ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Bilson Darku, Francis; Konietschke, Frank; Chattopadhyay, Bhargab

Article

Gini index estimation within pre-specified error bound: Application to Indian household survey data

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Bilson Darku, Francis; Konietschke, Frank; Chattopadhyay, Bhargab (2020) : Gini index estimation within pre-specified error bound: Application to Indian household survey data, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 8, Iss. 2, pp. 1-20, https://doi.org/10.3390/econometrics8020026

This Version is available at:
https://hdl.handle.net/10419/247574

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

econometrics

MDPI

Article

Gini Index Estimation within Pre-Specified Error Bound: Application to Indian Household Survey Data

Francis Bilson Darku [1,†], Frank Konietschke [2,3] and Bhargab Chattopadhyay [4,*]

1 Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, USA; FBilsonD@nd.edu
2 Institute of Biometry and Clinical Epidemiology, Charité—Universitätsmedizin Berlin, 10117 Berlin, Germany; Frank.Konietschke@charite.de
3 Berlin Institute of Health, Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany
4 Department of Decision Sciences and Information Systems, Indian Institute of Management Visakhapatnam, Visakhapatnam, Andhra Pradesh 530003, India
* Correspondence: Bhargab@iimv.ac.in
† This work is part of the final dissertation of Francis Bilson Darku that was submitted to the Department of Mathematical Sciences at The University of Texas at Dallas.

Received: 23 July 2019; Accepted: 12 June 2020; Published: 18 June 2020

Abstract: The Gini index, a widely used economic inequality measure, is computed using data whose designs involve clustering and stratification, generally known as complex household surveys. Under complex household survey, we develop two novel procedures for estimating Gini index with a pre-specified error bound and confidence level. The two proposed approaches are based on the concept of sequential analysis which is known to be economical in the sense of obtaining an optimal cluster size which reduces project cost (that is total sampling cost) thereby achieving the pre-specified error bound and the confidence level under reasonable assumptions. Some large sample properties of the proposed procedures are examined without assuming any specific distribution. Empirical illustrations of both procedures are provided using the consumption expenditure data obtained by National Sample Survey (NSS) Organization in India.

Keywords: complex household survey; confidence interval; income distribution; inequality; sequential analysis

1. Introduction

Economic measures based on income levels of the residents of a specific region play an important role in social, economic and socio-economic sciences. They are used to quantify both the actual balance of the economy as well as the wealthiness and poverty of the people. One of the most prominent candidates is the (normalized) Gini index,

$$G_F = G_F(X) = \frac{2}{\mu}\int_0^\infty xF(x)\,dF(x) - 1, \quad \mu = \mathrm{E}(X), \tag{1}$$

which quantifies the economic inequality of a region, state, country or the world. Here, the random variable X denotes the income level, $F(x)$ its cumulative distribution function, and $\mu = \mathrm{E}(X)$ its expected value. If $G_F = 0$, then the economic system has maximal equality (e.g., everyone has the same income), while $G_F = 1$ represents perfect inequality (e.g., one individual has everything while the rest have nothing). For example, according to the Organization for Economic Cooperation and Development (2017), the Gini indices of the USA, Germany and South Africa were $G_F = 0.39, 0.29, 0.62$ in 2017, respectively. These values suggest income inequality in these regions. Therefore, Gini index serves as a measure of economic balance that allows comparison across regions. Roughly speaking,

income levels were more balanced (equal) in Germany than they were in the USA and in Brazil, respectively. Thus, the Gini index serves as an important measure in economics, social and political sciences. The estimation of the Gini index G_F of a country or a region, however, is a rather challenging task, because income is usually measured on household levels and thus in a clustered and stratified way. In most countries (e.g., United States, European Union, India and others), complex household surveys are conducted annually, the data of which can be used for the estimation of the Gini index as given in (1) see (Bhattacharya 2005, 2007).

The single computation of a point estimator of G_F as being reported in most available resources is, however, rather unsatisfactory, because neither the variability in the sample nor sample/cluster sizes visualize the estimator in an informative manner. Therefore, computing $100(1-\alpha)\%$ confidence intervals for G_F as point estimators are much more informative for making both descriptive as well as comparative conclusions. Binder and Kovacevic (1995) and Bhattacharya (2007) proposed point estimators of the Gini index as well as of their standard errors in such complex survey designs (see Section 2 for details), which can be used for the computation of $100(1-\alpha)\%$ confidence intervals for G_F. Furthermore, Peng (2011) proposed an empirical likelihood-based approach to construct such confidence intervals (as well as the confidence interval for the difference of two Gini indices). Clearly, for a desired confidence level, a narrower confidence interval will be more accurate about the parameter of interest. Therefore, it is the aim of the present article to develop confidence intervals for the Gini index G_F in complex survey designs that both control the nominal confidence level $(1-\alpha)$ and the confidence interval width. To guarantee that these criteria will be fulfilled, the optimal number of clusters will be computed using an innovative 'learn-as-you-go' or sequential procedure. We refer the readers to Ghosh and Sen (1991); Ghosh et al. (1997); Chattopadhyay and Kelley (2017); Kelley et al. (2018) and others for more on sequential analysis literature.

The first known application of sequential analysis in surveys was done by Mahalanobis (1940), who described the design and implementation of the method (in a different context) for estimating acreage of jute crop in the whole state of Bengal in undivided India. This was even before the seminal works of Stein (1945, 1949) on sequential analysis area. Kanninen (1993); Greene (1998); Arcidiacono and Jones (2003); Aguirregabiria and Mira (2007) and many others contributed to application of sequential analysis in the field of economics, data analysis, medicine, and other areas. Recently, Chattopadhyay and De (2016) and De and Chattopadhyay (2017) developed a sequential procedure for inference problems related to the Gini index under independent and identically distributed (i.i.d.) conditions, but the proposed methodology cannot be used for finding a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the population Gini index under a complex household survey design. We propose a two stage procedure and a purely sequential procedure to find an estimate of the minimum number of clusters which is required to find a sufficiently narrow confidence interval under a distribution-free scenario. Both the two-stage and purely sequential procedures are applied to the 64th round of household survey data collected in India. Further, a simulation study is carried out on observations collected in the Indian household survey data and from known income distributions to explore the properties of the procedures.

The remainder of this paper is organized as follows: Section 2 describes the sampling framework of the complex survey design that is considered in this work. In Section 3, we formulate the problem of finding a sufficiently narrow confidence interval for the Gini index and the reason for non-applicability of a procedure with fixed cluster size. In Section 4, we develop the purely sequential, as well as the two-stage, procedure followed by a discussion on the characteristics of our procedure in Section 5. Furthermore, an application of both of our procedures to real and synthetic data sets can be found in Section 6, while Section 7 describes an extension of the problem to the multivariate setup. We discuss the advantage and drawbacks of the proposed procedures in Section 8, and provide concluding comments in Section 9.

2. Survey Design and Point Estimation

In this section, the complex household survey design along with the used notations will be described: Assume that the population is divided into $s = 1, 2, \ldots, S$ strata, whereas the sth stratum is divided into $c_s = 1, \ldots, H_s$ clusters. Under the c_s^{th} cluster in stratum s, there is a group of M_{sc_s} households with ν_{sc_sh} individuals or members, $h = 1, 2, ..., M_{sc_s}$. Therefore, the total number of clusters in the population is $H = \sum_{s=1}^{S} H_s$. The number of households in a stratum will be $M_s = \sum_{c_s=1}^{H_s} M_{sc_s}$ and the total number of households in the population is denoted by $M = \sum_{s=1}^{S} M_s = \sum_{s=1}^{S} \sum_{c_s=1}^{H_s} M_{sc_s}$.

