# Sequential Confidence Set and Point Estimation of the Population Gini Index by Controlling Accuracies Relative to the Population Mean


Nitis Mukhopadhyay[1], Shyamal K. De[2], and Tae Yang[3]

[1]Department of Statistics, University of Connecticut-Storrs, Connecticut, USA

[2]School of Mathematical Sciences, National Institute of Science Education and Research, HBNI, Odisha, India

[3]Department of Mathematics, College of Science: Myongji University, Seoul, Korea



**Abstract:** Chattopadhyay and De (2016) developed both theory and methodology for constructing a confidence interval for Gini index ($G$) with a specified confidence coefficient $1-\alpha$ and a specified width ($=2d$). In a sequel, De and Chattopadhyay (2017) developed both theory and methodology for constructing a minimum risk point estimator for $G$. In either problem, these authors formulated the part of their loss due to estimation as the *absolute difference* between $\widehat{G}$ (estimator of $G$) and $G$ without assuming any specific distribution of the data. In the present paper, we provide both economic persuasion and motivation behind a new formulation of the part of the loss due to estimation as a *weighted* (relative to the mean) *absolute difference* between $\widehat{G}$ and $G$. We have treated both confidence set estimation as well as minimum risk point estimation problems for $G$ by controlling accuracies relative to the population mean without assuming any specific distribution of the data. Optimality properties of the proposed purely sequential method under confidence set estimation include first-order asymptotic efficiency and first-order asymptotic consistency properties. Characteristics of the purely sequential method under minimum risk point estimation relative to the population mean are examined and a number of desirable asymptotic optimality properties are proved without assuming any specific distribution of the data. In the second problem, the optimality properties of the proposed purely sequential method include first-order asymptotic efficiency and first-order asymptotic risk efficiency properties. Overall, our proofs of the requisite asymptotic properties reported here require less stringent sufficient moment conditions than those used by Chattopadhyay and De (2016) and De and Chattopadhyay (2017). Small and moderate sample size performances of the new methods are examined through extensive sets of simulation studies.



**Keywords:** Asymptotic consistency; Confidence set; First-order efficiency; First-order risk efficiency; Minimum risk; Point estimation; Relative-accuracy; Sequential sampling; Stopping rule; U-statistics.

**Subject Classifications:** 62L12; 62G05; 60G46; 60G40; 91B82.


## 1. INTRODUCTION

Economic inequality exists in almost all aspects of life. It is very important to measure the inequality periodically for a region or a country of interest to gauge the effect of current economic policies in order to critically consider possible paths for future policies. Among many available measures of inequality, Gini index or Gini coefficient is a leading measure for income or wealth distribution in a region or country.

Suppose $X$ is a non-negative non-degenerate random variable having *distribution function* (d.f.) $F$ that represents income or wealth of a person or a household. If $X_1$ and $X_2$ are two i.i.d. copies of $X$, then a population Gini index, as given in Arnold (2005), is defined as:


Address correspondence to Tae Yang, Department of Mathematics, College of Science, Myongji University, Geobukgol-ro 34, 03674, Seoul, Gyeonggi-do, South Korea ; E-mails: nitis.mukhopadhyay@uconn.edu; sde@niser.ac.in; tyang@mju.ac.kr

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

A population Gini index (Gini 1914,1921) for a country is typically calculated based on census data every ten years. However, for continuous monitoring of economic policies, it is often crucial to estimate Gini index by observing samples from a population during some intermediate time periods. Typically, to estimate the inequality index for larger regions, stratified sampling or some other appropriate complex sampling scheme is adopted.

For practical purposes, however, in a smaller region, simple random sampling may be used to produce nearly *independent and identically distributed* (i.i.d.) observations. For elaborate discussions on simple random sampling schemes in the context of household data, one may refer to miscellaneous sources including Beach and Davidson (1983), Chattopadhyay and De (2016), Davidson (2009), Davidson and Duclos (2000), De and Chattopadhyay (2017), Gastwirth (1972), and Xu (2007).

Let us consider $n$ i.i.d. observations $X_1, \ldots, X_n$ from the common distribution function $F$ with its support $(0, \infty)$. An estimator of Gini index $G$ 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 follows:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \widehat{\Delta}_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \left|X_i - X_j\right|. \tag{1.3}$$

Note that $\overline{X}_n$ and $\widehat{\Delta}_n$ are both U-statistics (Hoeffding 1948, Lee 1990) of degree one and two respectively. Exploiting various crucial properties of U-statistics, Chattopadhyay and De (2016) and De and Chattopadhyay (2017) recently developed purely sequential methodologies for estimation of Gini index.

With some prefixed half-width $d > 0$, Chattopadhyay and De (2016) constructed a fixed-width $(= 2d)$ confidence interval of the form $\left[\widehat{G}_N \pm d\right]$ based on the finally accrued data $\{N, X_1, \ldots, X_N\}$ such that asymptotically, we have:

$$P_F\left(G \in \left[\widehat{G}_N \pm d\right]\right) \approx 1 - \alpha, \tag{1.4}$$

for some pre-specified level $\alpha \in (0, 1)$ in the spirit of Chow and Robbins (1965). The associated stopping variable, that is, the final sample size of the sequential procedure $N$ was shown to enjoy a number of desirable asymptotic properties including the *first-order asymptotic efficiency* property, defined by Ghosh and Mukhopadhyay (1981), and the *asymptotic consistency* property (Chow and Robbins 1965).

In a sequel, De and Chattopadhyay (2017) developed a purely sequential methodology for minimum risk point estimation of Gini index $G$ in the spirits of Robbins (1959) and Ghosh and Mukhopadhyay (1979). The loss function under that methodology was a combination of the squared error loss due to wrong estimation plus a linear cost, namely it had the following form:

$$A\left(\widehat{G}_n - G\right)^2 + cn, \tag{1.5}$$

where $A$ is a known positive constant representing the weight assigned by the researcher regarding the probable cost per unit squared error loss due to estimation and $c$ is a known positive constant representing the cost for sampling one observation. De and Chattopadhyay (2017) constructed a minimum risk point estimator (MRPE) of the form $\widehat{G}_N$ based on finally accrued data $\{N, X_1, \ldots, X_N\}$. The associated sequential risk became asymptotically close to the minimum risk. The stopping variable of the purely sequential procedure $N$ enjoyed the *first-order asymptotic efficiency* and *first-order asymptotic risk efficiency* properties in the spirits of Ghosh and Mukhopadhyay (1981).

Both Chattopadhyay and De (2016) and De and Chattopadhyay (2017) developed ways to handle a number of crucial technical difficulties in assessing some of the desirable asymptotic optimality properties for their proposed sequential estimation methodologies, especially in the verification of the first-order asymptotic efficiency property. In these cited papers, the authors assumed that the support of $F$ was $(t, \infty)$ with some $t > 0$ in order to keep the required moment conditions within reason and practicality. We attempt to overcome those difficulties in this article by proposing some competing notions for measuring estimation error.

Note that Chattopadhyay and De (2016) and De and Chattopadhyay (2017) focused on their loss due to estimation as a function of $|\widehat{G}_N - G|$. If $F$ is assumed continuous, then $G \in (0,1)$ and $\widehat{G}_N \in (0,1)$ *with probability* (w.p.) 1. On the other hand, however, for any reasonable methodology, one should expect $|\widehat{G}_N - G|$ to be rather small to begin with especially since it is the absolute difference between two fractions. That is, $|\widehat{G}_N - G|$ may not be a truly great quantification of the discrepancy between $\widehat{G}_N$ and $G$. For the sake of arguments, consider two artificial datasets A and B, both of size $n = 5$, showing weekly income per family from an extremely poor suburb (with distribution function $F(x)$) and not-so-poor suburb (with distribution function $F(x/10)$) of a city respectively:

| | | | | | |
|---|---|---|---|---|---|
| Dataset A (d.f. $F(x)$): | \$1 | \$2 | \$3 | \$4 | \$5 |
| Dataset B (d.f. $F(x/10)$): | \$10 | \$20 | \$30 | \$40 | \$50 |

In the case of either dataset, $\widehat{G}_n = \frac{1}{3}$. Using the Gini index alone, one may be tempted to postulate that these two suburbs have the same level of income inequality based on observed data. However, such a conclusion may misrepresent the true situation on hand as explained next.

It is clear that the families from the extremely poor suburb have practically no weekly purchasing power to acquire anything of importance (e.g., a loaf of bread) because the sample mean from this dataset happens to be \$3 per family per week. However, the families from the not-so-poor suburb has more weekly purchasing power to acquire a loaf of bread because the sample mean from this dataset happens to be \$30 per family per week.

From this economic persuasion, we may not feel very comfortable and complacent to declare this quickly that these two suburbs have the same level of income inequality simply because they have equal estimated $G$ from observed data. One may refer to Mukhopadhyay and Chattopadhyay (2018) which put forward significant economic persuasions in conjunction with alternatives to GMD or Gini's inequality index in order to address a number of crucial issues.

From the aforementioned discussions, we suggest that one may alternatively consider the loss due to estimation, namely $|\widehat{G}_N - G|$, but look at its magnitude relative to the magnitude of an appropriate power of $\mu$. To be specific, we should compare the values of

$$\mu^a |\widehat{G}_N - G|, \text{ where } a(>0) \text{ is known,}$$

for the two suburbs under consideration.

The true Gini index from (1.1) in the two suburbs will both be obviously equal, say $G$, since one distribution is a scale transformation of the other. Hence, it may be reasonable to compare $\widehat{\mu}_n^a \widehat{G}_n$ obtained from the two suburbs. In order to fix ideas, we may suppose that $a = 2$. In the case of dataset A, we have $\widehat{\mu}_n^2 \widehat{G}_n = 3^2 \times \frac{1}{3} = 3$ whereas in the context of dataset B, we have $\widehat{\mu}_n^2 \widehat{G}_n = 30^2 \times \frac{1}{3} = 300$ which should order such weighted inequality measure associated with dataset B at a much higher level of income inequality than that from the dataset A. This prompts us to consider a revised loss function in the contexts of fixed-width interval estimation problem as well as the minimum risk point estimation problem.

