mean. Later, Sen and Ghosh (1981) extended the sequential procedure of Ghosh and Mukhopadhyay (1979) to accommodate the minimum risk point estimation of any estimable parameter using U-statistics. For more

details on MRPE, we refer our readers to Ghosh et al. (1997), Mukhopadhyay and de Silva (2009), Sen (1981), and others.

In minimum risk point estimation problems, a cost function is defined which depends on sample size and error in estimation. In this paper, we will use mean square error (MSE) of Gini index as an error in estimation. We are interested in finding an estimate of unknown optimal sample size which minimizes the asymptotic cost function to estimate Gini index of the population.

1.1.3. Contributions of this Paper. Several fixed-sample methods are developed for estimation of Gini index assuming that the incomes from the sampled individuals are independent and identically distributed random variables. Examples of such methods can be found in Beach and Davidson (1983), Davidson (2009), Davidson and Duclos (2000), Gastwirth (1972), and Xu (2007). However, these methods cannot be used for minimum risk point estimation of an inequality index (for e.g., Chattopadhyay and De 2014). Motivated by this, we propose a sequential procedure that yields an asymptotic minimum risk point estimator of Gini index by minimizing the asymptotic risk function defined as a cost function plus a risk term for estimation error. Under some mild assumptions, we prove that the estimated final sample size for our procedure approaches the theoretically optimal sample size that minimizes the cost function. Moreover, we prove that the expected cost for estimating the Gini index using the estimated final sample size is asymptotically close to theoretically expected cost for estimating the Gini index, that is with theoretically optimal sample size. All theoretical results are validated by extensive simulation study.

The remainder of this paper is organized as follows. Section 2 develops a purely sequential procedure which minimizes both the estimation error and the overall sampling cost. Section 3 presents the theoretical properties enjoyed by the proposed sequential procedure. Performance of our method is assessed via simulation study in Section 4. The next section explores the possibility of satisfying stronger asymptotic optimality properties. In Section 6, we provide some concluding remarks. The Appendix contains some auxiliary lemmas and detailed proofs of all theoretical results.

2. Sequential Method of Estimation

2.1. Problem Formulation Incomes from n randomly selected individuals are collected. Suppose the incomes of n individuals $X_1, \ldots, X_n$ are i.i.d. copies of non-negative random variable X with an unknown distribution function $F(.)$ and the support of the distribution is (t, ∞) with $t > 0$. The estimator $\widehat{G}_n$ is a biased estimator of the population Gini index, G_F,

and $E(\widehat{G}_n - G_F)^2$ is the mean square error (MSE) of $\widehat{G}_n$. The asymptotic expression for MSE of $\widehat{G}_n$ is given by the following lemma.

Lemma 1. *$E(\widehat{G}_n - G_F)^2 = \frac{\xi^2}{n} + O\left(\frac{1}{n^{3/2}}\right)$, where*

$$\xi^2 = \frac{\sigma_1^2}{\mu^2} + \frac{\Delta^2\sigma^2}{4\mu^4} - \frac{\Delta}{\mu^3}(\tau - \mu\Delta), \tag{2.1}$$

$\sigma_1^2 = Var\left[E\left(|X_1 - X_2| \,\middle|\, X_1\right)\right]$, $\tau = E\left(X_1|X_1 - X_2|\right)$, *and* $\sigma^2 = Var(X)$, *provided $E(X_1^6)$ is finite.*

Proof. See Appendix.

Note that the parameter ξ^2 in Eq. 1 is positive since Hoeffding (1948) establishes that ξ^2 is the asymptotic variance of $\widehat{G}_n$.

We know that as the sample size becomes large, we receive more and more information about G_F and, therefore, expect the squared error loss $(G_F - \widehat{G}_n)^2$ due to estimation to be small. However, higher sample size leads to higher sampling cost. Therefore, it is desirable to consider a loss function that takes into account of both loss due to error in estimation and the sampling cost. Suppose c is the known cost of sampling each observation. Our goal is to find an estimation procedure which minimizes the MSE and also the sampling cost. We define a cost function depending on the MSE and the cost of sampling, also known as the risk function, as

$$R_n(G_F) = A\,E(\widehat{G}_n - G_F)^2 + cn. \tag{2.2}$$

Here, A is a known positive constant and is expressed in monetary terms which represents the weight assigned by the researchers or analysts regarding the probable cost per unit squared error loss due to estimation. Thus, the first term $AE(G_F - \widehat{G}_n)^2$ represents the loss in estimating G_F by $\widehat{G}_n$, and the second term cn represents the cost of sampling n observations. The risk function thus gives the expected cost of estimating G_F using the estimator $\widehat{G}_n$ based on incomes from n individuals. Using the asymptotic expression of MSE of $\widehat{G}_n$ expressed in Eq. 2.1, the fixed-sample size risk defined in Eq. 2.2 becomes

$$R_n(G_F) = A\frac{\xi^2}{n} + cn + O\left(\frac{1}{n^{3/2}}\right). \tag{2.3}$$

Thus, Eq. 2.3 gives the expected cost or the risk, to estimate the unknown value of the population Gini index using $\widehat{G}_n$ based on n observations. Our goal is to find the sample size for which the approximate expected cost

(ignoring the $O\left(\frac{1}{n^{3/2}}\right)$ term) defined in Eq. 2.3, i.e., $h(n) = A\frac{\xi^2}{n} + cn$, is minimized for all distributions that satisfy the conditions of Lemma 1.

Considering n as a non-negative continuous variable, the strictly convex function $h(n)$ can be minimized at

$$n = n_c \left(= \sqrt{\frac{A}{c}}\,\xi \right). \tag{2.4}$$

Thus, n_c is the required optimal sample size that should be collected using simple random sampling from the population in order to minimize the expected cost to estimate G_F. The the approximate expected cost of estimating the Gini index using a sample of size n_c or the asymptotic minimum risk is

$$R^*_{n_c}(G_F) = A\frac{\xi^2}{n_c} + cn_c = 2cn_c. \tag{2.5}$$

If the parameter ξ were known in advance, one could simply collect a sample of size n_c which is the minimum sample size to attain the asymptotic minimum risk. Since ξ is not known in practice, we need to collect samples in at least two stages where the first stage is to estimate ξ and n_c based on a pilot sample. The following section presents a consistent estimator for ξ^2 that will be used in our proposed sequential procedure.

2.2. Strong Consistent Estimator of ξ^2 Proceeding along the lines of Sproule (1969), let us define a U-statistic, for each $j = 1, 2, \ldots, n$,

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

where $T_j = \{(i_1, i_2) : 1 \le i_1 < i_2 \le n \text{ and } i_1, i_2 \ne j\}$. Moreover, define

$$W_{jn} = n\widehat{\Delta}_n - (n-2)\widehat{\Delta}^{(j)}_n, \text{ for } j = 1, \ldots, n, \text{ and } \overline{W}_n = n^{-1}\sum_{j=1}^{n} W_{jn}. \tag{2.6}$$

According to Sproule (1969), a strongly consistent estimator of $4\sigma_1^2$ is

$$s^2_{wn} = (n-1)^{-1}\sum_{i=1}^{n}(W_{jn} - \overline{W}_n)^2. \tag{2.7}$$

Using Xu (2007),

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

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

$$V_n^* = \frac{\widehat{\Delta}_n^2 S_n^2}{4\overline{X}_n^4} - \frac{\widehat{\Delta}_n}{\overline{X}_n^3}\widehat{\tau}_n + \frac{\widehat{\Delta}_n^2}{\overline{X}_n^2} + \frac{s_{wn}^2}{4\overline{X}_n^2}. \tag{2.8}$$

Note that $\widehat{\tau}_n, S_n^2, \widehat{\Delta}_n$ are U-statistics of degree 2 (e.g. Xu 2007; Chattopadhyay and Mukhopadhyay 2013) and the sample mean $\overline{X}_n$ is a U-statistic of degree 1. Using Sproule (1969) and Theorem 3.2.1 of Sen (p. 50, 1981), we observe that V_n^* is a strongly consistent estimator of ξ^2. Since V_n^* is a moment type estimator, it is very difficult in general, if not impossible, to prove that $P(V_n^* < 0) = 0$. Even though our simulation study suggests that V_n^* is strictly positive, we do not have any theoretical proof of that fact. Therefore, to avoid any ambiguity, we consider the positive part of V_n^*, i.e.,

$$V_n^2 = \max\{V_n^*, 0\}. \tag{2.9}$$

as an estimator of ξ^2. Since $f(x) = \max\{x, 0\}$ is a continuous function, V_n^2 will remain a strongly consistent estimator of ξ^2 by continuous mapping theorem.

2.3. The Sequential Procedure Recall that we need at least n_c samples to achieve the minimum risk in Eq. 6.5. Since the optimal sample size n_c is unknown, we must collect samples in at least two stages where the first stage is to estimate n_c based on a pilot sample and the second stage (or more stages) is to collect additional samples to have the total sample size as large as the estimated n_c. It can be seen from Dantzig (1940) that fixed-sample size procedures cannot minimize the risk in Eq. 2.3, not even asymptotically. Therefore, we propose a purely sequential procedure that yields minimum risk at least asymptotically.

The proposed purely sequential procedure deals with estimation of the minimum sample size in which, after pilot stage, collections of further observations and estimation of the unknown parameter ξ proceeds in stages according to a stopping rule. The stopping rule is formulated by using the optimal sample size defined in Eq. 2.4. For estimation of Gini index with minimum overall risk, we need a sample size which is at least as large as the optimal sample size. Thus, the stopping rule N, for every $c > 0$, can be defined as

$$N \equiv N(c) \text{ is the smallest integer } n(\geq m) \text{ such that } n \geq \sqrt{\frac{A}{c}}\, V_n. \tag{2.10}$$

Here, m is the initial or pilot sample size. Here, the estimator V_n may be very small which may cause our procedure to stop too early. To avoid this problem, we propose a slightly modified stopping rule N_c as

$$N_c \text{ is the smallest integer } n(\geq m) \text{ such that } n \geq \sqrt{\frac{A}{c}}\left(V_n + n^{-\gamma}\right), \quad (2.11)$$

where $\gamma \in (0, 2)$ is some chosen constant. The inclusion of the term $n^{-\gamma}$ ensures that we do not stop too early due to small value of V_n. Clearly, $V_n + n^{-\gamma}$ is still a strongly consistent estimator of ξ.