For estimation purpose in such complex survey designs, a sample of n_s clusters is selected from the sth stratum by simple random sampling with replacement. A simple random sample of k households is then considered (without replacement) from each of the selected clusters. Let the total number of clusters being selected from the population be denoted by

$$n = \sum_{s=1}^{S} n_s \quad \text{with} \quad n_s = a_s n \quad \text{and} \quad a_s = \frac{H_s}{H}. \tag{2}$$

Thus, the total number of households in the sample will be $kn = k\sum_{s=1}^{S} n_s$. For the hth household in the c_s^{th} cluster from the sth stratum, the observed data (that is, the household monthly income, monthly expenditure, per capita income or others) are denoted as x_{sc_sh}. With the presence of stratification and clustering, the households are assigned different weights W_{sc_sh} as the probability of inclusion in the sample will vary. The assigned weight to the selected household is computed as the inverse of the probability of inclusion of the household in the sample (see Binder and Kovacevic 1995; Horvitz and Thompson 1952; Lee and Forthofer 2006). If researchers wish to increase (or decrease) the representation of a subgroup of the population that is of interest, they can employ oversampling (or undersampling) procedures and use appropriate weighting techniques. Wells (1998) discussed several weighting methods for such cases. For our survey framework, weights are assigned to the data (x_{sc_sh}) with respect to the number of observations in the population. The attached weight for all the ν_{sc_sh} members of the hth household belonging to the c_s^{th} sampled cluster from the sth stratum as given by Bhattacharya (2007) is

$$W_{sc_sh} = \frac{M_{sc_sh} H_s}{kn_s} \nu_{sc_sh}.$$

It should be noted that the computation of the sampling weights will change depending on the sampling design and also on whether the analysis is being done at the district-, household- or individual level (Bhattacharya 2007). If the cluster size is large, sampling with or without replacement will result in similar values for the weights. Moreover, Bhattacharya (2005, 2007) noted that using sampling with or without replacement does not affect the asymptotic results of this work as in most practical situations the number of clusters per stratum are usually large.

Under the above framework, let

$$W = \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} W_{sc_sh}$$

denote the total of per-household weights associated with the survey and define

$$w_{sc_sh} = W^{-1} W_{sc_sh}$$

the normalized weights, which will be used in the estimation of the average income $\mu = \mathrm{E}(X)$ and its cumulative distribution function $F(x)$ by

$$\hat{\mu} = \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} x_{sc_sh}$$

and

$$\hat{F}(x) = \sum_{i=1}^{S} \sum_{j=1}^{n_s} \sum_{l=1}^{k} w_{ijl} \mathbf{1}\left(x_{ijl} \leq x\right),$$

in order to take the relative household sizes into account. Then, under fairly mild conditions on the numbers of clusters, a consistent estimator of the Gini index G_F given in (1) is given by

$$\hat{G}_n = 1 - \frac{2}{\hat{\mu}} \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} x_{sc_sh} \left(1 - \hat{F}\left(x_{sc_sh}\right)\right). \tag{3}$$

It follows that, the estimated Gini index basically is a ratio of two weighted averages of the income levels, respectively (see Bhattacharya 2007). In the next section, we discuss the idea towards the construction of confidence intervals with bounded width.

3. Bounded Width Confidence Intervals

In order to derive bounded width confidence intervals, the (asymptotic) distribution of the empirical Gini index $\hat{G}_n$ must be tackled. It has been shown by Bhattacharya (2007), that if $\mathrm{E}\left(|X| \middle| s\right) < \infty$, and if $n_s \to \infty$ for each stratum $s = 1, \ldots, S$ at the same rate, then

$$\sqrt{n}\left(\hat{G}_n - G_F\right) \xrightarrow{D} N\left(0, \xi^2\right). \tag{4}$$

Here, ξ^2 denotes the (asymptotic) variance of $\sqrt{n}\hat{G}_n$. Due to its quite involved representation, we refer to Bhattacharya (2007) for the specific variance formula. The asymptotic distribution, however, can now be used for the computation of $100(1-\alpha)\%$ confidence intervals for the population Gini index the width of which does not exceed a pre-specified value ω, that is

$$\Pr\left(\hat{G}_n - z_{\alpha/2}\frac{\xi}{\sqrt{n}} < G_F < \hat{G}_n + z_{\alpha/2}\frac{\xi}{\sqrt{n}}\right) \geq 1 - \alpha,$$

and

$$L = 2z_{\alpha/2}\frac{\xi}{\sqrt{n}} \leq \omega.$$

Here, $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution $\mathrm{N}(0,1)$. Thus, the actual arising task is the computation of n that will guarantee that the width of the confidence interval is bounded by ω, i.e.,

$$\frac{\omega\sqrt{n}}{2\xi} \geq z_{\alpha/2} \implies n \geq \frac{4z_{\alpha/2}^2\xi^2}{\omega^2} = C. \tag{5}$$

Hence, C denotes the optimal total number of clusters from all strata needed such that $L \leq \omega$. Therefore, the optimal number of clusters that will be required to be sampled from the sth stratum $(s = 1, 2, \ldots, S)$ will be $C_s = Ca_s$. Here, the term optimal is used in the sense of minimum number of clusters to meet the requirements and not as in the sense of optimal allocation used in sample survey methods (see Cochran 1997). If C is known, one can find the sufficiently narrow confidence interval

$$\left(\hat{G}_C - z_{\alpha/z}\frac{\xi}{\sqrt{C}}, \hat{G}_C + z_{\alpha/z}\frac{\xi}{\sqrt{C}}\right),$$

that satisfies (5). However without knowing the underlying distribution of the income (or assets or expenditure), the value of ξ^2 is unknown in practical scenarios. Thus, the optimal cluster size from all the S strata, C, is also unknown. We note that supposed value (or previous survey estimate) of ξ^2 may be used to obtain the value of C. However, a potential problem that may arise is that the supposed value of ξ^2 may be different from the actual value. Moreover, using previous survey estimates in many situations is not advised as that may not be applicable in the current population. This is because of

a possible change in socio-economic conditions that may arise due to the change in distribution of income or expenditure as a result of change in economic policies or situations. Due to all these factors, the value of C may widely differ from what it would have been if ξ^2 is known and will not guarantee that (5) is satisfied. The (asymptotic) variance ξ^2 of the estimated Gini index is, however, unknown in practical applications and must be estimated in an appropriate way. Consistent estimators will now be discussed below.