### 1.1. Revised Loss Functions

In the context of confidence set estimation of $G$, with fixed $n$ and predefined $d > 0$, we suppose that the $0 - 1$ estimation loss is given by:

$$L_n(G, \widehat{G}_n) = \begin{cases} 0 & \text{if } \mu^a |\widehat{G}_n - G| \leq d \\ 1 & \text{if } \mu^a |\widehat{G}_n - G| > d. \end{cases} \tag{1.6}$$

with the associated risk function $P_F\left(\mu^a |\widehat{G}_n - G| > d\right)$. That is, instead of (1.4), we should require:

$$P_F\left(\mu^a |\widehat{G}_N - G| \leq d\right) \approx 1 - \alpha, \tag{1.7}$$

based on the finally accrued data $\{N, X_1, \ldots, X_N\}$. In (1.7), $N$ would be an observable and appropriate stopping time to be defined shortly.

Analogously, with a fixed sample of size $n$, in the context of minimum risk point estimation of $G$ we would work under the weighted squared error loss plus linear cost of sampling defined as follows:

$$L_n(G, \widehat{G}_n) = A\mu^a \left(\widehat{G}_n - G\right)^2 + cn, \tag{1.8}$$

with the associated risk function $A\mu^a E_F \left[\widehat{G}_n - G\right]^2 + cn$. That is, instead of minimizing the fixed sample size risk function associated with (1.5), we should minimize the fixed-sample-size risk function associated with (1.8), requiring that we handle the following expression based on finally accrued data $\{N, X_1, \ldots, X_N\}$:

$$R_N(G) = A\mu^a E_F \left[\left(\widehat{G}_N - G\right)^2\right] + cE_F[N]. \tag{1.9}$$

Here, $N$ is an observable stopping time or the final sample size of a sequential procedure to be defined shortly. We assume in (1.8) and (1.9) that $a > 0$, $A > 0$ and $c > 0$ are all pre-specified.

### 1.2. An Overview and the Layout of the Paper

The literature on sequential and multi-stage sampling methodologies in the constructions of fixed-width confidence intervals and minimum risk point estimators of various parameters of interest are as broad ranging as one may imagine spanning the field of both parametric and nonparametric inferences. Many relevant concepts, desirable properties, and techniques of proofs are far reaching. Initially, one may review from the original sources including Stein (1945,1949), Anscombe (1952,1953), Ray (1957), Robbins (1959), and Chow and Robbins (1965). In those methodologies, a population standard deviation was traditionally replaced by suitable consistent estimators in the form of a multiple of the sample standard deviation.

Mukhopadhyay and Chattopadhyay (2013a,b) and Mukhopadhyay and Hu (2017,2018) have proposed alternative sampling strategies by replacing a population standard deviation by consistent estimators in the form of a multiple of GMD as well as the *mean absolute deviation* (MAD) for normally distributed data with possible outliers. One may accomplish a broader review from the texts of Sen (1981), Mukhopadhyay and Solanky (1994), Ghosh et al. (1997), and Mukhopadhyay and de Silva (2009).

In Section 2, we formally state the relative-accuracy confidence set estimation problem based on loss function (1.6), develop a purely sequential procedure with some asymptotic optimality properties and study its asymptotic first-order optimality properties (Theorems 2.1 and 2.2). In Section 3, we formally discuss the minimum relative risk point estimation problem based on loss function (1.8), propose a purely sequential procedure and study its asymptotic optimality properties (Theorems 3.1 and 3.2). Section 4 presents summaries from simulation studies validating all theoretical results under the proposed methodologies. Section 5 provides a number of important subsidiary lemmas and shows brief outlines of proofs of Theorems 2.1-2.2 and Theorems 3.1-3.2

## 2. RELATIVE-ACCURACY CONFIDENCE SET ESTIMATION

The estimator $\widehat{G}_n$ in (1.2), based on $n$ i.i.d. observations $X_1, \ldots, X_n$ from $F$, is a strongly consistent estimator of Gini index $G$ (Chattopadhyay and De 2016) to be used throughout this paper. For a prescribed level $\alpha \in (0, 1)$, we would like to have a $(1-\alpha)$ relative-accuracy confidence set of $G$ in the form of $J_n = [\widehat{G}_n \pm d\mu^{-a}]$ such that

$$P_F \left(\mu^a \left|\widehat{G}_n - G\right| \le d\right) \ge 1 - \alpha. \tag{2.1}$$

It is understood that (2.1) may be satisfied only approximately when $n$ becomes large. To this end, we note that $\widehat{G}_n$ is asymptotically normal (Hoeffding 1948,1961; Xu 2007)) when $E_F[X_1^2] < \infty$. That is,

$$\sqrt{n}\left(\widehat{G}_n - G\right) \xrightarrow{\mathcal{L}} N\left(0, \xi^2\right) \quad \text{as } n \to \infty, \tag{2.2}$$

where the asymptotic variance $\xi^2$ is given by

$$\xi^2 \equiv \xi_F^2 = \frac{\Delta^2}{4\mu^4}\sigma^2 - \frac{\Delta\tau}{\mu^3} + \frac{\Delta^2}{\mu^2} + \frac{\sigma_1^2}{\mu^2}, \tag{2.3}$$

and

$$\sigma^2 = V_F(X_1),\ \tau = E_F[X_1\,|X_1 - X_2|],\ \text{and } \sigma_1^2 = V_F[E_F\{|X_1 - X_2| \,\big|\, X_1\}]. \tag{2.4}$$

Thus, for large $n$, we may conclude:

$$P_F\left(\mu^a\,|\widehat{G}_n - G| \leq d\right) \approx 2\Phi\left(\frac{\sqrt{n}d}{\xi\mu^a}\right) - 1 = 1 - \alpha$$

provided that

$$n \text{ is the smallest integer} \geq \beta(d)\mu^{2a}\xi^2 = n_d,\ \text{say.} \tag{2.5}$$

Here, $z_{\alpha/2}$ is the upper $(\alpha/2)^{th}$ quantile of the standard normal distribution and $\beta(d) = d^{-2}z_{\alpha/2}^2$.

Clearly, $n_d$ is the optimal (minimal) sample size required to construct a relative-accuracy confidence set for $G$ with approximately $1-\alpha$ coverage probability. We tacitly disregard that $n_d$ may not be an integer.

**Remark 2.1.** In customary applications, it is not unrealistic to expect that $\mu$ would exceed 1. Under this reasonable perception, as we compare the requirements proposed via (1.4) and (1.7), we readily observe that under (1.7) we are certainly controlling a larger estimation error namely, $\mu^a\,|\widehat{G}_n - G|$ instead of claiming that $|\widehat{G}_n - G|$ alone goes below the preassigned half-width $d > 0$.

The optimal sample size in (2.5) is unknown since $\mu$ and $\xi$ remain unknown. We must estimate them first to develop the methodology for relative-accuracy confidence interval estimation. We first formulate a strongly consistent estimator of $\xi^2$. Towards that end, we let:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \ \text{ and } \ \widehat{\tau}_n = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\frac{1}{2}(X_i + X_j)|X_i - X_j|, \tag{2.6}$$

that is, $S_n^2$ represents the sample variance which obviously estimates $\sigma^2$.

Next, recall that $\sigma_1^2 = V_F[E_F\{|X_1 - X_2| \,\big|\, X_1\}]$. For each $j = 1,\ldots,n$, we obtain:

$$\widehat{\Delta}_n^{(j)} = \binom{n-1}{2}^{-1}\sum_{T_j}|X_{i_1} - X_{i_2}|, \tag{2.7}$$

where $T_j = \{(i_1, i_2) : 1 \leq i_1 < i_2 \leq n \text{ and } i_1, i_2 \neq j\}$. Finally, we define:

$$S_{wn}^{\ 2} = (n-1)^{-1}\sum_{i=1}^{n}(W_{jn} - \overline{W}_n)^2, \tag{2.8}$$

given in Sproule (1969,1985), where $W_{jn} = n\widehat{\Delta}_n - (n-2)\widehat{\Delta}_n^{(j)}$ for $j = 1,\ldots,n$, and $\overline{W}_n = n^{-1}\sum_{j=1}^n W_{jn}$. We note that $S_{wn}^2$ is a strongly consistent estimator of $4\sigma_1^2$.

Now, combining (2.6)-(2.8) with (2.3)-(2.4), we propose to estimate $\xi^2$ with the following strongly consistent estimator:

$$V_n^{\ 2} = \max\{0, V_n^*\} \ \text{ where we let } V_n^* = \frac{\widehat{\Delta}_n^2 S_n^2}{4\overline{X}_n^4} - \frac{\widehat{\Delta}_n}{\overline{X}_n^3}\widehat{\tau}_n + \frac{\widehat{\Delta}_n^2}{\overline{X}_n^2} + \frac{S_{wn}^2}{4\overline{X}_n^2}. \tag{2.9}$$

This expression of $V_n^*$ was originally developed in Chattopadhyay and De (2016). For a number of other plug-in estimators of the asymptotic variance parameter $\xi^2$ from (2.3), one may refer to Davidson (2009) and Langel and Tille (2013).

### 2.1. Purely Sequential Sampling Methodology

Since the optimal sample size $n_d$ given by (2.5) is unknown up to $\mu$ and $\xi$, to achieve approximate or asymptotic $1-\alpha$ coverage probability, we must draw samples at least in two stages. In the first stage, we estimate $n_d$ by estimating $\mu$ and $\xi$ based on a pilot sample of size $m$, and then in the subsequent stages we should collect samples until some stopping criteria are satisfied.