Below, we outline the purely sequential estimation procedure of the Gini index of the population according to the Stopping rule defined in Eq. 2.11.

Step 1: Collect a sample of incomes $X_1, \ldots, X_m$ from m randomly selected individuals. This sample is called pilot sample. Based on this pilot sample, estimate ξ^2 by V_m^2 defined in Eq. 2.9. If $m \geq \sqrt{\frac{A}{c}}\,(V_m + m^{-\gamma})$, then stop sampling and set the final sample size equal to m. Otherwise, go to the next step.

Step 2: Obtain one more income observation X_{m+1} from a randomly selected individual. Update the estimate of ξ^2 by computing V_{m+1}^2 based on $X_1, \ldots, X_{m+1}$. If $m + 1 \geq \sqrt{\frac{A}{c}}\,(V_{m+1} + (m+1)^{-\gamma})$ stop sampling and set the final sample size equal to $m + 1$. Otherwise, continue the sampling process by sampling one more individual.

The sampling process is continued until a stopping criteria is satisfied.

3 Main Results

For a given cost c per observation, the risk or the expected cost for estimating the Gini index G_F using an estimator based on the final sample size N_c is given by

$$R_{N_c}(G_F) = A\,E(\widehat{G}_{N_c} - G_F)^2 + cE(N_c). \quad (3.1)$$

Thus, the estimator $\widehat{G}_{N_c}$ is asymptotically minimum risk point estimator (AMRPE) if the ratio regret is asymptotically 1, i.e., if

$$\lim_{c \downarrow 0} R_{N_c}(G_F)/R_{n_c}(G_F) = 1. \quad (3.2)$$

In other words, estimator $\widehat{G}_{N_c}$ is AMRPE (c.f. Sen 1981) if the expected cost for estimating the Gini index G_F using an estimator based on the final

sample size N_c is asymptotically close to expected cost for estimating G_F using the optimal sample size, n_c. In decision theoretic framework, the ratio in Eq. 3.2 is known as ratio regret which is the ratio between the actual payoff and the minimum payoff due to some optimal strategy (Loomes and Sugden 1982).

Before discussing the asymptotic optimality properties of our method, we prove in the following lemma that if observations are collected using (2.11), sampling will stop at some finite time with probability one.

Lemma 2. *Under the assumption that $\xi < \infty$, for any $c > 0$, the stopping time N_c is finite, i.e., $P(N_c < \infty) = 1$.*

Proof. See Appendix.

This lemma is crucial for any sequential procedure because it assures that the practitioner will stop sampling eventually. Below, we provide the main theorem related to the asymptotic optimality properties of our procedure.

Theorem 1. *If $E(X_1^6)$ is finite, the stopping rule (2.11) yields:*

(i) $N_c/n_c \to 1$ *almost surely as* $c \downarrow 0$.

(ii) $E(N_c/n_c) \to 1$ *as* $c \downarrow 0$. *[Asymptotic First-order Efficiency]*

(iii) For $\gamma \in (0, 2)$, $R_{N_c}(G_F)/R_{n_c}(G_F) \to 1$ *as* $c \downarrow 0$. *[Asymptotic First-order Risk Efficiency]*

Proof. See Appendix.

Remark 1. *All the asymptotic properties of our sequential procedure as in Theorem 1 will hold even if we replace the assumption of truncated support (t, ∞) for $F(.)$ by the assumption of positive support $(0, \infty)$. In that case, we will need extra assumptions on $F(.)$ that higher order (twelfth) positive moment and (twentieth) negative moment are finite. The proofs of the corresponding lemmas and the theorem are along the same lines as in the Appendix.*

Remark 2. *The parts (i) and (ii) of Theorem 1 imply that the final sample size of our procedure is asymptotically same as the minimum sample size required to minimize the asymptotic risk defined in Eq. 2.3. The part (iii) proves that the risk attained by our procedure is asymptotically same as the minimum risk. Therefore, the Gini index estimator $\widehat{G}_{N_c}$ is indeed AMRPE. The optimality properties in part (ii) and (iii) are well known in the sequential literature as asymptotic first-order efficiency and asymptotic first-order risk efficiency respectively (see Mukhopadhyay and de Silva 2009).*

Table 1: Estimated average final sample sizes and overall risks with $A = \$500000, \gamma = 1$ and $c = \$0.5$

Distribution	$\widehat{G}_{N_c}$ G_F	$\overline{N_c}$ $\mathbf{s}(\overline{N_c})$	n_c	$\overline{N_c}/n_c$	$\overline{r}_{N_c}$ $\mathrm{s}(\overline{r}_{N_c})$	$\frac{\overline{r}_{N_c}}{R^*{}_{n_c}}$
Exponential	0.4996	292.6794	289	1.0127	288.5219	0.9983
(rate=5, t=0.001)	0.4997	0.2891			0.2917	
Gamma (shape=2.649,	0.3301	220.6912	217	1.0170	215.5002	0.9931
rate=0.84, t=0.001)	0.3308	0.2265			0.2309	
Log-normal (mean=2.185,	0.3082	228.5236	231	0.9893	223.3612	0.9670
sd=0.562, t=0.001)	0.3089	0.4423			0.4495	
Pareto	0.0891	128.699	122	1.0549	120.5224	0.9879
(scale=20000, shape=6.1)	0.0893	0.1284			0.1349	

4 Simulation Study

In this section, we evaluate the performance of our estimation method for moderate sample size (i.e., c is small but not too small) via simulation study. In this study, we consider the stopping rule given in Eq. 2.11. To implement the sequential procedure, we fix the pilot sample size $m = 10$, $A = \$500000$ and c at two different values: $c = \$0.5$ and $c = \$5$. The results in Tables 1 and 2 are based on random samples from 4 different income distributions: exponential with rate 5 truncated at $t = 0.001$ ($G_F = 0.4997, \xi^2 = 0.0832$), gamma with shape 2.649 and rate 0.84 truncated at $t = 0.001$ ($G_F = 0.3308, \xi^2 = 0.0468$), log-normal with mean (on the

Table 2: Estimated average final sample sizes and overall risks with $A = \$500000, \gamma = 1$ and $c = \$5$

Distribution	$\widehat{G}_{N_c}$ G_F	$\overline{N_c}$ $\mathbf{s}(\overline{N_c})$	n_c	$\overline{N_c}/n_c$	$\overline{r}_{N_c}$ $\mathrm{s}(\overline{r}_{N_c})$	$\frac{\overline{r}_{N_c}}{R^*{}_{n_c}}$
Exponential	0.4995	96.2140	92	1.0458	922.3382	1.0025
(rate=5, t=0.001)	0.4997	0.1363			1.3989	
Gamma (shape=2.649,	0.3298	72.7860	69	1.0549	678.7001	0.9836
rate=0.84, t=0.001)	0.3308	0.1027			1.0768	
Log-normal (mean=2.185,	0.3060	73.3188	73	1.0044	682.8903	0.9355
sd=0.562, t=0.001)	0.3089	0.1925			2.0117	
Pareto	0.0885	44.3318	39	1.1367	371.2755	0.9520
(scale=20000, shape=6.1)	0.0893	0.0371			0.4111	

log-scale) 2.185 and standard deviation (on the log scale) 0.562 truncated at $t = 0.001$ ($G_F = 0.3089, \xi^2 = 0.0532$), and Pareto with scale 20000 and shape 6.1 ($G_F = 0.0893, \xi^2 = 0.0148$). So, all randomly generated values from each of the above distributions are greater than 0.001. The number of replications used in Monte Carlo simulations is 5000.

Tables 1 and 2 present the average final sample size $\overline{N_c}$ (estimates $E(N_c)$) from 5000 replications, and the average risk $\overline{r}_{N_c}$ (estimates $R_{N_c}(G_F)$) obtained from the sample of size N_c. Moreover, $s(\overline{N_c})$ and $s(\overline{r}_{N_c})$ represent the standard errors of $\overline{N_c}$ and $\overline{r}_{N_c}$ respectively. Both tables show that the average sample size $\overline{N_c}$ is almost the same as the optimal sample size n_c. Therefore, on average, our procedure requires only the minimum sample size n_c. The last columns of both tables illustrate that, on average, the overall cost for estimating the Gini index G_F using an estimator based on the estimated final sample size N_c is asymptotically close to expected cost for estimating G_F using the optimal sample size n_c. In other words, the ratio regret is nearly 1. This implies that the risk incurred by our method is almost the same as the minimum possible risk $R^*_{n_c}$ defined in Eq. 6.5. Thus, we find that the proposed sequential procedure performs well for the income distributions considered above.

Table 3 compares the true values of the parameters, ξ^2, for the four income distributions considered above with their estimated values and their standard errors based on the final sample sizes N_c from Table 1. Here, $s\left(V^2_{N_c}\right)$ represents the standard error of the estimator, $V^2_{N_c}$, of ξ^2. Table 3 shows that the average value of the estimator are close to the true value of the parameter and, therefore, it indicates that $V^2_{N_c} \to \xi^2$ as $c \downarrow 0$.