Estimation of ξ^2

Several articles published in statistics and economics journals have proposed different estimators of the asymptotic variance parameter of the estimator of the Gini index under different sampling schemes. Zitikis and Gastwirth (2002) proposed explicit formulas for the asymptotic variance of a general class of the Gini index (i.e., the S-Gini index) for simple random sampling with observations coming from the Exponential and Pareto distributions. We refer to Langel and Tillé (2013) for a discussion on several techniques used in estimating the asymptotic variance of the Gini index for various sampling designs. Under the current framework, Binder and Kovacevic (1995) proposed an estimator of ξ^2 using the empirical variance

$$V_{n,1}^2 = \sum_{s=1}^{S} \frac{n_s}{n_s - 1} \sum_{c_s=1}^{n_s} (u_{sc_s} - \bar{u}_s)^2 \tag{6}$$

of the values

$$u_{sc_s} = \frac{2}{\hat{\mu}} \sum_{h=1}^{k} w_{sc_sh} \left[A(x_{sc_sh}) x_{sc_sh} + B(x_{sc_sh}) - \frac{\hat{\mu}}{2} \left(\hat{G}_n + 1 \right) \right].$$

Here, $\bar{u}_s = n_s^{-1} \sum_{c_s=1}^{n_s} u_{sc_s}$ denote the empirical mean of u_{sc_s} and

$$A(x_{sc_sh}) = \hat{F}(x_{sc_sh}) - \frac{\hat{G}_n + 1}{2}, \text{ and}$$
$$B(x_{sc_sh}) = \sum_{a=1}^{S} \sum_{b=1}^{n_s} \sum_{c=1}^{k} w_{abc} x_{abc} \mathbf{1}(x_{abc} \geq x_{sc_sh}),$$

are weighted placements and averages of the income values obtained from n clusters, respectively. It should be noted that Bhattacharya (2007) proposed an alternative estimator of ξ^2 which is given by

$$\begin{aligned} V_{n,2}^2 = & \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh}^2 \hat{\psi}_{sc_sh}^2 + \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} \sum_{h' \neq h} w_{sc_sh} \hat{\psi}_{sc_sh} w_{sc_sh'} \hat{\psi}_{sc_sh'} \\ & - \sum_{s=1}^{S} \frac{1}{n_s} \left(\sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} \hat{\psi}_{sc_sh} \right)^2, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \hat{\psi}_{sc_sh} = & -\frac{2}{\mu} \sum_{g=1}^{kn} w_g \left[x_{sc_sh} \mathbf{1}(x_{sc_sh} \leq x_{(g)}) + x_{(g)} \left(\hat{F}(x_{(g)}) - \mathbf{1}(x_{sc_sh} \leq x_{(g)}) \right) \right] \\ & + \frac{2}{\hat{\mu}^2} \sum_{g=1}^{kn} \left[\left\{ \sum_{a=1}^{S} \sum_{b=1}^{n_s} \sum_{c=1}^{k} w_{abc} x_{abc} \mathbf{1}(x_{abc} \leq x_{(g)}) \right\} x_{sc_sh} \right], \end{aligned}$$

$kn = k \sum_{s=1}^{S} n_s$ is the total number of observations, and

$x_{(g)}$ is the gth ordered observation (among all x_{sc_sh}).

However, Hoque and Clarke (2015) showed that the estimators in (6) and (7) are numerically the same, i.e., $V_{n,1}^2 = V_{n,2}^2$. We therefore chose $V_{n,1}^2$ as a consistent estimator of ξ^2 and drop the second subscript, without loss of generality (i.e., we use V_n^2 as the estimator of ξ^2). Having found a consistent estimator of the (asymptotic) variance ξ^2, it follows that the optimal number of clusters C defined in (5) that lead to the bounded width confidence interval can now be estimated from the data. In order to do so, different sequential methodologies will be discussed in the next section.

4. Sequential Methodology

In this section, different sequential methodologies including two-stage and purely sequential approaches will be discussed to find the sufficiently narrow confidence interval. First, purely sequential methods will be introduced.

4.1. Purely Sequential Procedure

The purely sequential confidence interval computation is based on consecutive sampling until a certain stopping rule is met which ensured that the width of the confidence interval is smaller than or equal to the given bound. This sampling process begins with a pilot sample the sizes of which will be specified in Section 4.3. However, recall that computing a bounded width confidence interval requires at least C_s clusters from the sth stratum ($s = 1, 2, \ldots, S$). Therefore, choose a pilot cluster size of t_s from each stratum s, which results in a total number of clusters in the pilot stage of $t = \sum_{s=1}^{S} t_s$. Within each selected cluster, there are k randomly selected households (without replacement). Now, collect pilot observations $x_{s11}, \ldots, x_{s1k}, \ldots, x_{st_s1}, \ldots, x_{st_sk}$ on each stratum $s = 1, \ldots, S$. Now, the estimator V_n^2 of ξ^2 is computed to examine the following stopping rule

$$\begin{gathered} N = N_\omega(\leq H) \text{ is the smallest integer } n(\geq t) \text{ such that} \\ n \geq \frac{4z_{\alpha/2}^2}{\omega^2}\left(V_n^2 + \frac{1}{n}\right) = \hat{C} \quad \text{and} \quad n_s \geq \hat{C}_s = \hat{C}a_s, \text{for all } s. \end{gathered} \tag{8}$$

If the condition in the stopping rule is not satisfied, the surveyor collects data from additional $m'(\geq 1)$ clusters, with k randomly chosen households, from each stratum that has $n_s \leq \hat{C}_s$. Then ξ^2 is estimated based on all the observations collected up to that stage and the stopping condition is checked. This process is repeated until the condition in the stopping rule is satisfied. It should be noted that $m'(\geq 1)$ can be any integer that is appropriate, suitable or feasible for the survey.

The term $1/n$ in (8) is a correction term incorporated to avoid early stopping of the sequential procedure as V_n^2 (the estimator of ξ^2) may be very small in the early stages. Without this term, the stopping rule in (8) can be satisfied for very small sample sizes due to sampling error. In general, any null-sequence, e.g., $1/n^\gamma$, where $\gamma(> 0)$ is a fixed number, can be used as a correction term, because it does not affect the the consistency of the variance estimator (see Mukhopadhyay and De Silva 2009, p. 260, for more details). The use of a correction term can be seen in several articles, e.g., Chattopadhyay and De (2016), Chattopadhyay and Kelley (2017), and Kelley et al. (2019). The final cluster size N constitutes N_s clusters from each stratum s where

$$N_s = Na_s, \text{for } s = 1, 2, \ldots, S.$$

Based on the sampled data x_{sc_sh} and their corresponding standardized weights w_{sc_sh}, where $s = 1, \ldots, S$, $c_s = 1, \ldots, N_s$, and $h = 1, \ldots, k$, the $100(1-\alpha)\%$ bounded width confidence interval for the Gini index G_F is given by

$$\left(\hat{G}_N - z_{\alpha/2}\frac{V_N}{\sqrt{N}}, \hat{G}_N + z_{\alpha/2}\frac{V_N}{\sqrt{N}}\right). \tag{9}$$

The purely sequential procedure may be numerically cumbersome due to the consecutive sampling and repeated computations of the variance estimators. Therefore, a less numerically intensive method—a two-stage procedure—will be examined in the next section.

4.2. Two-Stage Procedure

Unlike the purely sequential procedure, the two-stage procedure comprises of two stages. The first stage is called the pilot stage, wherein a sample is drawn from the population. That is, first a pilot sample of clusters, t_s (with $\sum_{s=1}^{S} t_s = t$), is selected from each stratum s. Based on the sample from the pilot stage, ξ^2 is estimated as in (6). Then, the total final cluster size from all strata can be estimated as

$$Q = \min\left\{H, \max\left\{t, \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2} V_t^2 \right\rceil\right\}\right\} = \min\{H, Q^*\} \tag{10}$$

where Q^* is the (unbounded) optimal cluster size and $\lceil\cdot\rceil$ is the ceiling function, that is, $\lceil x \rceil$ is the smallest integer that is greater than or equal to x. Thus, the estimated number of clusters to be sampled from the sth stratum is given by

$$Q_s = \min\{H_s, [Qa_s]\},$$

with a_s as defined in (2) and $[\cdot]$ being the nearest integer function. So, in the second stage, observations from k households will be collected from $Q_s - t_s$ clusters from each stratum s. Using the combined data from the two stages, the estimator of ξ^2 is updated and the approximate $100(1-\alpha)\%$ confidence interval for the Gini index is given by

$$\left(\hat{G}_Q - z_{\alpha/2}\frac{V_Q}{\sqrt{Q}}, \hat{G}_Q + z_{\alpha/2}\frac{V_Q}{\sqrt{Q}}\right). \tag{11}$$

We note that the final cluster size using either the two-stage procedure or the purely sequential procedure can be shown to be always finite. In addition, the number of clusters per stratum are mutually dependent as they all depend on the same stopping rule. In the next subsection, we derive the pilot cluster size formula.