In this paper, we propose a purely sequential procedure for $1-\alpha$ relative-accuracy confidence set estimation of $G$ in the spirit of Chattopadhyay and De (2016). Suppose $m$ denotes the pilot sample size. We discuss a specific choice of a pilot sample size in Remark 2.2.

The purely sequential procedure collects $X_1,\ldots,X_m$ i.i.d. observations from $F$ in the first stage and then collects one observation (independent of the previous ones) at-a-time until the current sample size is more or equal to the estimate of the optimal sample size $n_d$ from (2.5). Thus, we may define the following stopping rule or the final sample size of the procedure as follows:

$$N \equiv N_d = \inf\left\{n \geq m : n \geq \beta(d)\left(\overline{X}_n^{2a} V_n^2 + n^{-\gamma}\right)\right\} \text{ with } \beta(d) = d^{-2} z_{\alpha/2}^2, \tag{2.10}$$

where $\gamma$ is some known positive constant.

After termination of sampling, based on the finally accrued data $\{N, X_1,\ldots,X_N\}$, we propose to estimate $G$ with

$$\widehat{G}_{N_d} \equiv \widehat{G}_N = \widehat{\Delta}_N/2\overline{X}_N \text{ giving rise to the confidence set } J_N = [\widehat{G}_N \pm d\mu^{-a}]. \tag{2.11}$$

Note that the estimator $V_n^2$ from (2.9) or $\overline{X}_n$ can be close to zero with positive probability leading to early termination of sequential sampling, which in turn may cause loss of coverage probability. The term $n^{-\gamma}$ in (2.10) is introduced to avoid such a potential problem of early stopping. Chattopadhyay and De (2016) used $\gamma = 1$ in the definition of their stopping rule and for their ensuing sequential estimation methodology.

**Remark 2.2.** From (2.10), note that $N \geq \beta(d)N^{-\gamma}$ since $\overline{X}_n^{2a} V_n^2 \geq 0$ w.p.1. That is, the final sample size $N$ must be at least $\beta(d)^{1/(1+\gamma)}$. Therefore, in order to implement the purely sequential procedure (2.10), we select the pilot sample size as $m = \max\{4, \lceil \beta(d)^{1/(1+\gamma)} \rceil\}$ where $\lceil x \rceil$ denotes the smallest integer $\geq x$.

### 2.2. Asymptotic First-Order Properties

Before establishing the major asymptotic properties (Theorems 2.1-2.2) under the proposed sequential procedure (2.10)-(2.11), we state the following lemma that ensures termination w.p. 1 under the sequential sampling procedure (2.10).

**Lemma 2.1.** *For any fixed $d > 0$, the stopping time $N_d$ from (2.10) is finite w.p. 1, that is, $P_F(N_d < \infty) = 1$ if we assume $0 < \xi_F < \infty$.*

We omit its proof since it follows directly from Lemma A1 of Chattopadhyay and De (2016). Next, we establish the asymptotic properties of the proposed sequential procedure in the spirits of Theorems 1 and 2 of Chattopadhyay and De (2016).

**Theorem 2.1.** *If the parent distribution $F$ of $X$ is such that $E_F[X^2] < \infty$, then the stopping rule in (2.10) yields as $d \downarrow 0$:*

*(i)* $N_d/n_d \to 1$ w.p. 1 under $F$;

*(ii)* $P_F\left(\mu^a|\widehat{G}_{N_d} - G| \leq d\right) \to 1-\alpha\,;$

*where $G$, $n_d$ and $\widehat{G}_{N_d}$ respectively come from (1.1), (2.5),* and *(2.11).*

We omit its proof noting that it would follow very similarly along the lines of the proof of Theorem 1 in Chattopadhyay and De (2016). The following remark summarizes important information.

**Remark 2.3.** Hoeffding(1948) suggested that we only need finiteness of second moment, i.e., $E_F[X^2] < \infty$ to ensure the existence of asymptotic variance $\xi^2$. Moreover, it follows from Sproule (1969,1985) that $S_{wn}^2$ is a strong consistent estimator of $4\sigma_1^2$ if $E_F[X^2] < \infty$, and hence, we only need finiteness of second moment to guarantee that $V_n^2$ is a strongly consistent estimator of $\xi^2$. This consistency property is the key to prove part $(i)$ and part $(i)$ is used in the proof of part $(ii)$ of Theorem 2.1. In this context, note that a sufficient condition for Theorem 1 in Chattopadhyay and De (2016) was given as $E_F[X^4] < \infty$. In light of the above discussions, it is concluded that the moment condition in Theorem 1 of Chattopadhyay and De (2016) can be relaxed and the existence of second moment is sufficient for their theorem to hold.

**Theorem 2.2.** *If the parent distribution $F$ of $X$ is such that $E_F[X^b] < \infty$ for some $b = \max\{4, 4(a-1)\}$ and $a \geq 1$, then the stopping rule in (2.10) yields as $d \downarrow 0$:*

$$E_F\left[N_d/n_d\right] \to 1 \text{ as } d \downarrow 0 \quad [\textit{asymptotic first-order efficiency}],$$

*where $n_d$ comes from (2.5).*

*Proof*: Note that applying part (i) of Theorem 2.1 and the Lebesgue dominated convergence theorem, it suffices to show that $E_F[\sup_{d>0} N_d/n_d] < \infty$.

From the stopping rule in (2.10), we obtain the following upper bound for $N_d$ w.p. 1:

$$N_d \leq m + \beta(d)\{\overline{X}_{N_d-1}^{2a} V_{N_d-1}^2 + (N_d-1)^{-\gamma}\}. \tag{2.12}$$

Dividing both sides of (2.12) by $n_d$, using $(N_d-1)^{-\gamma} < (m-1)^{-\gamma}$, and then taking supremum over $d > 0$, it remains to show that $E_F[\sup_{d>0} \overline{X}_{N_d-1}^{2a} V_{N_d-1}^2] < \infty$, that is, to show that $E_F[\sup_{n\geq m} \overline{X}_n^{2a} V_n^2] < \infty$. Using Lemma A2 of Chattopadhyay and De (2016), we observe w.p. 1 for all fixed $n \geq m$:

$$\overline{X}_n^{2a} V_n^2 \leq \overline{X}_n^{2a}\left\{\frac{S_n^2}{\overline{X}_n^2} + 4 + \frac{s_{wn}^2}{4\overline{X}_n^2}\right\} = T_{1n} + 4T_{2n} + \frac{1}{4}T_{3n}, \tag{2.13}$$

where we denote $T_{1n} = \overline{X}_n^{2(a-1)} S_n^2$, $T_{2n} = \overline{X}_n^{2a}$, and $T_{3n} = \overline{X}_n^{2(a-1)} S_{wn}^2$.

We need to show that $E_F[\sup_{n\geq m} T_{in}] < \infty$ for $i = 1, 2, 3$. First, using Cauchy-Schwartz inequality, we have:

$$E_F\left[\sup_{n\geq m} T_{1n}\right] \leq \left\{E_F\left[\sup_{n\geq m} \overline{X}_n^{4(a-1)}\right] E_F\left[\sup_{n\geq m} S_n^4\right]\right\}^{1/2}.$$

Since $S_n^2$ and $\overline{X}_n$ are both U-statistics, a direct application of Lemma 9.2.4 from Ghosh et al. (1997) yields:

$$E_F\left[\sup_{n\geq m} \overline{X}_n^{4(a-1)}\right] \leq E_F\left[\sup_{n\geq m} \overline{X}_n^{b}\right] \leq \left(\frac{b}{b-1}\right)^b E_F\left[\overline{X}_m^b\right] < \infty$$

provided $E_F[X^b] < \infty$.

Similarly, we can claim:

$$E_F\left[\sup_{n\geq m} S_n^4\right] \leq 4E_F[S_m^4] < \infty \text{ and } E_F\left[\sup_{n\geq m} T_{2n}\right] \leq \left(\frac{2a}{2a-1}\right)^{2a} E_F\left[\overline{X}_m^{2a}\right] < \infty$$

provided $E_F[X^4]$ and $E_F[X^{2a}]$ are finite respectively. Note that $2a > 1$ as $a \geq 1$.

To deal with the term $E_F[\sup_{n\geq m} T_{3n}]$, we may again use Cauchy-Schwartz inequality, Lemma 9.2.4 of Ghosh et al. (1997), and Sen and Ghosh (1981) to argue that $E_F[\sup_{n\geq m} T_{3n}] < \infty$ provided $E_F[X^b] < \infty$ and $E_F[X^4] < \infty$. Note that

$$\max\{4, 4(a-1), 2a\} = \max\{4, 4(a-1)\} = b \text{ for } a \geq 1.$$

This completes the proof. ■

Theorem 2.2 and the first part of Theorem 2.1 establish that the estimated final sample size $N_d$ is asymptotically close to the unknown optimal sample size $n_d$. In the terminology of Ghosh and Mukhopadhyay (1981), the result highlighted by Theorem 2.2 is well-known in the literature as the first-order asymptotic efficiency property of the stopping rule. The last part of Theorem 2.1 proves that the coverage probability attained by the relative-accuracy confidence set $J_{N_d}$ from (2.11) is asymptotically close to the target $1-\alpha$. This property is customarily referred to as asymptotic consistency of the constructed confidence set (Chow and Robbins 1965). One may gain additional insights from Anscombe (1952,1953) and Ray (1957).

These asymptotic properties hold as the cost $d$ tends to zero. However, in practice, $d$ may not be close to zero, and thus, it becomes essential to assess the performance of the proposed estimation methods for some reasonable choices of $d$ via simulation studies. We present our summary findings in Section 4.