5 Application on Real Data

In this section, we provide a real life example based on a recent case study by Das and Rout (2015) on the monthly income of earning members of a tribal village in Odisha, India. The case study recorded monthly income of each of the 295 earning members of the village with the distant goal of

Table 3: Estimated values of ξ^2 and their standard errors

	exponential ($\xi^2 = 0.0833$)	gamma ($\xi^2 = 0.0468$)	log-normal ($\xi^2 = 0.0526$)	Pareto ($\xi^2 = 0.0148$)
$\overline{V^2_{N_c}}$	0.0837	0.0467	0.0509	0.0145
$\mathbf{s}\left(\overline{V^2_{N_c}}\right)$	0.0002	0.0001	0.00009	0.00003

universalizing elementary education. The cost of collecting income data per person, denoted earlier by c, is estimated to be ₹ 100. For this data, our objective is to measure the income inequality of the village. Figure 1 shows the Lorenz curve based on incomes of 295 earning members of Bhagabatipur. In order to test the applicability and usefulness of our proposed approach, we aim to collect an optimal sample from the given complete income data of the village to estimate population Gini index and compare the estimate with the population value of Gini index. The Gini index for the population of Bhagabatipur village came out to be 0.4586 and the optimal sample size came out as 176. For our study, we assume that the value of A, representing the loss due to wrong estimation, to be ₹ 65,00,000. The minimized risk of estimating the population Gini index with sample Gini index using a sample of size 176 turns out to be ₹ 35,200.00. Applying the sequential procedure, we first select randomly a pilot sample of size $m = 10$ and then in subsequent stages, added one observation in every stage until the stopping rule (2.10) is satisfied.

The optimal sample size n_c is estimated by the stopping variable N_c as 173. Using these 173 observations, the estimated value of Gini index $\widehat{G}_{N_c}$ turns out to be 0.44677 and the total risk $R_{N_c}(G_F)$ of estimating the population Gini index with sample Gini index using a sample of size 173 turns out to be ₹ 34,116.48. Thus, if our estimation technique were used, we could have saved taking at least $295 - 173 = 122$ observations, which would lead to saving cost of study.

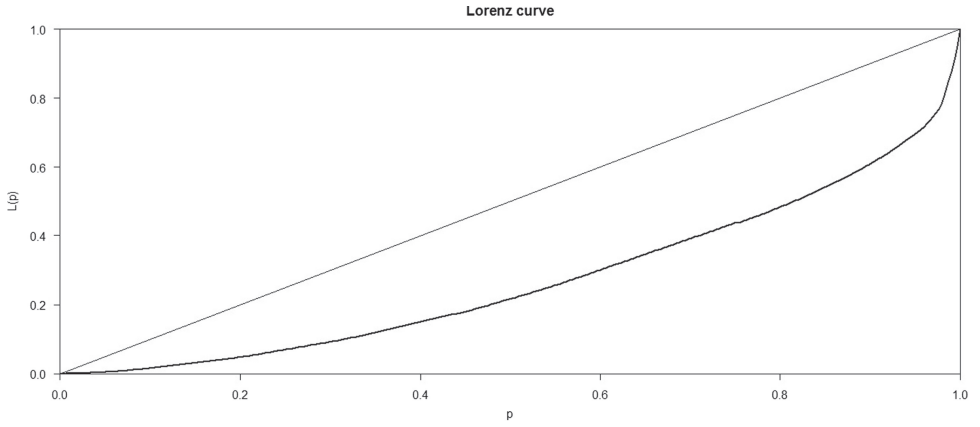

Figure 1: Lorenz Curve Relating to the Income of people residing at Bhagabatipur, Orissa

Table 4: Comparison of mean final sample sizes and over/under-sampling rates for using W_n^2 versus V_n^2 with $A = \$500000, \gamma = 1$ and $c = \$0.5$

Distribution	Estimator of ξ^2 Used	Estimate of ξ^2(SE$\times 10^4$)	Mean Final Sample Size	Over-/Under-Sampling Rate
Exp ($G_F = 0.4997$)	V_n^2	0.08285(2.72)	287.8596	0.39%
$\xi^2 = 0.0832, n_c = 289$	W_n^2	0.08288(2.73)	283.5776	1.88%
Gam($G_F = 0.3308$)	V_n^2	0.04650(1.03)	215.6922	0.60%
$\xi^2 = 0.0468, n_c = 217$	W_n^2	0.04652(1.04)	213.482	1.62%
Lnorm($G_F = 0.3089$)	V_n^2	0.0505 (2.18)	223.2190	3.37%
$\xi^2 = 0.0532, n_c = 231$	W_n^2	0.0504(2.19)	219.2792	5.07%
Par($G_F = 0.0893$)	V_n^2	0.0121 (1.13)	115.0246	5.72%
$\xi^2 = 0.0148, n_c = 122$	W_n^2	0.0112 (1.17)	96.8366	20.62%

6 Some Discussions and Extensions

6.1. Comparison with Sequential Procedure Based on Davidson's Estimator In our proposed sequential procedure, we used V_n^2, given in Eq. 2.9, as a strong consistent estimator of ξ^2. As an alternative, one may use another estimator of ξ^2 proposed in Davidson (2009). Davidson's estimator, say W_n^2, is simpler to compute. However, the strong consistency property of W_n^2 is not known. The strong consistency property is essential to prove the asymptotic optimality properties of Theorem 1. Moreover, we found that V_n^2 can also be computed easily with available modern computing facilities.

Nevertheless, we compare the performance of the proposed sequential procedure using V_n^2 with that of the Davidson's estimator W_n^2.

In Tables 4 and 5, we consider random samples from four different income distributions: Exp represents exponential distribution with rate 5 truncated at $t = 0.001$, Gam represents gamma distribution with shape 2.649 and rate 0.84 truncated at $t = 0.001$, Lnorm represents log-normal distribution with

Table 5: Comparison of mean final sample sizes and over/under-sampling rates for using W_n^2 versus V_n^2 with $A = \$500000, \gamma = 1$ and $c = \$5$

Distribution	Estimator of ξ^2 Used	Estimate of ξ^2(SE$\times 10^4$)	Mean Final Sample Size	Over-/Under-Sampling Rate
Exp ($G_F = 0.4997$)	V_n^2	0.0846 (1.72)	95.5168	3.83%
$\xi^2 = 0.0832, n_c = 92$	W_n^2	0.0845 (1.74)	88.2072	4.12%
Gam($G_F = 0.3308$)	V_n^2	0.0461(1.66)	68.0666	1.35%
$\xi^2 = 0.0468, n_c = 69$	W_n^2	0.0462(1.68)	66.0862	4.22%
Lnorm($G_F = 0.3089$)	V_n^2	0.0468(3.08)	67.6768	7.29%
$\xi^2 = 0.0532, n_c = 73$	W_n^2	0.0466(3.13)	64.1432	12.13%
Par($G_F = 0.0893$)	V_n^2	0.01035 (1.48)	38.7720	5.85%
$\xi^2 = 0.0148, n_c = 39$	W_n^2	0.00891(1.36)	26.4948	32.06%

mean (on the log-scale) 2.185 and standard deviation (on the log scale) 0.562 truncated at $t = 0.001$, and Par represents Pareto distribution with scale 20000 and shape 6.1. We fixed a seed value for both sequential procedures so that both procedures can be applied on the same set of observations up to the minimum final sample size value, the minimum being the smallest final sample size given by the two procedures. The number of replications used in the Monte Carlo simulations is 5000. The over- or under-sampling rate is computed as:

$$\text{Over-/under-sampling rate} = \frac{|\text{mean final sample size} - n_c|}{n_c} \times 100\%. \quad (6.1)$$

Tables 4 and 5 indicate that the over-sampling or under-sampling rate of the sequential procedure based on W_n^2 is higher than the sequential procedure using V_n^2. From Tables 4 and 5, we observe that the sequential procedure based on V_n^2 provides a final sample size which is closer to the optimal final sample size n_c.

Using the income data of Bhagabatipur village as in Section 5, we compare the under-sampling or over-sampling rate of our proposed sequential procedure with the sequential procedure based on Davidson's estimator. We found that for our procedure, the optimal sample size n_c (=176) is estimated as 173, whereas the estimated n_c turns out to be 166 using the sequential procedure based on Davidson's estimator. Thus, the estimator V_n^2 gave final sample size estimate closer to the true optimal sample size than that of W_n^2.

6.2. Two-stage Procedure As opposed to fixed-sample procedures, the final sample size is not fixed in advance in two-stage procedure. For details about the general two-stage confidence interval estimation procedures, we refer interested readers to Mukhopadhyay and De Silva (2009), Mukhopadhyay and Chattopadhyay (2012), Chattopadhyay and Mukhopadhyay (2013).

In a two-stage procedure, the sample size calculation is done in two stages. In the first stage, a pilot sample of size m is observed which can be used to estimate the unknown parameter (ξ^2) in the expression of MSE of the estimator of Gini index using it's strongly consistent estimator V_m^2. Then, using the value of V_m^2, we find an estimate of the sample size n_c by

$$N_{2c} = max\left\{m, \sqrt{\frac{A}{c}}\left(V_m + m^{-\gamma}\right)\right\} \quad (6.2)$$

If $N_{2c} = m$, no more observations are needed, but if $N_{2c} > m$, record $N_{2c} - m$ additional observations. Finally, based on the combined data $X_1, ..., X_{N_{2c}}$, we estimate the Gini index, G_F by $\widehat{G}_{N_{2c}}$.

6.3. Multi-parameter Estimation The theory presented in this paper for the single parameter case can be extended to a multi-parameter setup. Suppose we would like to estimate a vector of parameters $\boldsymbol{\theta}_F = (\theta_1, \theta_2, \ldots, \theta_p)$ for $p \geq 2$. This situation arise when we want to estimate the population mean, standard deviation, Gini index, and other characteristics related to income variable. Let the vector of estimators be defined as $\mathbf{T}_n = (T_{1n}, \ldots, T_{pn})$ based on $X_1, X_2, \ldots, X_n$ which are incomes of n individuals.

Our methodology can be easily extended to the multivariate set up in the spirit of Ghosh and Sen (pp. 346-352, 1991). Suppose under some regularity conditions, the asymptotic mean square error, $E[(\mathbf{T}_n - \boldsymbol{\theta}(F))(\mathbf{T}_n - \boldsymbol{\theta}(F))'] \approx \boldsymbol{\xi}/n$, where, $\boldsymbol{\xi}$ is a matrix.