4.3. Pilot Cluster Size

Using (8) and proceeding along the lines of Chattopadhyay and De (2016), we have

$$n \geq \frac{4z_{\alpha/2}^2}{\omega^2}\left(V_n^2 + \frac{1}{n}\right) \geq \frac{4z_{\alpha/2}^2}{\omega^2}\frac{1}{n} \implies n \geq \frac{2z_{\alpha/2}}{\omega}. \tag{12}$$

Thus the total number of sampled clusters is at least $2z_{\alpha/2}/\omega$. The maximum number of clusters from the sth stratum is H_s and also the minimum number of clusters to estimate ξ^2 is 2. Considering all the constraints in (8), the number of clusters recommended to be sampled from the sth stratum at the pilot stage is

$$t_s = \min\left\{H_s, \max\left\{2, \left\lceil \frac{2a_s z_{\alpha/2}}{\omega} \right\rceil\right\}\right\}. \tag{13}$$

We note that this ensures that the minimum cluster size is met as well as the total possible cluster size is not exceeded.

5. Characteristics of the Procedures and Simulation Study

The purely sequential procedure and the two-stage procedure for constructing a sufficiently narrow confidence interval for the Gini index—unlike fixed cluster size procedures—require cluster sizes which are obtained from data. So, the respective cluster sizes N and Q are random in nature. In the following subsection, we will look at the characteristics of the random cluster sizes viz. N and Q.

5.1. Characteristics

The following theorem provides some asymptotic properties (as $\omega \to 0$) of the final cluster sizes of the above procedures with sufficiently large H.

Theorem 1. *If the parent distribution(s) is(are) such that $E[V_n^2]$ exists and H_s (fixed) are sufficiently large for all $s \in S$, then as $\omega \to 0$,*

(i) $\dfrac{N}{C} \to 1$ *in probability,*
(ii) $\dfrac{Q}{C} \to 1$ *in probability, and*
(iii) $\dfrac{2z_{\alpha/2}V_N}{\sqrt{N}} \leq \omega$.

Proof of Theorem 1. (i) The definition of stopping rule N associated with the purely sequential procedure in (8) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_N^2 \leq N \leq t\mathbf{1}(N = t) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(V_{N-1}^2 + \frac{1}{N-1}\right). \tag{14}$$

Since $N \to \infty$ as $\omega \downarrow 0$ and $V_n^2 \to \xi^2$ in probability as $n \to \infty$, by applying Theorem 2.1 of Gut (2009), $V_N^2 \to \xi^2$ in probability.

Furthermore, $t\Pr(N = t)/C \leq t/C \to 0$ as $\omega \downarrow 0$. Hence, dividing all sides of (14) by C and letting $\omega \downarrow 0$, we prove $N/C \to 1$ in probability as $\omega \downarrow 0$.

(ii) The definition of final cluster size Q related to the two-stage procedure in (10) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_t^2 \leq Q \leq t\mathbf{1}(Q = t) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(V_t^2 + \frac{1}{t}\right). \tag{15}$$

Furthermore, $t\Pr(Q = t)/C \leq t/C \to 0$ as $\omega \downarrow 0$. Now, $V_t^2 \to \xi^2$ in probability as $\omega \downarrow 0$. Hence, dividing all sides of (15) by C and letting $\omega \downarrow 0$, we prove $Q/C \to 1$ in probability as $\omega \downarrow 0$.

(iii) Using stopping rule N in (8) we have, for all N,

$$\begin{aligned} \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_N^2 \leq N \quad &\Longrightarrow \quad \frac{4z_{\alpha/2}^2}{N} V_N^2 \leq \omega^2 \\ &\Longrightarrow \quad 2z_{\alpha/2}\frac{V_N}{\sqrt{N}} \leq \omega \end{aligned}$$

□

Parts (i) and (ii) of the theorem show that the final cluster size as obtained from the purely sequential and the two-stage procedure is a consistent estimator of the cluster size provided ξ^2 is known. Part (iii) of the theorem shows that the sufficiently narrow confidence interval (that is length less than or equal to ω) will be obtained by the purely sequential procedure. The same result cannot be proven for the two-stage procedure.

5.2. Simulation Study

We now use a detailed simulation study, presented in the Supplement, to illustrate and compare the properties of our purely sequential and the two stage procedures in constructing a $100(1-\alpha)\%$ confidence interval for the Gini index under a complex survey whose width is less than ω. We presented two different simulation studies with 5000 simulation runs—(a) simulation using the NSS survey data as the population and (b) a Monte Carlo simulation in which the observations are drawn from three different populations, each of which has been drawn using three different distributions,

namely; Pareto, Gamma and Lognormal distributions. The two simulation studies were performed in RStudio (RStudio Team 2018, version 1.2.1335) and codes are available upon request.

To begin with, we describe the simulation procedure for the purely sequential methodology. From the given populations, $t_s(s = 1, 2, ..., S)$ clusters are randomly sampled from the sth stratum without replacement. From there, four households are selected from each cluster using simple random sampling without replacement and these households from all t clusters will constitute the pilot sample. From the collected pilot sample, the asymptotic variance of the Gini index ξ^2 is estimated using (6), and from (8), the optimal number of clusters C is estimated. The stopping rule is checked and if it is satisfied, sampling is terminated. On the other hand, if the stopping rule is not satisfied, the strata whose number of clusters selected are less than the expected number, that is $\{s : t_s < \hat{C}_s\}$, are identified and additional m' number of clusters are randomly selected without replacement. Here, m' is chosen to be either 1, 10 or 20. In each of the selected m' clusters, four households are randomly selected without replacement. At this stage, with the total number of sampled clusters being n (say), the value of V_n^2 is updated and the stopping rule is checked. If the rule is met, sampling is stopped, otherwise the strata without enough clusters are identified again and additional m' clusters are collected from each of them. This process is continued until and unless the stopping rule is met. At that point, based on N (say) numbers of clusters sampled from all strata, the $100(1 - \alpha)\%$ confidence interval for the Gini index is constructed as given in (9).

Unlike the purely sequential procedure described above, the two-stage procedure has only two stages. The simulation algorithm for the two-stage is as follows. From a given population, t_s number of clusters are randomly selected without replacement from the sth stratum and four households are randomly sampled from each of the selected clusters without replacement. The per monthly capita expenditure x_{sc_sh} from the selected households, with their respective weight W_{sc_sh}, are used to estimate the asymptotic variance of the Gini index (from (6)). This is followed by using (10) to obtain the optimal number of clusters Q needed to achieve the desired confidence level and width. If $Q > t$, additional $Q_s - t_s$ number of clusters are randomly selected without replacement from each stratum s. In each of the additional clusters, four households are also randomly selected without replacement. Finally, per capita monthly expenditure of all households from the Q number of clusters are used to construct the $100(1 - \alpha)\%$ confidence interval for the Gini index as stated in (11).