## 3. MINIMUM RELATIVE RISK POINT ESTIMATION (MRRPE)

De and Chattopadhyay (2017) developed a minimum risk point estimation procedure for Gini index under the loss function $A(\widehat{G}_n - G)^2 + cn$. We recall that this is a combination of squared error loss plus linear cost. Motivated by the discussion in Section 1.1, we replace the squared loss $(\widehat{G}_n - G)^2$ by the weighted squared error loss, $\mu^a(\widehat{G}_n - G)^2$, and consider the problem of point estimation of $G$ under the modified loss function

$$L_n(G, \widehat{G}_n) = A\mu^a(\widehat{G}_n - G)^2 + cn,$$

given in (1.8).

Using equation (2.3) from De and Chattopadhyay (2017), with fixed sample size $n$, the associated risk function becomes:

$$R_n(G) \equiv A\mu^a E_F\left[\widehat{G}_n - G\right]^2 + cn = A\mu^a\frac{\xi^2}{n} + cn + \mathcal{O}(n^{-3/2}), \tag{3.1}$$

provided $E_F[X^6]$ is finite. Clearly, $R_n(G)$ represents the expected cost for estimating $G$ by means of $\widehat{G}_n$ based on $n$ observations $X_1, ..., X_n$.

Our objective is to determine the sample size for which the approximate expected cost (ignoring $\mathcal{O}(n^{-3/2})$ term) in (3.1) is minimum. Analogous to the minimum risk point estimation problem, coined by Robbins (1959), investigated further by Starr (1966), and later generalized by Ghosh and Mukhopadhyay (1979) in the nonparametric scenario, we now refer to this present problem as the *minimum relative risk point estimation* (MRRPE) problem. Following the lines of De and Chattopadhyay (2017), the approximate expected cost turns out to be (ignoring $\mathcal{O}(n^{-3/2})$ term):

$$h(n) = A\mu^a\frac{\xi^2}{n} + cn,$$

which is minimized if

$$n \text{ is the smallest integer} \geq \sqrt{\frac{A}{c}}\xi\mu^{a/2} = n_c, \text{ say.} \tag{3.2}$$

We refer to $n_c$ as the optimal (minimal) sample size to minimize the expected cost for point estimation of $G$ and tacitly disregard that $n_c$ may not be an integer. The approximate (asymptotically) minimum risk with $n_c$ samples is

$$R^*_{n_c}(G) = A\mu^a\frac{\xi^2}{n_c} + cn_c = 2cn_c, \tag{3.3}$$

up to $\mathcal{O}(c^{1/2})$ term.

However, the optimal sample size $n_c$ from (3.2) and the minimum risk in (3.3) are unknown up to the parameters $\mu$ and $\xi$. Therefore, we need to collect samples at least in two stages where $\mu$ and $\xi$ are estimated in the first stage based on a pilot sample followed by additional samples drawn in the subsequent stages until some stopping criteria is satisfied.

In this section, we propose a purely sequential procedure along the lines of De and Chattopadhyay (2017). With some predetermined pilot sample size $m$, we initially collect $X_1, \ldots, X_m$ i.i.d. observations from $F$ in the first stage and then collect one observation (independent of the previous ones) at-a-time until the current sample size is more or equal to the estimated optimal sample size from (3.2). Thus, the stopping rule or the final sample size of the procedure is defined as:

$$N_c = \inf\left\{n \geq m : n \geq \sqrt{\frac{A}{c}}\left(\overline{X}_n^{a/2} V_n + n^{-\gamma}\right)\right\}, \tag{3.4}$$

where $\gamma$ is some known positive constant and $V_n^2$ from (2.9) is a strong consistent estimator of $\xi^2$ shown in (2.3). Here, $\gamma$ plays the same role as it did in (2.10). After termination of sampling, based upon the finally accrued data $\{N_c, X_1, \ldots, X_{N_c}\}$, we propose:

$$\text{the point estimator of } G \text{ as } \widehat{G}_{N_c} = \widehat{\Delta}_{N_c}/2\overline{X}_{N_c}, \tag{3.5}$$

in the spirit of (2.11).

**Remark 3.1.** Note from (3.4) that $N_c \geq \sqrt{\frac{A}{c}}N_c^{-\gamma}$ w.p. 1. In other words, the final sample size must be at least $(A/c)^{1/2(1+\gamma)}$. This indicates that one should choose the pilot sample size as $m = \max\{4, \lceil (A/c)^{1/2(1+\gamma)} \rceil\}$ to implement the purely sequential procedure (3.4).

The following lemma, similar to Lemma 2.1, guarantees that sampling will stop at some finite time w. p. 1.

**Lemma 3.1.** *For any $c > 0$, the stopping time $N_c$ is finite, that is, $P_F(N_c < \infty) = 1$ provided $\xi < \infty$.*

*Proof*: Follows directly from Lemma 2 of De and Chattopadhyay (2017). ■

Below, we establish some asymptotic optimality properties of the proposed sequential procedure in the spirits of Theorem 1 from De and Chattopadhyay (2017).

**Theorem 3.1.** *Suppose observations $X_1, X_2, \ldots$ are i.i.d. copies of $X$ having distribution function $F$. The stopping rule in (3.4) yields the following properties as $c \downarrow 0$:*

(*i*) *$N_c/n_c \to 1$ w.p. 1 if $E_F[X^2] < \infty$;*

(*ii*) *For $m \geq 4$, if either (a) the support of $F$ is $(t, \infty)$ for some $t > 0$ and $E_F[X^{\max\{r,a\}}] < \infty$ for some $r > 2$ or (b) $E_F[X^{-s}] < \infty$ for some $s > 4$ and $E_F[X^{\max\{r,a\}}] < \infty$ for some $r > 4$, then the asymptotic first-order efficiency holds, that is, $E_F[N_c/n_c] \to 1$.*

*where $n_c$ comes from (3.2).*

*Proof*: See Appendix. ■

For a given cost $c > 0$, the risk function corresponding to the purely sequential estimation strategy (3.4)-(3.5) $N_c$ becomes:

$$R_{N_c}(G) = A\mu^a E_F\left[\left(\widehat{G}_{N_c} - G\right)^2\right] + cE_F[N_c]. \tag{3.6}$$

This is the final risk for estimating $G$ with $\widehat{G}_{N_c}$ using the proposed sequential procedure. The estimator $\widehat{G}_{N_c}$ is called an *asymptotic minimum relative risk point estimator* (AMRRPE) in the spirit of Sen (1981)), of $G$ if the final risk based on $N_c$ samples is asymptotically equivalent to the risk based on the optimal fixed-sample-size $n_c$, that is, if $\lim_{c\downarrow 0} R_{N_c}(G)/R_{n_c}(G) = 1$. This result is described as the *asymptotic first-order risk efficiency* property. The following theorem shows that the risk efficiency holds under some appropriate moment conditions.

**Theorem 3.2.** *Suppose observations $X_1, X_2, \ldots$ are i.i.d. copies of $X$ having distribution function $F$.* For any $\gamma \in (0, 2)$ *and different values of $a$ used in (3.1), the purely sequential estimation strategy (3.4)-(3.5) satisfies the asymptotic first-order risk efficiency property, that is,*

$$\lim_{c\downarrow 0} R_{N_c}(G)/R_{n_c}(G) = 1,$$

*where $n_c$ comes from (3.2), under the sufficient conditions laid down as follows:*

*When $a = 1$, 2 or 3: either (i) the support of $F$ is $(t, \infty)$ for some $t > 0$ and $E_F[X^6]$ is finite or (ii) $E_F[X^6]$ and $E_F[X^{-s}]$ are finite for some $s > 12$;*

*When $a \in \{4, 5, \ldots\}$: $E_F\left[X^{\max\{6,\, a-2\}}\right]$ is finite.*

*Proof*: See Appendix. ■

Without assuming a special nature of the distribution function $F$ generating the incoming data, Theorem 3.1 ensures that the average sample size of the proposed purely sequential procedure approaches the oracle optimal fixed-sample-size that minimizes the risk function which is a combination of expected squared error loss due estimation and linear cost. Theorem 3.2 establishes that the expected overall risk for estimating the population Gini index by the sample Gini index using the proposed sequential procedure is asymptotically close to the expected risk for estimating the population Gini index using the oracle optimal sample size. These asymptotic properties hold as the cost $c$ per unit observation tends to zero. However, in practice, $c$ may not be close to zero, and thus, it becomes essential to assess the performance of the proposed methods for some reasonable choices of $c$ via simulation studies. We present our summary findings in Section 4.

## 4. SIMULATION STUDIES

In Sections 2 and 3, we established that the proposed purely sequential procedures for relative-accuracy confidence set estimation as well as minimum relative risk point estimation enjoy a number of relevant asymptotic optimality properties as $d$ or $c$ approaches 0 respectively. In this section, we attempt to assess finite sample performances of our proposed estimation methodologies for moderate to small values of $d$ and $c$ via Monte Carlo simulations.

### 4.1. Confidence Set Estimation

Let us first consider the sequential procedure (2.10) for relative-accuracy confidence set estimation of Gini index from Section 2. With some preassigned values of $d$, we implemented the sequential procedure corresponding to the stopping rule (2.10) by choosing the pilot sample size $m = \max\{4, \lceil \beta(d)^{1/(1+\gamma)} \rceil\}$ by fixing $\gamma = 0.2$ (for the Pareto case) and $\gamma = 1$ (for the rest). We evaluated the performances in a number of cases, that is, by drawing independent samples from several distributions, specifically relevant to income or wealth data, namely gamma, log-normal, Pareto and exponential. Tables 1 and 2 correspond to the cases $a = 1$ and $a = 2$ respectively with $a$ given in (1.6) and show summaries in a limited number of scenarios.