Then, a multivariate extension of the cost function or the risk function, defined in Eq. 2.3, is given by

$$\begin{aligned} R(\mathbf{T}_n, \boldsymbol{\theta}(F)) &= AE[(\mathbf{T}_n - \boldsymbol{\theta}_F)'(\mathbf{T}_n - \boldsymbol{\theta}_F)] + cn \\ &= ATrace(E[(\mathbf{T}_n - \boldsymbol{\theta}_F)(\mathbf{T}_n - \boldsymbol{\theta}_F)']) + cn \\ &\approx ATrace(\boldsymbol{\xi})/n + cn \end{aligned} \tag{6.3}$$

The approximate multivariate risk can be minimized at

$$n = n_0 \left(= \sqrt{\frac{A}{c}} \, (Trace(\boldsymbol{\xi}))^{\frac{1}{2}} \right). \tag{6.4}$$

Thus, n_0 is the required optimal sample size that should be collected using simple random sampling from the population in order to minimize the expected cost to estimate $\boldsymbol{\theta}_F$ and the approximate expected cost of estimating the parameter vector using a sample of size n_0 or the asymptotic minimum risk is

$$R_{n_0}(\mathbf{T}_n, \boldsymbol{\theta}(F)) = \frac{A}{n_0} Trace(\boldsymbol{\xi}) + cn_0 = 2cn_0. \tag{6.5}$$

If the parameter $\boldsymbol{\xi}$ were known in advance, one could simply collect a sample of size n_0 which is the minimum sample size to attain the asymptotic minimum risk. Since $\boldsymbol{\xi}$ is not known in practice, it must be estimated by collecting a pilot sample. In order to estimate $\boldsymbol{\xi}$, we should use a strongly consistent estimator, $\mathbf{V}_n$ which can be obtained using the jackknife method

in the same spirit as in Eq. 2.7. The strong consistency result of the jackknife estimator follows from Sen (1988) which includes a general class of differentiable statistical functions. Thus, using the jackknife estimator we define a stopping rule similar to Eq. 2.11

$$N_0 \text{ is the smallest integer } n(\geq m) \ni n \geq \sqrt{\frac{A}{c}} \left(Trace(\boldsymbol{V}_n) + n^{-2\gamma}\right)^{\frac{1}{2}}. \tag{6.6}$$

A theorem similar to Theorem 1 is expected to hold under appropriate moment conditions. However, we do not explore this possibility in this paper.

7 Concluding Remarks

The Gini index or Gini concentration is a very popular measure of inequality. Apart from economics, there are other fields where researchers report Gini index. For instance, in social sciences and economics, the Gini index is used to measure inequality in education (see Thomas et al. 2001). In ecology, the Gini index is used as a measure of biodiversity (for e.g., see Wittebolle et al. 2009). Asada (2005) used Gini index as a measure of the inequality of health related quality of life in a population. Shi and Sethu (2003) used Gini index to evaluate the fairness achieved by internet routers in scheduling packet transmissions from different flows of traffic. Possible application of Gini index in so many fields such as sociology, health science, ecology, engineering including economics motivated us to develop the theory for the minimum risk point estimation problem of Gini index.

It is well known that error in estimation of Gini index decreases when the sample size increases. This inflates the overall cost of sampling. In order to compute Gini index for a region or a smaller country with lesser diversity at a specific point of time, we develop a procedure which computes the final sample size needed to minimize both the error of estimation as well as the cost of sampling via simple random sampling technique.

Without assuming any specific distribution of the data, we showed that the average final sample size using our procedure approaches the unknown optimal sample size that minimizes the cost function. Moreover, we proved that the expected cost for estimating the population Gini index using the estimated final sample size is asymptotically close to the expected cost for estimating the population Gini index using the unknown optimal sample size. Thus, based on the results, we conclude that the proposed sequential estimation strategy is quite efficient in reducing both sampling cost and estimation error.

Acknowledgements. We are grateful to the associate editor and the two anonymous referees for their valuable feedback that helped us improve this manuscript. We also thank Prof. Amarendra Das for providing the income data, which helped us show an application of the proposed method of this article.

Appendix: Auxiliary Results and Proofs

Proof of Lemma 2. Note that V_n is strongly consistent estimator of ξ. Therefore, for any fixed $c > 0$,

$$\begin{aligned} P(N_c > \infty) &= \lim_{n\to\infty} P(N_c > n) \\ &\leq \lim_{n\to\infty} P\left(n < \sqrt{A/c}\,(V_n + n^{-\gamma})\right) = 0. \end{aligned}$$

The last equality is obtained since $V_n \to \xi$ almost surely as $n \to \infty$. This completes the proof.

Lemmas to Prove the Main Result. This section is dedicated to prove some lemmas that are essential to establish the main Theorem 1. First, we introduce few notations. Note from Eq. 2.11 that $N_c \geq \sqrt{\frac{A}{c}}\, N_c^{-\gamma}$, i.e., $N_c \geq \left(\frac{A}{c}\right)^{\frac{1}{2(1+\gamma)}}$ with probability 1. For fixed ϵ, $\gamma > 0$, define

$$n_{1c} = \left(\frac{A}{c}\right)^{\frac{1}{2(1+\gamma)}}, \quad n_{2c} = n_c(1-\epsilon), \text{ and } n_{3c} = n_c(1+\epsilon), \text{ where } n_c = \sqrt{\frac{A}{c}}\,\xi. \tag{A.1}$$

Suppose $\boldsymbol{X}_{(n)}$ 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 Lee (1990), $\left\{\overline{X}_n, \mathcal{F}_n\right\}$, $\left\{S_n^2, \mathcal{F}_n\right\}$, $\left\{\widehat{\tau}_n, \mathcal{F}_n\right\}$, $\left\{\widehat{\Delta}_n, \mathcal{F}_n\right\}$, and their convex functions are all reverse submartingales. Using reverse submartingale properties of U-statistics, we prove the following lemmas.

Lemma 3. *If $X_1, \ldots, X_n$ are i.i.d. random variables such that $E(X_1^{2r}) < \infty$ for some positive integer r, then for any $k > 0$,*

$$P\left(\max_{n_{1c}\leq n\leq n_{2c}} \left|\widehat{\Delta}_n^2 - \Delta^2\right| \geq k\right) \leq O(n_{1c}^{-r/2}) \quad \text{as } c \downarrow 0.$$

Proof. Note that

$$\begin{aligned} \left|\widehat{\Delta}_n^2 - \Delta^2\right| &= \left|\left(\widehat{\Delta}_n^2 - \Delta^2\right) I(\widehat{\Delta}_n \geq \Delta) + \left(\widehat{\Delta}_n^2 - \Delta^2\right) I(\widehat{\Delta}_n < \Delta)\right| \\ &\leq \left(\widehat{\Delta}_n^2 - \Delta^2\right)^+ + 2\Delta\left|\widehat{\Delta}_n - \Delta\right|. \end{aligned} \tag{A.2}$$

Here, the notation x^+ is used to represent $\max\{x, 0\}$. Therefore, for $k > 0$,

$$P\left(\max_{n_{1c}\le n\le n_{2c}}\left|\widehat{\Delta}_n^2-\Delta^2\right|\ge k\right)$$
$$\le P\left(\max_{n_{1c}\le n\le n_{2c}}\left(\widehat{\Delta}_n^2-\Delta^2\right)^+\ge\frac{k}{2}\right)+P\left(\max_{n_{1c}\le n\le n_{2c}}\left|\widehat{\Delta}_n-\Delta\right|\ge\frac{k}{4\Delta}\right).$$

Since $\left(\widehat{\Delta}_n^2-\Delta^2\right)^+$ and $\left|\widehat{\Delta}_n-\Delta\right|$ are reverse submartingales, using maximal inequality for reverse submartingales (Ghosh et al. 1997, p. 25), we write

$$\begin{aligned}
&P\left(\max_{n_{1c}\le n\le n_{2c}}\left|\widehat{\Delta}_n^2-\Delta^2\right|\ge k\right) \qquad \text{(A.3)}\\
&\le\left(\frac{2}{k}\right)^r E\left[\left(\left(\widehat{\Delta}_{n_{1c}}^2-\Delta^2\right)^+\right)^r\right]+\left(\frac{4\Delta}{k}\right)^{2r}E\left[\left|\widehat{\Delta}_{n_{1c}}-\Delta\right|^{2r}\right]\\
&\le\left(\frac{2}{k}\right)^r\left\{E\left[\left(\widehat{\Delta}_{n_{1c}}-\Delta\right)^{2r}\right]E\left[\left(\widehat{\Delta}_{n_{1c}}+\Delta\right)^{2r}\right]\right\}^{\frac{1}{2}}+O\left(n_{1c}^{-r}\right)\\
&=O\left(n_{1c}^{-r/2}\right)+O\left(n_{1c}^{-r}\right)=O\left(n_{1c}^{-r/2}\right).
\end{aligned}$$

The last two inequalities are obtained by Lemma 2.2 of Sen and Ghosh (1981) and Cauchy–Schwarz inequality. The moment conditions in this lemma are needed to ensure that all expectations exist in the inequalities.