From the simulations, we find that the coverage probability for the confidence intervals for both purely sequential procedure and the two-stage procedure are approximately close to the desired confidence level provided that the cluster size (in all strata) is large, which is also a basic criterion while proving the asymptotic normality in (4). However, the width of the confidence intervals for the two stage procedure, unlike the purely sequential procedure, may result in confidence intervals of width larger than the pre-specified value of ω. For details, one may look at Tables S21–S24 of the supplementary material. This outcome is not surprising since the two-stage procedure is based on only the pilot sample which is usually taken to be small. So, the variability of the variance estimator V_n^2 is higher. The optimal cluster sizes obtained by the purely sequential procedure is less than the one obtained by the two-stage procedure. The newly developed methods can now be applied using real data. This will be explained in the next section.

6. Gini Index Estimation in India

We now apply the sequential procedures to construct bounded width confidence intervals for the Gini index in India using the per capita monthly expenditures obtained via the 64th Round National Sample Survey (NSS) (a stratified multi-staged survey design between July 2007 and June 2008). In 2008, the country was divided into 28 states and seven union territories thereof each was subdivided into districts. Within each district, two basic sectors were formed; all rural areas constituted the rural sector while all urban areas constituted the urban sector. Nonetheless, for the urban areas in a district, separate basic strata were formed for each town that had at least a population of 10 lakhs (1 lakh is 100,000). The remaining areas were grouped as another basic stratum (National Sample Survey Office

2007). For the rural sector, the sampling frame was made up of villages while for the urban sector, it was towns/blocks.[1]

Census villages and the Urban Frame Survey blocks were the first stage units (FSU) in the rural and urban sectors respectively. From each strata, FSUs are selected from the rural sector with probability proportional to size with replacement and from the urban sector by using simple random sampling without replacement. Within the FSU, the households in each sector were considered as the smallest unit of grouping, which is also referred to as the ultimate stage units. Households were selected by simple random sampling without replacement and various information about the households were recorded during the survey. Some of the information include the demographics, household size, expenditure on education, food, clothing, corresponding weights etc. A detailed description of the NSS Data can be found online at National Sample Survey Office (2015).

The "Stratum" variable in the 64th NSS data set will be used to stratify the states/sectors while "FSUno" (First Stage Unit Number) variable will be used to cluster the households under each stratum. We discuss the results obtained from applying the proposed sequential methodologies which were applied to the data collected from two of the most populous states in India, namely Uttar Pradesh and West Bengal. Additionally, the report includes the results for the whole state as well as rural and urban sectors of the state. Here, all the households in each cluster were considered since we are sampling from a survey that already has few number of households per cluster. However, the weight per household is adjusted at each sampling stage to reflect the actual weight that would be used during a survey.

In applying the sequential methodologies, the pilot cluster sizes t_s for each stratum s are computed using (13). At the outset, t_s number of clusters are selected from stratum s for $s = 1, \ldots, S$. Where t_s is same for both the purely sequential procedure and the two-stage methodology. We apply each of the procedures considering the survey data as our population.

6.1. Application of Purely Sequential Procedure (PSP)

The proposed purely sequential procedure, with observations from one cluster collected at each stage after the pilot stage, is applied to the NSS 64th round data. The results for different combinations of pre-specified width ($\omega \in \{0.020, 0.025\}$) and confidence level ($1 - \alpha, \alpha \in \{0.05, 0.10\}$) can be found in Tables 1–4. The first column of the tables indicates the region on which we applied our procedure. The PSP was applied on the entire data from Uttar Pradesh (denoted as *All*) and then separately applied on the rural and urban sectors of Utter Pradesh (denoted as *Rural* and *Urban* respectively). The same process was also repeated for West Bengal. The second column of the tables shows the estimated Gini index ($\hat{G}_H$) and its standard error ($se\left(\hat{G}_H\right)$) using the entire number of clusters (H) available in the data set for that region (i.e., all of the state, rural sector of the state, or the urban sector of the state). In the third column is the total number of clusters (H) available in the data set for that region. The fourth column shows the value of $\hat{C}$ when the procedure ended, $\hat{C}$ being the estimated optimal cluster size as in (8). The fifth column of the tables shows the collected cluster size N using the stopping rule in (8) and the pilot cluster size t. The values of $\hat{G}_N$ and $se(\hat{G}_N)$ in the sixth column are the estimated Gini index and its standard error respectively based on N clusters. The next two columns are respectively the lower and upper limits of the confidence intervals obtained with the stopping rule in (8). The ninth column is w_N which is the estimated width of the confidence interval. The last column $\Pr(N_s < \hat{C}_s)$ shows the proportion of strata that had their collected cluster size N_s from the purely sequential procedure being less than their estimated optimal cluster size $\hat{C}_s$ (N_s is the

[1] The survey excluded "(i) Leh (Ladakh) and Kargil districts of Jammu and Kashmir (for central sample), (ii) interior villages of Nagaland situated beyond 5 km of the bus route and (ii) villages of Andaman and Nicobar Islands which remain inaccessible throughout the year." (National Sample Survey Office 2007).

final number of clusters selected from stratum s while $\hat{C}_s$ is the estimated optimal number of clusters to be sampled from stratum s).

Table 1. Application results for PSP on NSS 64th round data for $\alpha = 0.1, \omega = 0.02$.

Region	$\hat{G}_H$ $se(\hat{G}_H)$	H	$\hat{C}$	N (t)	$\hat{G}_N$ $se(\hat{G}_N)$	Lower CI	Upper CI	w_N	$\Pr(N_s < \hat{C}_s)$
Uttar Pradesh									
All	0.2163 (0.0042)	1262	622	672 (321)	0.2116 (0.0057)	0.2023	0.2209	0.0186	0.2138
Rural	0.1997 (0.0041)	903	505	523 (198)	0.2024 (0.0057)	0.1931	0.2117	0.0186	0.4
Urban	0.2229 (0.0092)	359	903	359 (180)	0.2229 (0.0092)	0.2077	0.2381	0.0304	1.0
West Bengal									
All	0.2320 (0.0051)	878	587	593 (190)	0.2334 (0.0058)	0.2239	0.2430	0.0191	0.1282
Rural	0.1812 (0.0048)	551	450	450 (172)	0.1816 (0.0057)	0.1723	0.1909	0.0186	0.2353
Urban	0.2609 (0.0077)	327	612	327 (185)	0.2609 (0.0077)	0.2482	0.2736	0.0254	1.0

Table 2. Application results for PSP on NSS 64th round data for $\alpha = 0.05, \omega = 0.02$.

Region	$\hat{G}_H$ $se(\hat{G}_H)$	H	$\hat{C}$	N (t)	$\hat{G}_N$ $se(\hat{G}_N)$	Lower CI	Upper CI	w_N	$\Pr(N_s < \hat{C}_s)$
Uttar Pradesh									
All	0.2163 (0.0042)	1262	834	878 (333)	0.2117 (0.0048)	0.2022	0.2212	0.0190	0.2138
Rural	0.1997 (0.0041)	903	643	667 (226)	0.2024 (0.0048)	0.1930	0.2117	0.0187	0.4
Urban	0.2229 (0.0092)	359	1282	359 (254)	0.2229 (0.0092)	0.2048	0.2410	0.0362	1.0
West Bengal									
All	0.2320 (0.0051)	878	906	878 (223)	0.2320 (0.0051)	0.2221	0.2419	0.0198	1.0
Rural	0.181 (0.0048)	551	552	551 (203)	0.1812 (0.0048)	0.1719	0.1906	0.01871	1.0
Urban	0.2609 (0.0077)	327	869	327 (207)	0.2609 (0.0077)	0.2458	0.2761	0.0303	1.0

Table 3. Application results for PSP on NSS 64th round data for $\alpha = 0.1, \omega = 0.025$.