**Table 1.** Performance of the purely sequential procedure for relative-accuracy confidence set estimation of $G$ for different distributions with $a = 1$: 10,000 replications

| Distribution | $d$<br>$1-\alpha$ | $\lceil n_d \rceil$ | $\overline{N}_d$<br>$s(\overline{N}_d)$ | $\overline{N}_d/n_d$ | $\widehat{CP}$<br>$s(\widehat{CP})$ | $G$<br>$\widehat{G}_{N_d}$ |
|---|---|---|---|---|---|---|
| Gamma<br>shape = 1.5, rate = 1 | 0.03<br>0.90 | 461 | 463.98<br>0.6133 | 1.0066 | 0.8981<br>0.0030 | 0.4244<br>0.4240 |
| Log-normal<br>meanlog = 0.75, sdlog = 0.25 | 0.025<br>0.95 | 301 | 314.33<br>0.4452 | 1.0466 | 0.9492<br>0.0022 | 0.1403<br>0.1400 |
| Pareto<br>scale = 5, shape = 4.01 | 0.09<br>0.90 | 721 | 693.92<br>2.6246 | 0.9635 | 0.8817<br>0.0032 | 0.1424<br>0.1401 |
| Exponential<br>rate = 0.4 | 0.05<br>0.95 | 801 | 799.7<br>0.9167 | 0.9992 | 0.9513<br>0.0021 | 0.5000<br>0.4998 |

These tables present the estimated expected sample size $\overline{N}_d$ (estimator of $E_F[N_d]$), its standard error $s(\overline{N}_d)$, the estimated coverage probability $\widehat{CP}$ (estimator of the probability in part (ii) of Theorem 2.1), its standard error $s(\widehat{CP})$, and the final estimator of Gini index $\widehat{G}_{N_d}$ based on 10,000 replications. For different values of $d$, $\alpha$ and the given parameters of respective distributions, we exhibit the values of population Gini index $G$ and the optimal oracle sample size $\lceil n_d \rceil$. Fifth columns in Tables 1 and 2 illustrate that the ratios of average sample sizes and optimal sample sizes are nearly 1 under all scenarios. These validate the asymptotic first-order efficiency property of $N_d$ given in Theorem 2.2.

The standard errors of the estimator $\overline{N}_d$ are small in all scenarios. The estimated coverage probabilities in column six of Tables 1 and 2 are very close to the target level $1 - \alpha$ for all chosen distributions, which validates the asymptotic consistency property of the sequential procedure. In the Pareto case, we observe that the expected final sample sizes are slightly less, on an average, than the corresponding optimal sample sizes leading to little loss in the attained coverage probabilities. This may be due to the fact that for the given shape and scale parameters of Pareto distributions, the values of $\xi^2$ and its estimator $V_n^2$ are close to zero leading to early stopping of the purely sequential procedure. One way to deal with this problem may be to choose a smaller value of $\gamma$ to prevent undersampling. However, if the chosen value of $\gamma$ is too small, it may also lead to oversampling. The question of optimal selection of $\gamma$ for a given scenario, or in general, is an open question and is beyond the scope of this article. The last columns in Tables 1 and 2 illustrate that the final estimator $\widehat{G}_{N_d}$ is very accurate in estimating the population Gini index, $G$. Overall, the simulation results validate the theoretical properties and the finite-sample-size performances of the purely sequential procedure of section 2 are clearly very encouraging.

**Table 2.** Performance of the purely sequential procedure for relative-accuracy confidence set estimation of $G$ for different distributions with $a = 2$: $10,000$ replications

| Distribution | $d$<br>$1-\alpha$ | $\lceil n_d \rceil$ | $\overline{N}_d$<br>$s(\overline{N}_d)$ | $\overline{N}_d/n_d$ | $\widehat{CP}$<br>$s(\widehat{CP})$ | $G$<br>$\widehat{G}_{N_d}$ |
|---|---|---|---|---|---|---|
| Gamma<br>shape = 2, rate = 1.5 | 0.04<br>0.90 | 305 | 304.25<br>0.6249 | 0.9987 | 0.8929<br>0.0031 | 0.3750<br>0.3744 |
| Log-normal<br>meanlog = 0.6, sdlog = 0.3 | 0.03<br>0.90 | 587 | 582.86<br>0.8005 | 0.9932 | 0.8952<br>0.0031 | 0.1680<br>0.1676 |
| Pareto<br>scale = 2.25, shape = 4.01 | 0.09<br>0.90 | 1311 | 1268.54<br>4.2764 | 0.9680 | 0.8805<br>0.0032 | 0.1424<br>0.1409 |
| Exponential<br>rate = 0.5 | 0.07<br>0.95 | 1046 | 1036.86<br>1.5595 | 0.9919 | 0.9439<br>0.0023 | 0.5000<br>0.4998 |

### 4.2. Point Estimation

We implemented the purely sequential procedure corresponding to (3.4) for minimum relative risk point estimation with $A = \$500000$, $c = \$0.5$, $\gamma = 1$ and pilot sample size $m = \max\{4, \lceil (A/c)^{1/2(1+\gamma)} \rceil\}$. The performances of the sequential procedure were evaluated for small to moderate sample sizes by drawing random samples from four different income distributions, namely exponential, gamma, log-normal and Pareto. Table 3 and Table 4 correspond to $a = 1$ and $a = 2$ respectively where $a$ is given in (1.8).

**Table 3.** Performance of the purely sequential procedure for minimum relative risk point estimation of $G$ with $a = 1$, $A = \$500000$, $\gamma = 1$ and $c = \$0.5$: $10,000$ replications

| Distribution | $G$<br>$\widehat{G}_{N_c}$ | $\lceil n_c \rceil$ | $\overline{N}_c$<br>$s(\overline{N}_c)$ | $\overline{N}_c/n_c$ | $\overline{R}_{N_c}$<br>$s(\overline{R}_{N_c})$ | $\frac{\overline{R}_{N_c}}{R^*_{n_c}}$ |
|---|---|---|---|---|---|---|
| Exponential<br>rate = 0.5, $t = 0.001$ | 0.4998<br>0.4996 | 409 | 411.12<br>0.2665 | 1.0073 | 408<br>0.2680 | 0.9996 |
| Gamma<br>shape = 7.5, rate = 1.5, $t = 0.001$ | 0.2026<br>0.2023 | 314 | 315.82<br>0.1986 | 1.0075 | 312.01<br>0.2006 | 0.9953 |
| Log-normal<br>meanlog = 2, sdlog = 0.25 | 0.1403<br>0.1401 | 280 | 280.92<br>0.2271 | 1.005 | 276.71<br>0.2299 | 0.9902 |
| Pareto<br>scale = 45, shape = 6.01 | 0.0907<br>0.0903 | 913 | 884.56<br>1.2633 | 0.9692 | 882.69<br>1.2650 | 0.9672 |

Since negative moments do not exist for exponential and gamma distributions under consideration, we assumed truncated support in such cases, that is, assumed that the data came from truncated exponential or truncated gamma distributions having support $(t, \infty)$ with $t = 0.001$. In the cases of log-normal and Pareto, since all negative moments exist, we assumed full support $(0, \infty)$. Tables 3 and 4 summarize the estimated expected sample size $\overline{N}_c$ (estimates $E_F[N_c]$), its standard error $s(\overline{N}_c)$, the average risk $\overline{R}_{N_c}$ (estimates $R_{N_c}(G)$), its standard error $s(\overline{R}_{N_c})$, and the final estimator $\widehat{G}_{N_c}$ for the Gini index $G$ from $10,000$ replications.

**Table 4.** Performance of the purely sequential procedure for minimum relative risk point estimation of $G$ with $a = 2$, $A = \$500000$, $\gamma = 1$ and $c = \$0.5$: $10,000$ replications

| Distribution | $G$<br>$\widehat{G}_{N_c}$ | $\lceil n_c \rceil$ | $\overline{N}_c$<br>$s(\overline{N}_c)$ | $\overline{N}_c/n_c$ | $\overline{R}_{N_c}$<br>$s(\overline{R}_{N_c})$ | $\frac{\overline{R}_{N_c}}{R^*_{n_c}}$ |
|---|---|---|---|---|---|---|
| Exponential<br>rate = 1.5, $t = 0.001$ | 0.4992<br>0.4982 | 61 | 66.70<br>0.0979 | 1.0959 | 613.31<br>1.0388 | 1.0077 |
| Gamma<br>shape = 0.9, rate = 1.1, $t = 0.001$ | 0.5194<br>0.5182 | 77 | 80.99<br>0.1195 | 1.0618 | 763.63<br>1.2478 | 1.0012 |
| Log-normal<br>meanlog = 1.55, sdlog = 0.35 | 0.1955<br>0.1949 | 226 | 223.19<br>0.2445 | 0.9912 | 2211.01<br>2.4597 | 0.9819 |
| Pareto<br>scale = 16, shape = 6.01 | 0.0907<br>0.0902 | 754 | 726.37<br>1.1460 | 0.9635 | 7252.22<br>11.46651 | 0.9619 |

For the given distributions, we also provide the values of population Gini index $G$ and the optimal sample size $\lceil n_c \rceil$ that minimized the expected cost. The ratios of the estimated expected sample sizes and optimal sample sizes under given scenarios are nearly 1 validating the assertion of asymptotic first-order efficiency property stated in Theorem 3.1. The last columns of Tables 3 and 4 show that, on the average, the overall risk of estimating $G$ using the final sample size $N_c$ and the accrued data is approximately equal to the minimum possible risk $R^*_{n_c}(G)$ given in (3.3). This validates the asymptotic first-order risk efficiency property stated in Theorem 3.2. Moreover, we observed that the final estimator $\widehat{G}_{N_c}$ accurately estimated the population Gini index $G$ funder all four scenarios. Based on the large set of our broad ranging simulations, we feel comfortable to conclude that the finite-sample-size performances of the proposed purely sequential procedure of (3.4) are clearly very encouraging.