Lemma 4. *Suppose $X_1, \ldots, X_n$ are i.i.d. random variables from the distribution $F(.)$ such that $E(X_1^{2p})$ is finite for some $p > 1$. Then, for any positive integer r and $k > 0$,*

$$P\left(\max_{n_{1c}\le n\le n_{2c}}\left|\frac{1}{\overline{X}_n^r}-\frac{1}{\mu^r}\right|\ge k\right)\le O(n_{1c}^{-p})\quad as\ \ c\downarrow 0.$$

PROOF. By Taylor expansion of $\overline{X}_n^{-r}=\frac{1}{\mu^r}\left(1+(\overline{X}_n-\mu)/\mu\right)^{-r}$, we have

$$\begin{aligned}
&\left|\left(\frac{1}{\overline{X}_n^r}-\frac{1}{\mu^r}\right)I\left(\frac{1}{\overline{X}_n}<\frac{1}{\mu}\right)\right|\\
&=\frac{1}{\mu^r}\left|\left\{-\frac{r}{\mu}(\overline{X}_n-\mu)+\frac{r(r+1)}{2\mu^2}\frac{(\overline{X}_n-\mu)^2}{z^{r+2}}\right\}I\left(\frac{1}{\overline{X}_n}<\frac{1}{\mu}\right)\right|,
\end{aligned}$$

where $z \in [1, \overline{X}_n/\mu]$. Since $z^{-(r+2)} I(\overline{X}_n^{-1} < \mu^{-1}) \leq 1$, proceeding along the lines of Eq. A.2

$$\begin{aligned} \left| \frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right| &= \left| \left(\frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right) I\left(\frac{1}{\overline{\overline{X}}_n} \geq \frac{1}{\mu} \right) + \left(\frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right) I\left(\frac{1}{\overline{\overline{X}}_n} < \frac{1}{\mu} \right) \right| \\ &\leq \left(\frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right)^+ + \frac{r}{\mu^{r+1}} \left| \overline{X}_n - \mu \right| + \frac{r(r+1)}{2\mu^{r+2}} (\overline{X}_n - \mu)^2. \quad \text{(A.4)} \end{aligned}$$

Let $U_{1n} = \left(\frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right)^+$, $U_{2n} = \frac{r}{\mu^{r+1}} \left| \overline{X}_n - \mu \right|$, and $U_{3n} = \frac{r(r+1)}{2\mu^{r+2}} (\overline{X}_n - \mu)^2$. Using (A.4), we can write

$$\begin{aligned} P\left(\max_{n_{1c} \leq n \leq n_{2c}} \left| \frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right| \geq k \right) &\leq P\left(\max_{n_{1c} \leq n \leq n_{2c}} U_{1n} \geq \frac{k}{3} \right) + P\left(\max_{n_{1c} \leq n \leq n_{2c}} U_{2n} \geq \frac{k}{3} \right) \\ &\quad + P\left(\max_{n_{1c} \leq n \leq n_{2c}} U_{3n} \geq \frac{k}{3} \right). \quad \text{(A.5)} \end{aligned}$$

Since $\left(\frac{1}{\overline{\overline{X}}_n^r} - \frac{1}{\mu^r} \right)$ is a reverse submartingale and $f(x) = x^+$ is a non-decreasing convex function of x, U_{1n} is a reverse submartingale. Moreover, U_{1n}^p is a reverse submartingale for $p > 1$. Therefore, using maximal inequality for reverse submartingales

$$\begin{aligned} & P\left(\max_{n_{1c} \leq n \leq n_{2c}} U_{1n} \geq \frac{k}{3} \right) \\ \leq & \left(\frac{3}{k} \right)^{2p} E\left[\left(\frac{1}{\overline{\overline{X}}_{n_{1c}}^r} - \frac{1}{\mu^r} \right)^+ \right]^{2p} \\ \leq & \left(\frac{3}{k} \right)^{2p} E\left[\left(\frac{1}{\overline{\overline{X}}_{n_{1c}}} - \frac{1}{\mu} \right) \left(\frac{1}{\overline{\overline{X}}_{n_{1c}}^{r-1}} + \frac{1}{\mu \overline{\overline{X}}_{n_{1c}}^{r-2}} + \ldots + \frac{1}{\mu^{r-1}} \right) I\left(\overline{X}_{n_{1c}} < \mu \right) \right]^{2p} \\ \leq & \left(\frac{3}{k} \right)^{2p} r^{2p} E\left[\left(\frac{1}{\overline{\overline{X}}_{n_{1c}}} - \frac{1}{\mu} \right)^{2p} \overline{X}_{n_{1c}}^{-2p(r-1)} \right] \\ \leq & \left(\frac{3}{k} \right)^{2p} r^{2p} E\left[\left(\frac{1}{\overline{\overline{X}}_{n_{1c}}} - \frac{1}{\mu} \right)^{2p} t^{-2p(r-1)} \right] \\ \leq & \left(\frac{3r}{k} \right)^{2p} \left\{ E\left[\left(\overline{X}_{n_{1c}} - \mu \right)^{2p} \right] \left(\frac{1}{t^r \mu} \right)^{2p} \right\} \\ = & O(n_{1c}^{-p}). \quad \text{(A.6)} \end{aligned}$$

The last two inequalities are obtained by Lemma 2.2 of Sen and Ghosh (1981). Since $|\overline{X}_n - \mu|$ and $(\overline{X}_n - \mu)^2$ are reverse submartingales, we can write

$$P\left(\max_{n_{1c}\le n\le n_{2c}} U_{2n} \ge \frac{k}{3}\right) \le \left(\frac{3r}{k\mu^{r+1}}\right)^{2p} E\left(\overline{X}_{n_{1c}} - \mu\right)^{2p} = O(n_{1c}^{-p}), \quad \text{(A.7)}$$

$$P\left(\max_{n_{1c}\le n\le n_{2c}} U_{3n} \ge \frac{k}{3}\right) \le \left(\frac{3r(r+1)}{2k\mu^{r+2}}\right)^{2p} E\left(\overline{X}_{n_{1c}} - \mu\right)^{2p} = O(n_{1c}^{-p}). \quad \text{(A.8)}$$

Apply (A.6), (A.7), and (A.8) in (A.5) to complete the proof.

Lemma 5. *Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables such that $E(X_1^{4r})$ is finite for some $r \ge 1$. For any $\epsilon \in (0,1)$ and $\gamma > 0$,*

(i) $P(N_c \le n_c(1-\epsilon)) = O\left(n_{1c}^{-r}\right) = O\left(c^{\frac{r}{2(1+\gamma)}}\right)$ *as* $c \downarrow 0$,

(ii) $P(N_c > n_c(1+\epsilon)) = O\left(n_{1c}^{-r}\right) = O\left(c^{\frac{r}{2(1+\gamma)}}\right)$ *as* $c \downarrow 0$.

PROOF. Using the definition of stopping rule N_c in Eqs. 2.11 and A.1, we have

$$\begin{aligned}
P(N_c \le n_{2c}) &\le P\left(n > \sqrt{\frac{A}{c}} V_n \text{ for some } n \in [n_{1c}, n_{2c}]\right) \\
&\le P\left(V_n^2 \le \left(\frac{c}{A}\right) n_{2c}^2 \text{ for some } n \in [n_{1c}, n_{2c}]\right) \\
&\le P\left(V_n^* \le \xi^2(1-\epsilon)^2 \text{ for some } n \in [n_{1c}, n_{2c}]\right) \\
&\le P\left(|V_n^* - \xi^2| \ge \xi^2\epsilon(2-\epsilon) \text{ for some } n \in [n_{1c}, n_{2c}]\right) \\
&\le P\left(\max_{n_{1c}\le n\le n_{2c}} \{|V_{1n}| + |V_{2n}| + |V_{3n}| + |V_{4n}|\} \ge \xi^2\epsilon(2-\epsilon)\right),
\end{aligned} \quad \text{(A.9)}$$

where $V_{1n} = \left(\frac{\widehat{\Delta}_n^2}{4\overline{X}_n^4} S_n^2 - \frac{\Delta^2}{4\mu^4}\sigma^2\right)$, $V_{2n} = \left(\frac{\widehat{\Delta}_n}{\overline{X}_n^3}\widehat{\tau}_n - \frac{\Delta}{\mu^3}\tau\right)$, $V_{3n} = \left(\frac{\widehat{\Delta}_n^2}{\overline{X}_n^2} - \frac{\Delta^2}{\mu^2}\right)$, and $V_{4n} = \left(\frac{s_{wn}^2}{4\overline{X}_n^2} - \frac{\sigma_1^2}{\mu^2}\right)$. Let $k = \xi^2\epsilon(2-\epsilon)$. Then, Eq. A.15 can be written as $P(N_c \le n_{2c}) \le P_1 + P_2 + P_3 + P_4$, where

$$P_i = P\left(\max_{n_{1c}\le n\le n_{2c}} |V_{in}| \ge \frac{k}{4}\right), \quad \text{for } i = 1, 2, 3, 4.$$

First, let us find an upper bound of P_1. Let $T_{1n} = \left(\widehat{\Delta}_n^2 - \Delta^2\right)$, $T_{2n} = \left(S_n^2 - \sigma^2\right)$, and $T_{3n} = \left(\frac{1}{4\overline{X}_n^4} - \frac{1}{4\mu^4}\right)$. Note that

$$V_{1n} = T_{1n}T_{2n}T_{3n} + \Delta^2 T_{2n}T_{3n} + \sigma^2 T_{1n}T_{3n} + \frac{1}{\mu^4}T_{1n}T_{2n} + \frac{\sigma^2}{\mu^4}T_{1n} + \frac{\Delta^2}{\mu^4}T_{2n} + \Delta^2\sigma^2 T_{3n}, \tag{A.10}$$

Let us consider the first term in the summation of Eq. A.10 and state the following inequalities.

$$\begin{aligned} P\left(\max_{n_{1c}\le n\le n_{2c}} |T_{1n}T_{2n}T_{3n}| \ge \frac{k}{28}\right) &\le \sum_{i=1}^{3} P\left(\max_{n_{1c}\le n\le n_{2c}} |T_{in}| \ge \left(\frac{k}{28}\right)^{\frac{1}{3}}\right) \\ &= O(n_{1c}^{-r}) + O(n_{1c}^{-r}) + O(n_{1c}^{-2r}) = O(n_{1c}^{-r}). \end{aligned} \tag{A.11}$$

The asymptotic orders in Eq. A.11 are obtained by using Lemma 3, maximal inequality for reverse martingales (Lee, p. 112, 1990), Lemma 2.2 of Sen and and Ghosh (1981), and Lemma 4 for $p = 2r$. The conditions of Lemma 5 are also used in Eq. A.11. Following the same argument as above, one can show that the aymptotic order of probability of large deviations (as in Eq. A.11) corresponding to the remaining six terms in the summation of Eq. A.10 are either $O(n_{1c}^{-r})$. Therefore,

$$P_1 = P\left(\max_{n_{1c}\le n\le n_{2c}} |V_{1n}| \ge \frac{k}{4}\right) \le O(n_{1c}^{-r}). \tag{A.12}$$

Note that all the estimators in V_{2n} and V_{3n} are U-statistics as we had in the case of V_{1n}. So, following similar arguments as in the proof of Eq. A.12, one can show that both P_2 and P_3 are $O(n_{1c}^{-r})$ as $c \downarrow 0$.