Region	$\hat{G}_H$ $se(\hat{G}_H)$	H	$\hat{C}$	N (t)	$\hat{G}_N$ $se(\hat{G}_N)$	Lower CI	Upper CI	w_N	$\Pr(N_s < \hat{C}_s)$
Uttar Pradesh									
All	0.2163 (0.0042)	1262	401	540 (302)	0.2138 (0.0063)	0.2035	0.2242	0.0207	0.0
Rural	0.1997 (0.0041)	903	386	400 (168)	0.2014 (0.0070)	0.1899	0.2130	0.0231	0.1714
Urban	0.2229 (0.0092)	359	578	359 (168)	0.2229 (0.0092)	0.2077	0.2381	0.0304	1.0
West Bengal									
All	0.2320 (0.0051)	878	324	319 (158)	0.2288 (0.0069)	0.2175	0.2401	0.0226	0.1795
Rural	0.1812 (0.00477)	551	276	289 (138)	0.1829 (0.0066)	0.1721	0.1937	0.0216	0.2353
Urban	0.2609 (0.0077)	327	392	327 (142)	0.2609 (0.0077)	0.2482	0.2736	0.0254	1.0

Table 4. Application results for PSP on NSS 64th round data for $\alpha = 0.05, \omega = 0.025$.

Region	$\hat{G}_H$ $se(\hat{G}_H)$	H	$\hat{C}$	N (t)	$\hat{G}_N$ $se(\hat{G}_N)$	Lower CI	Upper CI	w_N	$\Pr(N_s < \hat{C}_s)$
Uttar Pradesh									
All	0.2163 (0.0042)	1262	572	653 (728)	0.2123 (0.0058)	0.2010	0.2236	0.0226	0.2138
Rural	0.1997 (0.0041)	903	496	510 (197)	0.2010 (0.0060)	0.1893	0.2128	0.0234	0.1714
Urban	0.2229 (0.0092)	359	821	359 (717)	0.2229 (0.0092)	0.2048	0.2410	0.0362	1.0
West Bengal									
All	0.2320 (0.0051)	878	517	519 (186)	0.2318 (0.0061)	0.2199	0.2437	0.0238	0.1538
Rural	0.1812 (0.0048)	551	351	352 (163)	0.1815 (0.0057)	0.1703	0.1927	0.0223	0.2353
Urban	0.2609 (0.0077)	327	556	327 (162)	0.2609 (0.0077)	0.2458	0.2761	0.0303	1.0

In Tables 1–4, it can be seen that, when the maximum available (to be drawn from) cluster size (H_s) per stratum are large, the purely sequential procedure is able to achieve desired precision, i.e., a narrow confidence interval, ($w_N \leq \omega$) for the Gini index with relatively fewer number of clusters sampled while maintaining the desired confidence level. This is shown in the results where $N < H$ for all of Uttar Pradesh and West Bengal, as well as their individual rural sectors. The same cannot be said about their urban sectors as they do not have enough maximum available clusters from the onset. Thus, the procedure did not reach the optimal cluster size but stopped when there were no more clusters remaining to be sampled.

The results also show that, aside the fact that the entire urban regions did not have enough clusters ($N = H < C$), each of the strata in the regions also do not have enough clusters (that is, $\Pr(N_s < \hat{C}_s) = 1$) to obtain a narrow confidence interval width. However, in the other regions (i.e., *All* and *Rural* for Uttar Pradesh and West Bengal), even though $\hat{C} < N < H$, some strata had $N_s < \hat{C}_s$. This is because some strata have more than enough clusters while others do not and that offsets each

other at the end. For example, it can be seen from Table 1 that in the rural sector of Uttar Pradesh, 40% of the strata did not have enough clusters even though, at the end, the confidence interval was 0.0186 wide which was less than the desired width of 0.02.

Next, the the results will be compared with the two-stage procedure as discussed in Section 4.2.

6.2. Application of Two-Stage Procedure

First, the estimator V_n^2 of ξ^2 is obtained from the pilot stage and then the final cluster size Q^* is computed. Q^* is then adjusted to account for the limited availability of clusters per stratum in the NSS data to obtain the possible number of clusters Q that can be sampled (see (10)). Here, Q is distributed over S strata as Q_s for stratum s; rounding off where Q_s is not an integer. The sum of Q_s gives the actual number of clusters, $\tilde{Q} = \sum_{s=1}^{S} Q_s$, that are sampled from all strata. Using $\tilde{Q}$ clusters, the Gini index and ξ^2 are re-estimated (or updated) and a $100(1-\alpha)\%$ confidence interval is constructed according to (11).

Similar to the application of the purely sequential procedure, the two-stage procedure is applied to the NSS 64th round data for different combinations of pre-specified precision (ω) and accuracy ($1-\alpha$) with the results shown in Tables 5–8. The second column of the tables indicates the total number of clusters H in the unit (i.e., the whole state, rural sector, or urban sector) of the NSS data. The third column displays estimated optimal number of cluster (Q^*) that are required in order to achieve the desired precision and accuracy. Below Q^* is the pilot number of clusters t. The next column shows the estimated optimal cluster sizes Q taking into account the total number of clusters available in the data, because the number of clusters are finite and limited. Furthermore, $\tilde{Q}$ is the actual number of clusters that can be sampled from all strata considering the fact that we can only sample integer number of clusters from each strata (i.e., rounding off where there are decimals in the number of clusters to be sampled from a stratum). Using (3) and (6), the Gini index estimate, $\hat{G}_H$, for the unit is computed using all H clusters with its standard error as $se(\hat{G}_H)$ and these are shown in the fifth column. The selected clusters are used to estimate the Gini index and it is denoted as $\hat{G}_{\tilde{Q}}$, with its standard error as $se(\hat{G}_{\tilde{Q}})$, in the sixth column. In the seventh and eighth columns, Lower CI and Upper CI are the lower and upper limits of the $100(1-\alpha)\%$ confidence interval of the Gini index using $\tilde{Q}$ clusters, respectively. The last column shows the length of the confidence interval, $w_{\tilde{Q}}$. It must be noted that Q^* is unbounded while on the other hand, Q and $\tilde{Q}$ cannot exceed H. $\tilde{Q}$ can be less than, equal to, or greater than Q depending on the rounding off. Q^* will be equal to Q if and only if Q^* is less than or equal to H.

Table 5. Application results for the two-stage procedure on NSS 64th round data for $\alpha = 0.1$ and $\omega = 0.02$.

Region	H	Q^* (t)	$\tilde{Q}$ (Q)	$\hat{G}_H$ ($se(\hat{G}_H)$)	$\hat{G}_{\tilde{Q}}$ ($se(\hat{G}_{\tilde{Q}})$)	Lower CI	Upper CI	$w_{\tilde{Q}}$
Uttar Pradesh								
All	1262	1146 (321)	1171 (1146)	0.2163 (0.0042)	0.2137 (0.0040)	0.2072	0.2202	0.0131
Rural	903	398 (198)	406 (398)	0.1997 (0.0041)	0.2027 (0.0053)	0.1940	0.2114	0.0174
Urban	359	1177 (180)	359 (359)	0.2229 (0.0092)	0.2229 (0.0092)	0.2077	0.2381	0.0304
West Bengal								
All	878	624 (190)	626 (624)	0.2320 (0.0051)	0.2307 (0.0055)	0.2216	0.2398	0.0182
Rural	551	422 (173)	420 (422)	0.1812 (0.0048)	0.1785 (0.0047)	0.1707	0.1862	0.0155
Urban	327	857 (185)	327 (327)	0.2609 (0.0077)	0.2609 (0.0077)	0.2482	0.2736	0.0254

Table 6. Application results for the two-stage procedure on NSS 64th round data for $\alpha = 0.05$ and $\omega = 0.02$.