## 5. APPENDIX WITH SELECTED TECHNICALITIES

This section presents a number of auxiliary results which are needed to prove Theorems 3.1 and 3.2.

**Lemma 5.1.** *Let $\overline{X}_n$ be the sample mean based on non-negative i.i.d. observations $X_1, \ldots, X_n$ following $F$. For any $r \geq 1$, $E_F\left[\sup_{n \geq m} \overline{X}_n^{-r}\right] < \infty$ provided $E_F(X^{-s}) < \infty$ for some $s > r$.*

*Proof*: For some $\beta > 1$, we have:

$$E_F\left[\max_{n \geq m} \frac{1}{\overline{X}_n^r}\right] \leq 1 + \int_1^\infty P_F\left(\max_{n \geq m} \frac{1}{\overline{X}_n^r} \geq t\right) dt \leq 1 + \frac{E_F\left[\overline{X}_m^{-r\beta}\right]}{\beta - 1}. \tag{5.1}$$

The last upper bound in (5.1) is obtained by applying maximal inequality for reverse submartingales (Lee 1990). Let $s = r\beta$. Now, it is enough to show that if $E_F[X^{-s}] < \infty$, then $E_F[\overline{X}_n^{-s}] < \infty$. Note that $\overline{X}_n \geq \left(\prod_{i=1}^n X_i\right)^{1/n}$ as the observations are nonnegative and we may write:

$$E_F\left(\overline{X}_n^{-s}\right) \leq E_F\left[\left(\prod_{i=1}^n \frac{1}{X_i}\right)^{s/n}\right] = \left\{E_F\left[\left(\frac{1}{X_1}\right)^{s/n}\right]\right\}^n. \tag{5.2}$$

The last upper bound in (5.2) is due to the i.i.d. property of the observations.

We know that $\{E_F[|X|^p]\}^{1/p}$ is a nondecreasing function of $p$ for $p > 0$. Applying this result with $p = 1/n \leq 1$ in (5.2), we complete the proof of Lemma 5.1. ■

**Lemma 5.2.** *If the parent distribution $F$ of $X$ is such that either $(i)$ support of $F$ is $(t, \infty)$ for some $t > 0$ and $E_F[X^r] < \infty$ for some $r > 2$ or $(ii)$ $E_F[X^s]$ and $E_F[X^{-s}]$ exist for some $s > 4$, then $E_F\left[\sup_{n\geq m} V_n^2\right] < \infty$ for $m \geq 4$.*

*Proof*: The first condition (i) is already proven to be sufficient for $E_F\left[\sup_{n\geq m} V_n^2\right] < \infty$ in Lemma A3 of Chattopadhyay and De (2016). For (ii), we note that $V_n^2 = \max\{0, V_n^*\} \leq S_n^2\overline{X}_n^{-2} + 4 + \frac{1}{4}s_{wn}^2\overline{X}_n^{-2}$, and thus, it is enough to show that $E_F[\sup_{n\geq m} S_n^2\overline{X}_n^{-2}]$ and $E_F[\sup_{n\geq m} s_{wn}^2\overline{X}_n^{-2}]$ are finite. Now, Holder's inequality yields: $E_F[\sup_{n\geq m} S_n^2\overline{X}_n^{-2}]$ and $E_F[\sup_{n\geq m} s_{wn}^2\overline{X}_n^{-2}]$ are bounded from above by:

$$\left\{E_F\left[\sup_{n\geq m} S_n^4\right] E_F\left[\sup_{n\geq m} \overline{X}_n^{-4}\right]\right\}^{1/2} \text{ and } \left\{E_F\left[\sup_{n\geq m} s_{wn}^4\right] E_F\left[\sup_{n\geq m} \overline{X}_n^{-4}\right]\right\}^{1/2} \text{ respectively.}$$

By Lemma 9.2.4 from Ghosh et al. (1997), the term involving $S_n$ is finite provided $E_F[X^4] < \infty$ and by Lemma 5.1, the term involving $\overline{X}_n$ is finite provided $E_F[X^{-s}] < \infty$ for some $s > 4$. Then, following Sen and Ghosh (1981), we have: $E_F\left[\sup_{n\geq m} s_{wn}^4\right]$ is finite if $E_F[X^s]$ is finite for some $s > 4$ and $m \geq 4$. This completes the proof of lemma 5.2. ■

### 5.1. Proof of Theorem 3.1

A proof of part (i) requires strong consistency property of both $\overline{X}_n$ and $V_n^2$ which demands existence of the second moment as discussed in Remark 2.3. The rest of the proof of part (i) follows from Theorem 1 of De and Chattopadhyay (2017).

In order to prove part (ii), note from the definition of stopping rule (3.4) that we can express (w.p.1):

$$N_c/n_c \leq 1 + \left\{\sup_{n\geq m} \overline{X}_n^{a/2} V_n + (m-1)^{-\gamma}\right\} \mu^{-a/2}\xi^{-1} \text{ for all } c > 0,$$

and thus, it is enough to show that $E_F\left[\sup_{n\geq m} \overline{X}_n^{a/2} V_n\right] < \infty$. A convenient upper bound of $E_F\left[\sup_{n\geq m} \overline{X}_n^{a/2} V_n\right]$ is obtained via Holder's inequality which leads to

$$\left\{E_F\left[\sup_{n\geq m} \overline{X}_n^{a}\right] E_F\left[\sup_{n\geq m} V_n^2\right]\right\}^{1/2},$$

where $E_F\left[\sup_{n\geq m} V_n^2\right]$ is finite by Lemma 5.2, $E_F\left[\sup_{n\geq m} \overline{X}_n^{a}\right]$ is finite by Lemma 9.2.4 of Ghosh et al. (1997), and the sufficient moment conditions of Theorem 3.1. ■

### 5.2. Proof of Theorem 3.2

In the case when $a = 1$, 2, or 3, the condition (i) was already proved to be sufficient for the first-order risk efficiency property as shown in Theorem 1 in De and Chattopadhyay (2017). We only need to verify whether the condition (ii), specifically, the negative moment condition is sufficient for the first-order risk efficiency property. To this end, we need to establish some auxiliary results similar to the lemmas in the Appendix of De and Chattopadhyay (2017).

For some fixed $\gamma > 0$, $c > 0$ and $\epsilon > 0$, define:

$$n_{1c} = \left(\frac{A}{c}\right)^{1/2(1+\gamma)}, \; n_{2c} = (1-\epsilon)n_c, \text{ and } n_{3c} = (1+\epsilon)n_c,$$

where $n_c$ is the optimal sample size given in (3.2). Clearly, $N_c \geq n_{1c}$ w.p.1 as discussed in Remark 3.1. Let us first restate Lemma 4 from De and Chattopadhyay (2017) with appropriate modifications and negative moment conditions. The remainder of this proof will be indicated later.

**Lemma 5.3.** *Suppose that observations $X_1, X_2, \dots$ are i.i.d. copies of $X$ having a common distribution function $F$. Let $s$ be some positive integer and $k$ be some positive constant.*

*(I) If either (i) the support of $F$ is $(t,\infty)$ for some $t>0$ and $E_F[X^{2p}]<\infty$ for some $p>1$ or (ii) $E_F[X^{2p}]<\infty$ and $E_F[X^{-4ps}]<\infty$ for some $p>1$, then*

$$P_F\left(\max_{n_{1c}\le n\le n_{2c}} \left|\overline{X}_n^{-s}-\mu^{-s}\right| \ge k\right) \le \mathcal{O}\left(n_{1c}^{-p}\right) \text{ as } c\downarrow 0 \ .$$

*(II) If $E_F[X^r]<\infty$ for some $r\ge \max\{2,s\}$, then*

$$P_F\left(\max_{n_{1c}\le n\le n_{2c}} \left|\overline{X}_n^{s}-\mu^{s}\right| \ge k\right) \le \mathcal{O}\left(n_{1c}^{-r/2}\right) \text{ as } c\downarrow 0.$$

*Proof*: Part (I) follows from Lemma 4 of De and Chattopadhyay (2017) by applying Cauchy-Schwartz's inequality and Lemma 5.1 where appropriate. For part (II), note that the binomial expansion of $\overline{X}_n^s = \mu^s\left(1+(\overline{X}_n-\mu)/\mu\right)^s$ yields

$$|\overline{X}_n^s-\mu^s| \le s\mu^{s-1}|\overline{X}_n-\mu| + \binom{s}{2}\mu^{s-2}|\overline{X}_n-\mu|^2+\cdots+|\overline{X}_n-\mu|^s.$$

Applying maximal inequality for reverse submartingales and Lemma 2.2 of Sen and Ghosh (1981), we get

$$P_F\left(\max_{n_{1c}\le n\le n_{2c}} \left|\overline{X}_n^{s}-\mu^{s}\right| \ge k\right) \le \textstyle\sum_{i=1}^s \frac{E_F[|\overline{X}_{n_{1c}}-\mu|^r]}{k_i} = \mathcal{O}\left(n_{1c}^{-r/2}\right),$$

which completes its proof. ■

Now, let us restate Lemma 5 from De and Chattopadhyay (2017) with appropriate negative moment conditions.