To work with P_4, we note that the expression of V_{4n} involves s_{wn}^2 which is not a U-statistics. Therefore, arguments given in the case of P_1-P_3 may not work without additional result. Following the proof of Lemma 3.1 of Sen and Ghosh (1981) and noting that $E\left(X_1^{4r}\right) < \infty$ for $r \ge 1$,

$$P\left(\max_{n_{1c}\le n\le n_{2c}} \left|\frac{s_{wn}^2}{4} - \sigma_1^2\right| \ge K\right) \le O(n_{1c}^{-r}), \quad \text{for any positive constant } K. \tag{A.13}$$

Noting that $V_{4n} = W_{1n}W_{2n} + \sigma_1^2 W_{2n} + \mu^{-2}W_{1n}$, where $W_{1n} = \left(\frac{s_{wn}^2}{4} - \sigma_1^2\right)$ and $W_{2n} = \left(\frac{1}{\overline{X}_n^2} - \frac{1}{\mu^2}\right)$,

$$\begin{aligned} P_4 &\leq P\left(\max_{n_{1c}\leq n\leq n_{2c}} |W_{1n}W_{2n}| \geq \frac{k}{12}\right) + P\left(\max_{n_{1c}\leq n\leq n_{2c}} |W_{2n}| \geq \frac{k}{12\sigma_1^2}\right) \\ &\quad + P\left(\max_{n_{1c}\leq n\leq n_{2c}} |W_{1n}| \geq \frac{k\mu^2}{12}\right) \\ &\leq \sum_{i=1}^{2} P\left(\max_{n_{1c}\leq n\leq n_{2c}} |W_{in}| \geq \sqrt{\frac{k}{12}}\right) + O(n_{1c}^{-2r}) + O(n_{1c}^{-r}) = O(n_{1c}^{-r}). \end{aligned} \tag{A.14}$$

The asymptotic orders in Eq. A.14 are obtained by using Lemma 4 and the inequality in Eq. A.13. We complete the proof of (i) by adding all the upper bounds for P_1-P_4 and noting that $n_{1c} = O\left(c^{-1/(2+2\gamma)}\right)$. Now, we prove the part (ii) of this theorem. Using the definition of stopping rule N_c in Eqs. 2.11 and A.1, we have

$$\begin{aligned} P(N_c > n_{3c}) &\leq P\left(n < \sqrt{\frac{A}{c}} V_n \text{ for some } n > n_{3c}\right) \\ &= P\left(n^2 < \frac{A}{c} V_n^2 \text{ for some } n > n_{3c} \text{ and } V_n^2 = V_n^*\right) \\ &\quad + P\left(n^2 < \frac{A}{c} V_n^2 \text{ for some } n > n_{3c} \text{ and } V_n^2 = 0\right) \\ &\leq P\left(n_{3c}^2 < \frac{A}{c} V_n^* \text{ for some } n > n_{3c}\right) \\ &\leq P\left(V_n^* - \xi^2 \geq \xi^2\epsilon(2+\epsilon) \text{ for some } n > n_{3c}\right) \\ &\leq P\left(|V_n^* - \xi^2| \geq \xi^2\epsilon(2+\epsilon) \text{ for some } n > n_{3c}\right) \\ &\leq P\left(\max_{n>n_{3c}} \{|V_{1n}| + |V_{2n}| + |V_{3n}| + |V_{4n}|\} \geq \xi^2\epsilon(2+\epsilon)\right), \end{aligned} \tag{A.15}$$

where V_{in}, $i = 1, \ldots, 4$, is defined above. The rest of the proof for part (ii) of Lemma 5 is very similar to the proof of part (i).

Lemma 6. *If $X_1, \ldots, X_n$ are i.i.d. random variables from $F(.)$ such that $E(X_1^{2r})$ exists for some $r \geq 1$, then*

$$E\left(\max_{n_{1c}\leq n\leq n_{2c}} (\widehat{G}_n - G_F)^{2r}\right) = O(n_{1c}^{-r}) \quad \textit{as } c \downarrow 0.$$

PROOF. Applying C_r inequality, we can write

$$\begin{aligned}\left(\widehat{G}_n - G_F\right)^{2r} &= \frac{1}{2^{2r}\overline{X}_n^{2r}}\left\{\left(\widehat{\Delta}_n - \Delta\right) - \frac{\Delta}{\mu}\left(\overline{X}_n - \mu\right)\right\}^{2r} \\ &\leq \frac{1}{2t^{2r}}\left\{\left(\widehat{\Delta}_n - \Delta\right)^{2r} + \frac{\Delta^{2r}}{\mu^{2r}}\left(\overline{X}_n - \mu\right)^{2r}\right\} \quad \text{(A.16)}\end{aligned}$$

Using Lemma 9.2.4 of Ghosh et al. (1997), we have

$$2t^{2r}E\left[\max_{n_{1c}\leq n\leq n_{2c}}(\widehat{G}_n - G_F)^{2r}\right] \quad \text{(A.17)}$$

$$\begin{aligned}&\leq E\left[\max_{n_{1c}\leq n\leq n_{2c}}\left(\widehat{\Delta}_n - \Delta\right)^{2r}\right] + \frac{\Delta^{2r}}{\mu^{2r}}E\left[\max_{n_{1c}\leq n\leq n_{2c}}\left(\overline{X}_n - \mu\right)^{2r}\right] \\ &\leq \left(\frac{2r}{2r-1}\right)^{2r}E\left(\widehat{\Delta}_{n_{1c}} - \Delta\right)^{2r} + \frac{\Delta^{2r}}{\mu^{2r}}\left(\frac{2r}{2r-1}\right)^{2r}E\left(\overline{X}_{n_{1c}} - \mu\right)^{2r} \\ &= O(n_{1c}^{-r}). \quad \text{(A.18)}\end{aligned}$$

Lemma 7. *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 7, we refer to the proof of Lemma 3 in Chattopadhyay and De (2016).

Lemma 8. *Let U_n be a U-statistics for estimating θ based on n observations with kernel ϕ such that $E(\phi^4)$ is finite. For any $\epsilon \in (0,1)$,*

$$E\left(\max_{n_{2c}\leq n\leq n_c}(U_n - U_{n_c})^4\right) = O\left(\frac{\epsilon}{n_c^2}\right) \quad as \;\; c \downarrow 0.$$

PROOF. Since $\{U_n - U_{n_c}\}_{n=n_{2c}}^{n_c}$ is a reverse martingale, Lemma 9.2.4 of Ghosh et al. (1997) yields

$$E\left(\max_{n_{2c}\leq n\leq n_c}(U_n - U_{n_c})^4\right) \leq \left(\frac{4}{3}\right)^4 E\left(U_{n_{2c}} - U_{n_c}\right)^4. \quad \text{(A.19)}$$

Let $Q_n = U_n - \theta$. Using reverse martingale property of Q_n, i.e., $E(Q_{n_{2c}} \mid \mathcal{F}_{n_c}) = Q_{n_c}$, we have

$$E\left(Q_{n_{2c}} Q_{n_c}^3\right) = E\left(Q_{n_c}^4\right), \quad E\left(Q_{n_{2c}}^3 Q_{n_c}\right) \geq E\left(Q_{n_c}^4\right), \quad \text{and} \tag{A.20}$$

$$E\left(Q_{n_{2c}}^2 Q_{n_c}^2\right) \leq \left\{E\left(Q_{n_{2c}}^4\right) E\left(Q_{n_c}^4\right)\right\}^{\frac{1}{2}} \leq E\left(Q_{n_{2c}}^4\right). \tag{A.21}$$

Using (A.20)–(A.21) and asymptotic form of 4^{th} central moment of U-statistics (Sen, p. 55, 1981),

$$\begin{aligned} & E\left(U_{n_{2c}} - U_{n_c}\right)^4 \\ & = E\left(Q_{n_{2c}}^4\right) + E\left(Q_{n_c}^4\right) - 4E\left(Q_{n_{2c}} Q_{n_c}^3\right) - 4E\left(Q_{n_{2c}}^3 Q_{n_c}\right) + 6E\left(Q_{n_{2c}}^2 Q_{n_c}^2\right) \\ & \leq 7\left\{E\left(Q_{n_{2c}}^4\right) - E\left(Q_{n_c}^4\right)\right\} = O\left(\frac{1}{n_{2c}^2} - \frac{1}{n_c^2}\right) + o\left(\frac{1}{n_c^2}\right) = O\left(\frac{\epsilon}{n_c^2}\right). \end{aligned} \tag{A.22}$$

Equation A.22 is obtained by noting that $n_{2c} = n_c(1-\epsilon)$. Hence, the proof is complete.

Lemma 9. *If $X_1, \ldots, X_n$ are i.i.d. random variables from the distribution $F(.)$ such that $E(X_1^4)$ is finite, then for $\epsilon \in (0,1)$,*

$$E\left[\max_{n_{2c} \leq n \leq n_{3c}} \left(\widehat{G}_n - \widehat{G}_{n_c}\right)^2\right] = O\left(\frac{\sqrt{\epsilon}}{n_c}\right) \quad as \;\; c \downarrow 0.$$

PROOF. $E\left[\max_{n_{2c} \leq n \leq n_{3c}} \left(\widehat{G}_n - \widehat{G}_{n_c}\right)^2\right] \leq E_1 + E_2$, where

$$\begin{aligned} E_2 &= E\left[\max_{n_c \leq n \leq n_{3c}} \left(\widehat{G}_n - \widehat{G}_{n_c}\right)^2\right], \text{ and} \\ E_1 &= E\left[\max_{n_{2c} \leq n \leq n_c} \left(\widehat{G}_n - \widehat{G}_{n_c}\right)^2\right] \\ &= E\left[\max_{n_{2c} \leq n \leq n_c} \left\{\left(\frac{1}{\overline{X}_n} - \frac{1}{\overline{X}_{n_c}}\right)\frac{\widehat{\Delta}_n}{2} + \frac{1}{2\overline{X}_{n_c}}\left(\widehat{\Delta}_n - \widehat{\Delta}_{n_c}\right)\right\}^2\right] \leq \frac{E_{11} + E_{12}}{4}, \end{aligned}$$

where $E_{11} = E\left[\max_{n_{2c} \leq n \leq n_c} \left(\frac{1}{\overline{X}_n} - \frac{1}{\overline{X}_{n_c}}\right)^2 \widehat{\Delta}_n^2\right]$ and

$$E_{12} = E\left[\max_{n_{2c} \leq n \leq n_c} \frac{1}{\overline{X}_{n_c}^2}\left(\widehat{\Delta}_n - \widehat{\Delta}_{n_c}\right)^2\right].$$