Region	H	Q^* (t)	$\tilde{Q}$ (Q)	$\hat{G}_H$ ($se(\hat{G}_H)$)	$\hat{G}_{\tilde{Q}}$ ($se(\hat{G}_{\tilde{Q}})$)	Lower CI	Upper CI	$w_{\tilde{Q}}$
Uttar Pradesh								
All	1262	1665	1262	0.2163	0.2163	0.2081	0.2245	0.0164
		(333)	(1262)	(0.0042)	(0.0042)			
Rural	903	593	595	0.2000	0.2000	0.1914	0.2085	0.0171
		(226)	(593)	(0.0041)	(0.0044)			
Urban	359	1712	359	0.2229	0.2229	0.2048	0.2410	0.0362
		(254)	(359)	(0.0092)	(0.0092)			
West Bengal								
All	878	874	878	0.2320	0.2320	0.2221	0.2419	0.0198
		(223)	(874)	(0.0051)	(0.0051)			
Rural	551	535	534	0.1812	0.1814	0.1719	0.1910	0.0191
		(203)	(535)	(0.0048)	(0.0049)			
Urban	327	1110	327	0.2609	0.2609	0.2458	0.2761	0.0303
		(207)	(327)	(0.0077)	(0.0077)			

Table 7. Application results for the two-stage procedure on NSS 64th round data for $\alpha = 0.1$ and $\omega = 0.025$.

Region	H	Q^* (t)	$\tilde{Q}$ (Q)	$\hat{G}_H$ ($se(\hat{G}_H)$)	$\hat{G}_{\tilde{Q}}$ ($se(\hat{G}_{\tilde{Q}})$)	Lower CI	Upper CI	$w_{\tilde{Q}}$
Uttar Pradesh								
All	1262	688	680	0.2163	0.2104	0.2023	0.2185	0.0162
		(302)	(688)	(0.0042)	(0.0049)			
Rural	903	299	308	0.1997	0.2026	0.1927	0.2126	0.0199
		(168)	(299)	(0.0041)	(0.0061)			
Urban	359	1087	359	0.2229	0.2229	0.2077	0.2381	0.0304
		(168)	(359)	(0.0092)	(0.0092)			
West Bengal								
All	878	396	396	0.2320	0.2293	0.2171	0.2414	0.0243
		(158)	(396)	(0.0051)	(0.0074)			
Rural	551	275	275	0.1812	0.1750	0.1660	0.1840	0.0180
		(138)	(275)	(0.0048)	(0.0055)			
Urban	327	582	327	0.2609	0.2609	0.2482	0.2736	0.0254
		(142)	(327)	(0.0077)	(0.0077)			

From Tables 5–8, it can be observed that in all cases, except the urban sectors for both states, the confidence interval widths were less than ω. These results were achieved because the optimal number of clusters required (Q^*), according to the two-stage procedure, were less than the number available (H). On the other hand, in both Uttar Pradesh and West Bengal, the estimated optimal cluster sizes Q^* for the urban sector exceeded the available number of clusters H in the data. As a consequence of this, the confidence interval widths for the Gini index in the urban sectors were larger than the pre-specified bound, that is $w_{\tilde{Q}} > \omega$.

Table 8. Application results for the two-stage procedure on NSS 64th round data for $\alpha = 0.05$ and $\omega = 0.025$.

Region	H	Q^* (t)	$\tilde{Q}$ (Q)	$\hat{G}_H$ ($se(\hat{G}_H)$)	$\hat{G}_{\tilde{Q}}$ ($se(\hat{G}_{\tilde{Q}})$)	Lower CI	Upper CI	$w_{\tilde{Q}}$
Uttar Pradesh								
All	1262	976	947	0.2163	0.2124	0.2041	0.2207	0.0166
		(302)	(946)	(0.0042)	(0.0042)			
Rural	903	364	353	0.1997	0.2032	0.1922	0.2142	0.0220
		(197)	(364)	(0.0041)	(0.0056)			
Urban	359	1081	359	0.2229	0.2229	0.2048	0.2410	0.0362
		(177)	(359)	(0.0092)	(0.0092)			
West Bengal								
All	878	607	608	0.2320	0.2315	0.2204	0.2427	0.0224
		(186)	(607)	(0.0051)	(0.0057)			
Rural	551	391	392	0.1812	0.1759	0.1670	0.1849	0.0178
		(163)	(391)	(0.0048)	(0.0045)			
Urban	327	754	327	0.2609	0.2609	0.2458	0.2761	0.0303
		(162)	(327)	(0.0077)	(0.0077)			

7. Extension: Narrow Confidence Region

The methodology presented in this article for the Gini Index parameter can be extended to a multi-parameter setup in which we would like to make an inference about a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top$ for $p \geq 2$. This situation arises when we are interested in making joint inference related to a number of welfare related measures computed from socio-economic survey data (e.g., household consumer expenditure survey conducted by National Sample Survey, India). Thus, instead of a sufficiently narrow confidence interval, we would like to construct a narrow confidence region for a vector of parameters. Let the vector of estimators be defined as $\boldsymbol{T}_n = (T_{1n}, \ldots, T_{pn})^\top$ based on the data on n households collected using a complex household survey. We extend our proposed methodology for constructing the narrow confidence region in the spirit of Mukhopadhyay and De Silva (2009, pp. 284–89). We propose the following confidence region for $\boldsymbol{\theta}_F$:

$$\Re_n = \left\{ \boldsymbol{\theta} \in \mathbb{R}^p : (\mathbf{T}_n - \boldsymbol{\theta})^\top (\mathbf{T}_n - \boldsymbol{\theta}) \leq \omega^2 \right\}.$$

Using the regularity conditions by Bhattacharya (2005), we have,

$$\sqrt{n}\,(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \Sigma), \quad \text{i.e.,} \quad n(\boldsymbol{T}_n - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{T}_n - \boldsymbol{\theta}) \overset{a}{\sim} \chi^2_p,$$

with Σ being a positive definite matrix and χ^2_p being a chi-squared distribution with p degrees of freedom. If Σ is a positive definite matrix then there exist an orthogonal matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{\Delta}$ such that $\boldsymbol{P}^\top \Sigma \boldsymbol{P} = \Delta$. The diagonal elements of Δ contains the eigen values of Σ. If the positive eigen

values of Σ be $\lambda_1, \ldots, \lambda_p$ then $\Delta = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Furthermore, let $(\boldsymbol{PT}_n - \boldsymbol{P\theta}) = (Y_1, \ldots, Y_p)^\top$ and $\lambda_{(p)}$ is the maximum of the p eigen values of Σ. So, we have

$$\begin{aligned}
(\boldsymbol{T}_n - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{T}_n - \boldsymbol{\theta}) &= (\boldsymbol{T}_n - \boldsymbol{\theta})^\top \boldsymbol{P}^\top \Delta^{-1} \boldsymbol{P} (\boldsymbol{T}_n - \boldsymbol{\theta}) \\
&= (\boldsymbol{PT}_n - \boldsymbol{P\theta})^\top \Delta^{-1} (\boldsymbol{PT}_n - \boldsymbol{P\theta}) = \sum_{i=1}^{p} \frac{Y_i^2}{\lambda_i} \\
\lambda_{(p)} (\boldsymbol{T}_n - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{T}_n - \boldsymbol{\theta}) &\geq \sum_{i=1}^{p} Y_i^2 = (\boldsymbol{PT}_n - \boldsymbol{P\theta})^\top (\boldsymbol{PT}_n - \boldsymbol{P\theta}) \\
&= (\boldsymbol{T}_n - \boldsymbol{\theta})^\top (\boldsymbol{T}_n - \boldsymbol{\theta}) . \qquad (16)
\end{aligned}$$