**Lemma 5.4.** *Suppose that the observations $X_1, X_2, \dots$ are i.i.d. copies of $X$ having a common distribution function $F$. For different values of $a$ in (1.8), the stopping rule (3.4) yields the following asymptotic orders:*

*(I) Let $a=1,2$ or $3$. For any $\epsilon\in(0,1)$ and $\gamma>0$, we have*

$$P_F(N_c\le n_{2c})=\mathcal{O}(n_{1c}^{-r})=\mathcal{O}\left(c^{r/2(1+\gamma)}\right) \text{ and } P_F(N_c>n_{3c})=\mathcal{O}(n_{1c}^{-r})=\mathcal{O}\left(c^{r/2(1+\gamma)}\right) \text{ as } c\downarrow 0,$$

*provided either (i) the support of $F$ is $(t,\infty)$ for some $t>0$ and $E_F[X^{4r}]<\infty$ for some $r\ge 1$ or (ii) $E_F[X^{4r}]<\infty$ and $E_F\left[X^{-4\alpha r(4-a)}\right]<\infty$ for some $r\ge 1$ and $\alpha>1$.*

*(II) Let $a\in\{4,5,\dots\}$. If $E_F\left[X^{\max\{4r,\,a-2\}}\right]<\infty$ for some $r\ge 1$, then for any $\epsilon\in(0,1)$ and $\gamma>0$*

$$P_F(N_c\le n_{2c})=\mathcal{O}(n_{1c}^{-r})=\mathcal{O}\left(c^{r/2(1+\gamma)}\right) \text{ and } P_F(N_c>n_{3c})=\mathcal{O}(n_{1c}^{-r})=\mathcal{O}\left(c^{r/2(1+\gamma)}\right) \text{ as } c\downarrow 0.$$

*Proof*: For a proof of part (I), we only need to verify the condition (ii) as the condition (i) is already verified in Lemma 5 of De and Chattopadhyay (2017). Following the definition of stopping rule (3.4) and proceeding along the lines of (A.9) in De and Chattopadhyay (2017), we have:

$$\begin{aligned} P_F(N_c\le n_{2c}) &\le P_F(|\overline{X}_n^a V_n^* - \mu^a\xi^2| \ge \mu^a\xi^2\epsilon(2-\epsilon) \text{ for some } n\in[n_{1c},n_{2c}]) \\ &\le P_F\left(\max_{n_{1c}\le n\le n_{2c}} \{|V_{1n}|+|V_{2n}|+|V_{3n}|+|V_{4n}|\}\ge k\right), \end{aligned} \tag{5.3}$$

where

$$k=\mu^a\xi^2\epsilon(2-\epsilon),\ V_{1n}=\left(\frac{\widehat{\Delta}_n^2 S_n^2}{4\overline{X}_n^{4-a}}-\frac{\Delta^2\sigma^2}{4\mu^{4-a}}\right),\ V_{2n}=\left(\frac{\widehat{\Delta}_n\widehat{\tau}_n}{\overline{X}_n^{3-a}}-\frac{\Delta\tau}{\mu^{3-a}}\right),$$
$$V_{3n}=\left(\frac{\widehat{\Delta}_n^2}{\overline{X}_n^{2-a}}-\frac{\Delta}{\mu^{2-a}}\right), \text{ and } V_{4n}=\left(\frac{s_{wn}^2}{4\overline{X}_n^{2-a}}-\frac{\sigma_1^2}{\mu^{2-a}}\right).$$

Then (5.3) can be rewritten as

$$P_F(N_c \le n_{2c}) \le P_1 + P_2 + P_3 + P_4,$$

where

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

Let us first focus on $P_1$. Following equation (A.10) of De and Chattopadhyay (2017), $V_{1n}$ can be rewritten as:

$$\begin{array}{c} V_{1n} = T_{1n}T_{2n}T_{3n} + \Delta^2 T_{2n}T_{3n} + \sigma^2 T_{1n}T_{3n} + \frac{T_{1n}T_{2n}}{4\mu^{4-a}} \\ + \frac{\sigma^2 T_{1n}}{4\mu^{4-a}} + \frac{\Delta^2 T_{2n}}{4\mu^{4-a}} + \Delta^2\sigma^2 T_{3n}, \end{array} \tag{5.4}$$

where

$$T_{1n} = (\widehat{\Delta}_n^2 - \Delta^2),\ T_{2n} = (S_n^2 - \sigma^2) \text{ and } T_{3n} = \left(\frac{1}{4\overline{X}_n^{4-a}} - \frac{1}{4\mu^{4-a}}\right).$$

Let us consider the first term in (5.4) and bound it from above with

$$P_F\left(\max_{n_{1c}\le n\le n_{2c}} |T_{1n}T_{2n}T_{3n}| \ge k/28\right) \le \sum_{i=1}^{3} P_i' \ \text{ where } \ P_i' = P_F\left(\max_{n_{1c}\le n\le n_{2c}} |T_{in}| \ge (k/28)^{1/3}\right),$$

for $i = 1,2,3$.

Now, note that $P_1' = \mathcal{O}(n_{1c}^{-r})$ using the finiteness of $E_F[X^{4r}]$ in Lemma 3 of De and Chattopadhyay (2017), $P_2' = \mathcal{O}(n_{1c}^{-r})$ using the the maximal inequality for reverse martingales (Lee 1990), Lemma 2.2 of Sen and Ghosh (1981) and the finiteness of $E_F[X^{4r}]$. Finally, $P_3' = \mathcal{O}(n_{1c}^{-\alpha r})$ for some $\alpha > 1$ using finiteness of $E_F[X^{2\alpha r}]$ and $E_F[X^{-4\alpha r(4-a)}]$, that is, applying $p = \alpha r$ and $s = (4-a)$ in part (I) of Lemma 5.3. Therefore, we have $P_F\left(\max_{n_{1c}\le n\le n_{2c}} |T_{1n}T_{2n}T_{3n}| \ge k/28\right) \le \mathcal{O}(n_{1c}^{-r})$.

Following similar arguments, we can show that the asymptotic orders of probabilities of large deviations corresponding to the remaining six terms in (5.4) are either $\mathcal{O}(n_{1c}^{-r})$ or higher (that is, converges faster to 0 as $c$ tends to 0), and thus, $P_1 \le \mathcal{O}(n_{1c}^{-r})$. Note that $V_{1n}$ involves $\overline{X}_n^{-(4-a)}$ whereas $V_{2n}$, $V_{3n}$ and $V_{4n}$ involve terms $\overline{X}_n^{-(3-a)}$, $\overline{X}_n^{-(2-a)}$ and $\overline{X}_n^{-(2-a)}$ respectively.

When $a = 1,2,3$, applying part (I) of Lemma 5.3, the negative moment conditions needed for handling $V_{1n}$ would be sufficient to deal with $V_{2n}$, $V_{3n}$ and $V_{4n}$. Thus, the moment conditions in (ii) will be sufficient to establish the asymptotic order of $P_F(N_c \le n_{2c})$ and the remaining proof of part (I) of Lemma 5.4 will follow from the proof of Lemma 5 from De and Chattopadhyay (2017). The proof related to the asymptotic order of $P_F(N_c > n_{3c})$ is also similar, and hence omitted.

To prove part (II), that is, in the case when $a \in \{4,5,\dots\}$, we will neither need the truncated support assumption nor any negative moment assumption. Interestingly, note from the proof of part (I) that we separately obtained the asymptotic orders of the probabilities of large deviations of the following 8 terms, namely $Q_{1n} = |\widehat{\Delta}_n^2 - \Delta^2|$, $Q_{2n} = |S_n^2 - \sigma^2|$, $Q_{3n} = |\widehat{\tau}_n - \tau|$, $Q_{4n} = |\widehat{\Delta}_n - \Delta|$, $Q_{5n} = |s_{wn}^2/4 - \sigma_1^2|$, $Q_{6n} = |\overline{X}_n^{a-4} - \mu^{a-4}|$, $Q_{7n} = |\overline{X}_n^{a-3} - \mu^{a-3}|$, and $Q_{8n} = |\overline{X}_n^{a-2} - \mu^{a-2}|$. The minimum asymptotic order among these 8 terms became the asymptotic orders of both $P_F(N_c \le n_{2c})$ and $P_F(N_c > n_{3c})$.

For some $r \ge 1$, a sufficient condition to obtain the asymptotic order $\mathcal{O}(n_{1c}^{-r})$ corresponding to the first 5 terms $Q_{in}$, for $i = 1,\dots,5$, is the finiteness of $E_F[X^{4r}]$. This result is due to Lemma 3 in De and Chattopadhyay (2017), maximal inequality for reverse submartingales, Lemmas 2.2 and 3.1 from Sen and Ghosh (1981). In the case when $a \in \{4,5,\dots\}$, a sufficient positive moment condition will be determined by the last term $Q_{8n}$.

Applying part (II) of Lemma 5.3, we observe:

$$P_F\left(\max_{n_{1c}\le n\le n_{2c}} Q_{8n} \ge k\right) = \mathcal{O}\left(n_{1c}^{-\max\{r,(a-2)/2\}}\right) \text{ provided } E_F\left[X^{\max\{2r,a-2\}}\right] < \infty.$$

Clearly, to secure the minimum asymptotic order $\mathcal{O}(n_{1c}^{-r})$ among all these orders, a sufficient requirement would be the existence of $\max\{4r, 2r, a-2\}$ moments reducing to the existence of $\max\{4r, a-2\}$ moments. This completes the proof of Lemma 5.4. ■

Next, we restate Lemma 6, Lemma 8, and Lemma 9 from De and Chattopadhyay (2017) with appropriate modifications and negative moment conditions.