Applying Cauchy-Schwarz inequality as needed, we can write

$$\begin{aligned} E_{11} &\leq E\left[\max_{n_{2c}\leq n\leq n_c} t^{-4}\left(\overline{X}_n-\overline{X}_{n_c}\right)^2\widehat{\Delta}_n^2\right] \\ &\leq t^{-4}\left\{E\left[\max_{n_{2c}\leq n\leq n_c}\left(\overline{X}_n-\overline{X}_{n_c}\right)^4\right]\right\}^{\frac{1}{2}}\left\{E\left[\max_{n_{2c}\leq n\leq n_c}\widehat{\Delta}_n^4\right]\right\}^{\frac{1}{2}} \end{aligned}$$

Using Lemma 8, Lemma 9.2.4 of Ghosh et al. (1997) and the conditions of Lemma 9, we conclude that $E_{11} = O\left(\sqrt{\epsilon}/n_c\right)$. Similarly, using Cauchy-Schwarz inequality and Lemma 8, we have $E_{12} = O\left(\sqrt{\epsilon}/n_c\right)$. Therefore, $E_1 = O\left(\sqrt{\epsilon}/n_c\right)$. Following the same arguments as above, one can show that $E_2 = O\left(\sqrt{\epsilon}/n_c\right)$. Hence, Lemma 9 is proved.

Proof of Theorem 1.

(i) The definition of stopping rule N_c in Eq. 2.11 yields

$$\sqrt{\frac{A}{c}}\,V_{N_c} \leq N_c \leq 1+\sqrt{\frac{A}{c}}\left(V_{N_c-1}+(N_c-1)^{-\gamma}\right). \tag{A.23}$$

Since $N_c\to\infty$ a.s. as $c\downarrow 0$ and $V_n\to\xi$ a.s. as $n\to\infty$, by Theorem 2.1 of Gut (2009), $V_{N_c}\to\xi$ a.s.. Hence, dividing all sides of Eq. A.23 by n_c and letting $c\to 0$, we prove $N_c/n_c\to 1$ a.s. as $c\downarrow 0$.

(ii) Since $N_c\geq m$ a.s. and $n_c\geq 1$, dividing (A.23) by n_c yields

$$N_c/n_c \leq 1+\frac{1}{\xi}\left(\sup_{c>0} V_{N_c-1}+(m-1)^{-\gamma}\right) \quad \text{almost surely,} \tag{A.24}$$

where $E\left(\sup_{c>0} V_{N_c-1}\right) < \infty$ by Lemma 7. Since $N_c/n_c\to 1$ a.s. as $c\downarrow 0$, by the dominated convergence theorem, we conclude that $\lim_{c\downarrow 0} E(N_c/n_c)=1$.

(iii) We need to show

$$\lim_{c\downarrow 0} R_{N_c}(G_F)/R^*_{n_c}(G_F) = \lim_{c\downarrow 0}(A/2cn_c)E\left(\widehat{G}_{N_c}-G_F\right)^2+\frac{1}{2}\lim_{c\downarrow 0}E\left(N_c/n_c\right)=1.$$

It is enough to show that $\lim_{c\downarrow 0}(A/cn_c)E\left(\widehat{G}_{N_c}-G_F\right)^2=1$, i.e., $\lim_{c\downarrow 0} n_c\,E(\widehat{G}_{N_c}-G_F)^2=\xi^2$. Since we know that $\lim_{c\downarrow 0} n_c E\left(\widehat{G}_{n_c}-G_F\right)^2=\xi^2$, it is sufficient to show that

$$\lim_{c\downarrow 0} n_c\left\{E\left(\left(\widehat{G}_{N_c}-G_F\right)^2-\left(\widehat{G}_{n_c}-G_F\right)^2\right)\right\}=0. \tag{A.25}$$

Let $E_1 = E\left[(\widehat{G}_{N_c} - G_F)^2 I(N_c \leq n_{2c})\right]$. By Eq. A.1, Lemma 5, and Lemma 6, we have

$$\begin{aligned} n_c E_1 &\leq n_c E\left[\max_{n_{1c}\leq n\leq n_{2c}} (\widehat{G}_n - G_F)^2 I(N_c \leq n_{2c})\right] \\ &\leq n_c \left\{E\left[\max_{n_{1c}\leq n\leq n_{2c}} (\widehat{G}_n - G_F)^4\right] P(N_c \leq n_{2c})\right\}^{\frac{1}{2}} = O\left(c^h\right), \end{aligned} \tag{A.26}$$

where $h = (2-\gamma)/(2+2\gamma) > 0$ using $\gamma \in (0,2)$. Here, we assume that $E(X_1^4)$ exists. Following the same arguments as in Lemma 6, we can show that $E\left(\widehat{G}_{n_c} - G_F\right)^4 = O\left(n_c^{-2}\right)$ provided $E(X_1^4)$ exists. Let $E_2 = E\left[(\widehat{G}_{n_c} - G_F)^2 I(N_c \leq n_{2c})\right]$. By Cauchy-Schwarz inequality and Lemma 5, we have

$$n_c E_2 \leq n_c \left\{E\left[(\widehat{G}_{n_c} - G_F)^4\right] P(N_c \leq n_{2c})\right\}^{\frac{1}{2}} = O\left(c^{\frac{2+\gamma}{2(1+\gamma)}}\right) \tag{A.27}$$

provided $E(X_1^4)$ exists. Therefore, combining (A.26) and (A.27), we have

$$\lim_{c\downarrow 0} n_c E\left[\left\{(\widehat{G}_{N_c} - G_F)^2 - (\widehat{G}_{n_c} - G_F)^2\right\} I(N_c \leq n_{2c})\right] = 0. \tag{A.28}$$

Using the same arguments as in Lemma 6, one can show that

$$E\left[\max_{n\geq n_{3c}} \left(\widehat{G}_n - G_F\right)^4\right] = O\left(n_{3c}^{-2}\right) \text{ provided } E(X_1^4) \text{ exists.}$$

Let $E_3 = E\left[(\widehat{G}_{N_c} - G_F)^2 I(N_c > n_{3c})\right]$. Cauchy-Schwarz inequality and Lemma 5 yields

$$n_c E_3 \leq n_c \left\{E\left[\max_{n\geq n_{3c}} (\widehat{G}_n - G_F)^4\right] P(N_c > n_{3c})\right\}^{\frac{1}{2}} = O\left(c^{\frac{2+\gamma}{2(1+\gamma)}}\right). \tag{A.29}$$

Following the same approach as in Eq. A.27, $n_c E\left[(\widehat{G}_{n_c} - G_F)^2 I(N_c > n_{3c})\right] = O\left(c^{\frac{2+\gamma}{2(1+\gamma)}}\right)$. Thus,

$$\lim_{c\downarrow 0} n_c E\left[\left\{(\widehat{G}_{N_c} - G_F)^2 - (\widehat{G}_{n_c} - G_F)^2\right\} I(N_c > n_{3c})\right] = 0. \tag{A.30}$$

Hence, it remains to prove that

$$\lim_{c\downarrow 0} n_c E\left[\left\{(\widehat{G}_{N_c}-G_F)^2-(\widehat{G}_{n_c}-G_F)^2\right\} I(n_{2c}\le N_c\le n_{3c})\right]=0. \quad \text{(A.31)}$$

Let $W=\left\{(\widehat{G}_{N_c}-G_F)^2-(\widehat{G}_{n_c}-G_F)^2\right\} I(n_{2c}\le N_c\le n_{3c})$. Note that

$$\begin{aligned} W &= \left\{(\widehat{G}_{N_c}-G_F)+(\widehat{G}_{n_c}-G_F)\right\}(\widehat{G}_{N_c}-\widehat{G}_{n_c})I(n_{2c}\le N_c\le n_{3c}) \\ &\le 2\left\{\max_{n_{2c}\le n\le n_{3c}}\left|\widehat{G}_n-G_F\right|\right\}\left\{\max_{n_{2c}\le n\le n_{3c}}\left|\widehat{G}_n-\widehat{G}_{n_c}\right|\right\} I(n_{2c}\le N_c\le n_{3c}). \end{aligned}$$

Using Cauchy-Schwarz inequality, Lemma 9, and following the lines of Lemma 6,

$$\begin{aligned} n_c E(W) &\le 2n_c\left\{E\left(\max_{n_{2c}\le n\le n_{3c}}\left(\widehat{G}_n-G_F\right)^2\right)E\left(\max_{n_{2c}\le n\le n_{3c}}\left(\widehat{G}_n-\widehat{G}_{n_c}\right)^2\right)\right\}^{\frac{1}{2}} \\ &\le 2n_c\left\{O\left(n_c^{-1}\right)O\left(\frac{\sqrt{\epsilon}}{n_c}\right)\right\}^{\frac{1}{2}}=O\left(\epsilon^{1/4}\right). \quad \text{(A.32)} \end{aligned}$$

Since (A.32) is true for any $\epsilon\in(0,1)$, taking limit on both sides of Eq. A.32 as $\epsilon\to 0$, (A.31) is proved. Hence, the proof of Theorem 1 is complete.