**Lemma 5.5.** *Suppose that the observations $X_1, X_2, \ldots$ are i.i.d. copies of $X$ having a common distribution function $F$ such that either $(i)$ support of $F$ is $(t, \infty)$ for some $t > 0$ and $E_F[X^{2r}] < \infty$ for some $r \geq 1$ or $(ii)$ $E_F[X^{2r}]$ and $E_F(X^{-s})$ are finite for some $s > 4r$ and $r \geq 1$. Then, we have:*

$$E_F\left[\max_{n_{1c}\leq n\leq n_{2c}}\left|\widehat{G}_n - G\right|^{2r}\right] = \mathcal{O}(n_{1c}^{-r}) \text{ as } c \downarrow 0.$$

*Proof*: The proof will proceed along the lines of Lemma 6 in De and Chattopadhyay (2017) with additional application of Holder's inequality and Lemma 5.1. We leave out the rest. ■

**Lemma 5.6.** *Let $U_n$ be a U-statistic for estimating some parameter $\theta$ based on $n$ i.i.d.* copies *$X$ having a common distribution function $F$ with kernel $\phi$ such that $E_F[|\phi|^6] < \infty$. Then, for any $\epsilon \in (0, 1)$, we have*:

$$E_F\left[\max_{n_{2c}\leq n\leq n_c}|U_n - U_{n_c}|^6\right] = \mathcal{O}\left(\epsilon n_c^{-3}\right) \text{ as } c \downarrow 0.$$

*Proof*: Its proof is similar to the proof of Lemma 8 in De and Chattopadhyay (2017) and hence omitted. ■

**Lemma 5.7.** *Suppose that the observations $X_1, X_2, \ldots$ are i.i.d. copies of $X$ having a common distribution function $F$ such that either $(i)$ support of $F$ is $(t, \infty)$ for some $t > 0$ and $E_F[X^4] < \infty$ or $(ii)$ $E_F[X^6]$ and $E_F(X^{-s})$ are finite for some $s > 12$. Then for any $\epsilon \in (0, 1)$, we have*:

$$E_F\left[\max_{n_{2c}\leq n\leq n_{3c}}(\widehat{G}_n - \widehat{G}_{n_c})^2\right] = \mathcal{O}\left(\epsilon^{1/3}n_c^{-1}\right) \text{ as } c \downarrow 0.$$

*Proof*: Consider the term $E_{11} = E_F\left[\max_{n_{2c}\leq n\leq n_c}(\overline{X}_n^{-1} - \overline{X}_{n_c}^{-1})^2\widehat{\Delta}_n^2\right]$ given in Lemma 9 of De and Chattopadhyay (2017) and bound it from above by using the following general form of Holder's inequality:

$$E_{11} \leq \left\{E_F\left[\max_{n_{2c}\leq n\leq n_c}\overline{X}_n^{-6}\overline{X}_{n_c}^{-6}\right]E_F\left[\max_{n_{2c}\leq n\leq n_c}(\overline{X}_n - \overline{X}_{n_c})^6\right]E_F\left[\max_{n_{2c}\leq n\leq n_c}\widehat{\Delta}_n^6\right]\right\}^{1/3}.$$

Applying Cauchy-Schwartz's inequality and Lemma 5.1, $E_F\left[\max_{n_{2c}\leq n\leq n_c}\overline{X}_n^{-6}\overline{X}_{n_c}^{-6}\right]$ is finite provided $E_F[X^{-s}]$ is finite for some $s > 12$. The second term in (5.5) is $\mathcal{O}\left(\epsilon^{1/3}n_c^{-1}\right)$ from Lemma 5.6 whereas the last term in (5.5) is finite due to Lemma 9.2.4 in Ghosh et al. (1997). Thus, $E_{11} \leq \mathcal{O}\left(\epsilon^{1/3}n_c^{-1}\right)$. The rest of the proof follows from Lemma 9 of De and Chattopadhyay (2017) by proceeding along the lines as above and bounding other terms $E_{12}$ and $E_2$ analogously. ■

**Proof of Theorem 3.2 Continued**

We need to show

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

Using part (ii) of Theorem 3.1, It is enough to show that $\lim_{c\downarrow 0}\{(A\mu^a)/(cn_c)\}E_F\left[(\widehat{G}_{N_c} - G)^2\right] = 1$, that is, to verify: $\lim_{c\downarrow 0} n_c E_F\left[(\widehat{G}_{N_c} - G)^2\right] = \xi^2$. This leads to equation (A.25) of De and Chattopadhyay (2017). For $a = 1, 2, 3$, the rest can be proved by proceeding along the lines of the proof of Theorem 1 (part (iii)) of De and Chattopadhyay (2017) and applying Lemmas 5.4, 5.5, and 5.7 as appropriate. ■

**REFERENCES**


Anscombe, F. J. (1952). Large-Sample Theory of Sequential Estimation, *Proceedings of Cambridge Philosophical Society* 48: 600-607.

Anscombe, F. J. (1953). Sequential Estimation, *Journal of Royal Statistical Society, Series B* 15: 1-29.

Arnold, B. C. (2005). Inequality Measures for Multivariate Distributions, *Metron* 63: 317–327.

Beach, C. M. and Davidson, R. (1983). Distribution-Free Statistical Inference with Lorenz Curves and Income Shares, 50: 723–735.

Chattopadhyay, B. and De, S. K. (2016). Estimation of Gini Index Within Pre-specified Error Bound, 4:30, DOI 10.3390/econometrics4030030.

Chow, Y. S. and Robbins, H. (1965). On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean, *Annals of Mathematical Statistics* 36: 457-462.

Davidson, R. (2009). Reliable Inference for the Gini Index, *Journal of Econometrics* 150: 30–40.

Davidson, R. and Duclos, J. Y. (2000). Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality, 68: 1435–1464.

De, S. K. and Chattopadhyay, B. (2017). Minimum Risk Point Estimation of Gini Index, *Sankhyā, Series B* 79: 247–277.

Gastwirth, J. L. (1972). The Estimation of the Lorenz Curve and Gini Index, *Review of Economics and Statistics* 54: 306–316.

Ghosh, M. and Mukhopadhyay, N. (1979). Sequential Point Estimation of the Mean When the Distribution Is Unspecified, *Communications in Statistics - Theory & Methods* 8: 637–652.

Ghosh, M. and Mukhopadhyay, N. (1981). Consistency and Asymptotic Efficiency of Two-Stage and Sequential Procedures, *Sankhyā, Series A* 43: 220–227.

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*, New York: Wiley.

Gini, C. (1914). *L'Ammontare la Composizione della Ricchezza delle Nazioni*, Boca, Torino.

Gini, C. (1921). Measurement of Inequality of Incomes, *Economic Journal* 31: 124-126.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution, *Annals of Mathematical Statistics* 19: 293–325.

Hoeffding, W. (1961). The Strong Law of Large Numbers for U-Statistics, Institute of Statistics Mimeo Series #302, University of North Carolina, Chapel Hill.

Langel, M. and Tille, Y. (2013). Variance Estimation of the Gini Index: Revisiting a Result Several Times Published, *Journal of Royal Statistical Society, Series* A 176: 521–540.

Lee, A. J. (1990). *U-statistics: Theory and Practice*, New York: Dekker.

Mukhopadhyay, N. and Chattopadhyay, B. (2013a).Asymptotic Expansion of Percentiles for a Sample Mean Standardized by GMD: An Application in the Normal Case, *Journal of Japan Statistical Society* 42: 165-184.

Mukhopadhyay, N. and Chattopadhyay, B. (2013b).Two-Stage Fixed-Width Confidence Interval Procedures for the Mean of a Normal Distribution, *Sequential Analysis* 32: 134-157.

Mukhopadhyay, N. and Chattopadhyay, B. (2018). Introducing Informal Inequality Measures (IIMs) Constructed from U-statistics of Degree Three or Higher in Analyzing Economic Disparity, in *Contemporary Issues on Globalization & Sustainable Development*, P. P. Sengupta, ed., pp. 1-39, New Delhi: Serial Publications.

Mukhopadhyay, N. and de Silva, B. M. (2009). *Sequential Methods and Their Applications*, Boca Raton: CRC.

Mukhopadhyay, N., and Hu, J. (2017). Confidence Intervals and Point Estimators for a Normal Mean Under Purely Sequential Strategies Involving Gini's Mean Difference and Mean Absolute Deviation, *Sequential Analysis* 36: 210-239.

Mukhopadhyay, N., and Hu, J. (2018). Gini's Mean Difference and Mean Absolute Deviation Based Two-Stage Estimation for a Normal Mean with Known Lower Bound of Variance, *Sequential Analysis*, 37: in press.

Mukhopadhyay, N. and Solanky, T. K. S. (1994). *Multistage Selection and Ranking Procedures: Second-Order Asymptotics*, New York: Dekker.

Ray, W. D. (1957). Sequential Confidence Intervals for the Mean of a Normal Population with Unknown Variance, *Journal of Royal Statistical Society, Series B* 19: 133-143.

Robbins, H. (1959). Sequential Estimation of the Mean of a Normal Population, in *Probability and Statistics*, H. Cramér volume, U. Grenander, ed., pp. 235–245, Uppsala: Almquist and Wiksell.

Sen, P. K. (1981). *Sequential Nonparametrics*, New York: Wiley.

Sen, P. K. and Ghosh, M. (1981). Sequential Point Estimation of Estimable Parameters Based on U-statistics, *Sankhyā, Series A* 43: 331–344.

Sproule, R. (1969). *A Sequential Fixed-Width Confidence Interval for the Mean of a U-statistic*, *Ph. D. dissertation*, University of North Carolina, Chapel Hill.

Sproule, R. (1985). Sequential Nonparametric Fixed-Width Confidence Intervals for U-statistics, *Annals of Statistics* 13: 228–235.

Starr, N. (1966). On the Asymptotic Efficiency of a Sequential Procedure for Estimating the Mean, *Annals of Mathematical Statistics* 37: 1173-1185.

Stein, C. (1945). A Two Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance, *Annals of Mathematical Statistics* 16: 243-258.

Stein, C. (1949). Some Problems in Sequential Estimation (Abstract), *Econometrica* 17: 77-78.

Xu, K. (2007). U-statistics and Their Asymptotic Results for Some Inequality and Poverty Measures, *Econometric Reviews* 26: 567–577.