Proof of Lemma 1 . By bivariate Taylor expansion of $f(\widehat{\Delta}_n, 2\overline{X}_n)=\frac{\widehat{\Delta}_n}{2\overline{X}_n}$ around $(\Delta, 2\mu)$,

$$\frac{\widehat{\Delta}_n}{2\overline{X}_n}-\frac{\Delta}{2\mu}=\frac{\widehat{\Delta}_n-\Delta}{2\mu}-\frac{\Delta}{2\mu^2}(\overline{X}_n-\mu)+R_{1n}, \quad \text{(A.33)}$$

where $R_{1n}=-2(\widehat{\Delta}_n-\Delta)(\overline{X}_n-\mu)/b^2+4a(\overline{X}_n-\mu)/b^3$, $a=\Delta+p(\widehat{\Delta}_n-\Delta)$, $b=2\mu+p(2\overline{X}_n-2\mu)$, and $p\in(0,1)$. Let $E_{1n}=E(R_{1n}^2)$, $E_{2n}=\frac{1}{\mu}E\left(R_{1n}\left(\widehat{\Delta}_n-\Delta\right)\right)$, and $E_{3n}=-\frac{\Delta}{\mu^2}E\left(R_{1n}\left(\overline{X}_n-\mu\right)\right)$. Squaring both sides of Eq. A.33 and taking expectation,

$$E\left(\frac{\widehat{\Delta}_n}{2\overline{X}_n}-\frac{\Delta}{2\mu}\right)^2=\frac{1}{4\mu^2}V(\widehat{\Delta}_n)+\frac{\Delta^2\sigma^2}{4n\mu^4}-\frac{\Delta}{2\mu^3}cov(\widehat{\Delta}_n,\overline{X}_n)+\sum_{i=1}^{3}E_{in}. \quad \text{(A.34)}$$

Using variance and covariance formulas for U-statistics (Lee, 1990), it is simple to show that $V\left(\widehat{\Delta}_n\right)=\frac{4\sigma_1^2}{n}+O\left(n^{-2}\right)$ and $cov(\widehat{\Delta}_n,\overline{X}_n)=\frac{2}{n}(\tau-\mu\Delta)$.

Therefore, it remains to show that $\sum_{i=1}^{3} E_{in} = O(n^{-3/2})$. First, we work on E_{1n}. Note that $R_{1n}^2 = 4W_{1n} + 16W_{2n} - 16W_{3n}$, where

$$W_{1n} = \frac{\left(\widehat{\Delta}_n - \Delta\right)^2 \left(\overline{X}_n - \mu\right)^2}{b^4} \leq \frac{1}{16t^4} \left(\widehat{\Delta}_n - \Delta\right)^2 \left(\overline{X}_n - \mu\right)^2, \quad \text{(A.35)}$$

$$W_{2n} = \frac{a^2}{b^6}(\overline{X}_n - \mu)^4 \leq \frac{a^2}{64t^6}(\overline{X}_n - \mu)^4, \text{ and} \quad \text{(A.36)}$$

$$W_{3n} = \frac{a}{b^5} \left(\widehat{\Delta}_n - \Delta\right) \left(\overline{X}_n - \mu\right)^3 \leq \frac{a}{32t^5} \left(\widehat{\Delta}_n - \Delta\right) \left(\overline{X}_n - \mu\right)^3 \quad \text{(A.37)}$$

By Cauchy-Schwarz inequality and Lemma 2.2 of Sen and Ghosh (1981),

$$E\left|W_{1n}\right| \leq \frac{1}{16} E\left(\left(\widehat{\Delta}_n - \Delta\right)^2 \left(\overline{X}_n - \mu\right)^2\right) = O(n^{-2}), \quad \text{(A.38)}$$

$$E\left|W_{2n} I(\widehat{\Delta}_n > \Delta)\right| \leq \frac{1}{64t^6} E\left((\overline{X}_n - \mu)^4 \widehat{\Delta}_n^2\right) = O(n^{-3}) \quad \text{(A.39)}$$

$$E\left|W_{2n} I(\widehat{\Delta}_n \leq \Delta)\right| \leq \frac{\Delta^2}{(2t)^6} E\left(\overline{X}_n - \mu\right)^4 = O(n^{-2}), \quad \text{(A.40)}$$

provided $E(X_1^6)$ is finite. The asymptotic order in Eq. A.38 is obtained by using Holder's inequality. Similarly, we can show that $E(W_{3n}) = O(n^{-2})$ provided $E(X_1^6)$ exists. Therefore, $E_{1n} = E(R_{1n}^2) = O(n^{-2})$. By Cauchy-Schwarz inequality and Lemma 2.2 of Sen and Ghosh (1981), we obtain $E_{2n} = O(n^{-3/2})$ and $E_{3n} = O(n^{-3/2})$. Hence, Lemma 1 is proved.

References

AGUIRREGABIRIA, V. and MIRA, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica* **75**, 1–53.

ALLISON, P.D. (1978). Measures of inequality. *Am. Sociol. Rev.*, 865–880.

ARCIDIACONO, P. and JONES, J.B. (2003). Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica* **71**, 933–946.

ASADA, Y. (2005). Assessment of the health of americans: the average health-related quality of life and its inequality across individuals and groups. *Popul. Health Metrics* **3**, 7.

BEACH, C.M. and DAVIDSON, R. (1983). Distribution-free statistical inference with lorenz curves and income shares. *Rev. Econ. Stud.* **50**, 723–735.

CHATTOPADHYAY, B. and DE, S.K. (2014). *Estimation accuracy of an inequality index*. Recent advances in applied mathematics, modelling and simulation. In: MASTORAKIS, N.E., DEMIRALP, M., MUKHOPADHYAY, N. and MAINARD, F. (eds.) Recent Advances in Applied Mathematics, Modelling and Simulation, WSEAS, p. 318–321.

CHATTOPADHYAY, B. and DE, S.K. (2016). Estimation of Gini index within pre-specified error bound. *Econometrics* **4**, 30. doi: 10.3390/econometrics4030030.

CHATTOPADHYAY, B. and MUKHOPADHYAY, N. (2013). Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Seq. Anal.* **32**, 134–157.

COCHRAN, W.G. (1977). *Sampling techniques, 98*. Wiley, New York.

DANTZIG, G.B. (1940). On the non-existence of tests of "student's" hypothesis having power functions independent of σ. *Ann. Math. Stat.* **11**, 186–192.

DAS, A. and ROUT, H.S. (2015). *The social sector in India: issues and challenges*. Cambridge Scholars Publishing, UK. chap 10.

DAVIDSON, R. (2009). Reliable inference for the Gini index. *J. Econ.* **150**, 30–40.

DAVIDSON, R. and DUCLOS, J.Y. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* **68**, 1435–1464.

DOOB, J.L. (1953). *Stochastic processes*. Wiley, New York.

GASTWIRTH, J.L. (1972). The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.*, 306–316.

GHOSH, B.K. and SEN, P.K. (1991). *Handbook of sequential analysis*, vol 118. CRC Press.

GHOSH, M. and MUKHOPADHYAY, N. (1979). Sequential point estimation of the mean when the distribution is unspecified. *Commun. Stat.-Theory Methods* **8**, 637–652.

GHOSH, M., MUKHOPADHYAY, N. and SEN, P.K. (1997). *Sequential estimation*. Wiley, New York.

GREENE, W.H. (1998). Gender economics courses in liberal arts colleges: further results. *J. Econ. Educ.* **29**, 291–300.

GUT, A. (2009). *Stopped random walks: Limit theorems and applications*. Springer.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19**, 293–325.

HOEFFDING, W. (1961). The strong law of large numbers for u-statistics. Institute of Statistics mimeo series 302.

HOLLANDER, M. and WOLFE, D.A. (1999). *Nonparametric statistical methods*. Wiley, New York.

KANNINEN, B.J. (1993). Design of sequential experiments for contingent valuation studies. *J. Environ. Econ. Manag.* **25**, S1–S11.

LEE, A.J. (1990). *U-statistics: theory and practice*. CRC Press.

LOÈVE, M. (1963). *Probability theory*. Van Nostrand, Princeton.

LOOMES, G. and SUGDEN, R. (1982). Regret theory: an alternative theory of rational choice under uncertainty. *Econ. J.* **92**, 805–824.

MUKHOPADHYAY, N. and CHATTOPADHYAY, B. (2012). A tribute to Frank Anscombe and random central limit theorem from 1952. *Seq. Anal.* **31**, 265–277.

MUKHOPADHYAY, N. and DE SILVA, B.M. (2009). *Sequential methods and their applications*. CRC Press.

ROBBINS, H. (1959). *Sequential estimation of the mean of a normal population*. Almquist and Wiksell, Uppsala, p. 235–245.

SEN, P.K. (1981). *Sequential nonparametrics: invariance principles and statistical inference*. Wiley, New York.

SEN, P.K. (1988). Functional jackknifing: rationality and general asymptotics. *Ann. Stat.*, 450–469.

SEN, P.K. and GHOSH, M. (1981). Sequential point estimation of estimable parameters based on u-statistics. Sankhyā: The Indian Journal of Statistics, Series A pp. 331–344.

shi, h. and sethu, h. (2003). *Greedy fair queueing: a goal-oriented strategy for fair real-time packet scheduling*. IEEE, p. 345–356.

sproule, r. (1969). *A sequential fixed-width confidence interval for the mean of a u-statistic*. PhD thesis Ph, D. dissertation, Univ. of North Carolina.

thomas, v., wang, y. and fan, x. (2001). *Measuring education inequality: Gini coefficients of education*, vol. 2525. World Bank Publications.

wittebolle, l., marzorati, m., clement, l., balloi, a., daffonchio, d., heylen, k., de vos, p., verstraete, w. and boon, n. (2009). Initial community evenness favours functionality under selective stress. *Nature* **458**, 623–626.

xu, k. (2007). U-statistics and their asymptotic results for some inequality and poverty measures. *Econ. Rev.* **26**, 567–577.

yitzhaki, s. and schechtman, e. (2013). *The Gini methodology: a primer on a statistical methodology*. Springer.

Shyamal K. De
School of Mathematical Sciences,
National Institute of Science
Education and Research, HBNI
Jatni, Odisha, 752050, India
E-mail: sde@niser.ac.in

Bhargab Chattopadhyay
Indian Institute of Information
Technology Vadodara,
Gandhinagar, Gujarat 382028, India
Department of Mathematical
Sciences,
University of Texas at Dallas,
Richardson, TX 75080, USA
E-mail: bhargab@iiitvadodara.ac.in
bhargab@utdallas.edu

Paper received: 6 September 2015.