Thus, using (16), we say,

$$\begin{aligned}
\Pr(\boldsymbol{\theta} \in \Re_n) &= \Pr\left[(\boldsymbol{T}_n - \boldsymbol{\theta})^\top (\boldsymbol{T}_n - \boldsymbol{\theta}) \leq \omega^2\right] \\
&\geq \Pr\left[\lambda_{(p)} (\boldsymbol{T}_n - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{T}_n - \boldsymbol{\theta}) \leq \omega^2\right] \\
&= \Pr\left[(\boldsymbol{T}_n - \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{T}_n - \boldsymbol{\theta}) \leq \frac{\omega^2}{\lambda_{(p)}}\right].
\end{aligned}$$

Provided $\chi^2_{\alpha;p}$ being the $100(1-\alpha)$th percentile of χ^2_p, we claim that the coverage probability of the confidence region $\Re_n$ is more than $(1-\alpha)$ if

$$\frac{n\omega^2}{\lambda_{(p)}} \geq \chi^2_{\alpha;p}, \quad \text{i.e.,} \quad n \geq \frac{\chi^2_{\alpha;p}\lambda_{(p)}}{\omega^2} = C.$$

Here, C is the required optimal cluster size that should be used provided the covariance matrix (Σ) is known. If the parameter Σ is known in advance, one could simply collect observations belonging to cluster $C_s, s = 1, \ldots, S$ of the each of the S Strata. Since Σ is not known in practice, we can estimate Σ, using a consistent estimator ($\boldsymbol{V}_n$, say) which can be obtained using the jackknife method. The consistency result of the jackknife estimator follows from Sen (1988). Thus using the jackknife estimator, we may propose either a two-stage or a sequential procedure. Similar results associated with the procedures described earlier is expected to hold under appropriate regularity conditions.

8. Discussion

At the outset, we would like to caution readers not to confuse two-stage sampling with the two-stage procedure discussed in Section 4.2, in the sequential sampling literature. For two-stage procedure, we refer Chattopadhyay and Mukhopadhyay (2013); Stein (1945) and others. A two-stage sampling (e.g., see Fuller (2009)) is a sampling technique in which a sample of clusters is selected and within those selected clusters, a sample of units are selected assuming the units to be independent of one another, and the selection rule depends only on the cluster. Under this two-stage sampling, Fuller (2009) discussed the use of Horvitz-Thompson estimator to estimate the total number and mean of the population and their respective variances. In addition, Fuller (2009) elaborated on the use of Horvitz–Thompson estimators and their (asymptotic) variances for functions of means and complex estimators, in general, under the assumption that the population distribution has a finite fourth moment. However, in the asymptotic framework of Fuller (2009), it was assumed that observations are independently and identically distributed (iid) which is a stronger assumption when compared to the framework of Bhattacharya (2007), also used in this work. Furthermore, Fuller (2009) also discussed the classical optimal sample allocation problem under the two-stage sampling technique for estimating the mean per element in a population. In his discussion, he assumed an equal number of units to be sampled from each cluster as well as an equal total number of units in each cluster and also known

population variances of the cluster size and the sampling units. Under these assumptions, Fuller (2009) obtained the optimal number of units to be sampled per cluster by minimizing the variance of the mean per element subject to a cost constraint.

Our work is different from the survey procedures discussed in Fuller (2009). Our work, as indicated earlier, is based on the survey framework used in Bhattacharya (2007). In order to get such a confidence interval for Gini index, we are interested in estimating the unknown optimum number of clusters in each of the stratum, prefixing the number of strata. Apart from the survey framework, in our work, optimal cluster size depends on the data unlike the procedures discussed in Fuller (2009). The total cluster size (as well as the cluster size per each stratum) is a random variable that depends on a stopping criterion. This procedure also makes the estimated cluster sizes mutually dependent as they are all estimated based on the same stopping rule. Thus, the method discussed in Fuller (2009) or any other existing work can not be applied to find such a confidence interval.

We believe, this is the first work to make developments on having sufficiently narrow confidence interval of economic inequality index based on complex household survey. Now we discuss some issues or limitations of our proposed procedures because our proposed (a) procedures depend on the pre-specified number of households in each cluster (b) sequential procedure depends on pre-specified m' (c) procedures consider large cluster size scenario (d) procedures do not consider the sampling cost and/or a fixed budget.

To begin with, the purely sequential procedure requires observations from additional m' clusters, after the first stage, every time the condition in the stopping rule is not met. Thus, there is a need to fix the value of m'. In some situations, it is as easy to collect observations from more than one cluster as it is to collect observations from a single cluster at every stage. So, as per convenience, the value of m' should be accordingly decided based on economic considerations. In fact, the purely sequential procedure is not affected by the choice of m', the larger the value of m', the fewer number of stages, and the higher the chances of overestimating the optimal number of clusters. On the other hand, the smaller the value of m', the more number of stages and the higher the chances of accurately stopping at the optimal number of clusters. Thus, there is a trade off between the number of stages and stopping accurately at the optimal cluster size when choosing m'.

Furthermore, our proposed procedures are based on the central limit theorem (when the cluster sizes per stratum are large). If the number of clusters is small, the confidence interval for Gini index cannot be constructed using Bhattacharya (2005, 2007) (fixed-cluster size method) and narrow confidence interval for Gini index using our proposed procedures. For smaller number of available clusters (H_s) for few strata, the sequence of the sampling distributions of the empirical Gini indices may not reach asymptotic limiting normal distribution. In a situation when limiting normality cannot be reached, our proposed procedures should not be applied. If one of our proposed procedures are applied, because of not having enough clusters in a few strata, one may not achieve desired confidence interval for the population Gini Index. This scenario was encountered in the application section of this work, for both the purely sequential and two-stage procedures, when there were not enough available clusters in the urban sectors, and as such, resulted in confidence intervals that were wider than desired.

Lastly, a very important question raised by the Bhattacharya (2005) was about developing a survey design taking the economic factors into account. Both our proposed procedures can be extended to include cost factors whereby optimization will be done at several levels for construction of a narrow confidence interval or confidence region under cost constraints. However, we do not explore that possibility in this article. A related issue is the fact that usually a budget is allocated by a country to its survey agency to carry out the survey. Under such budget constraints, the funding agency is not likely to willingly hand out more money if stopping rule is not met with the available amount. Without question, issue of budget constraint is important. Here, we do not discuss the estimation of cluster sizes under a fixed budget. We feel that our current work is a first step towards addressing the important issue in the sense of achieving a sufficiently narrow confidence interval or region and may

yield different outcomes under cost constraints. We believe our work will lead to further research on this topic.

9. Conclusions

Working within the asymptotic purview for complex survey data, developed by Bhattacharya (2005, 2007), we have developed purely sequential and two-stage procedures for constructing sufficiently narrow confidence intervals for the Gini index which is one of the most popular measure of economic inequality. Our procedure may be applied for surveys when stratified clustered sample data are drawn from a large number of clusters per stratum, which is a reasonable assumption to make. More so, our procedure may also be applied to special cases of multi-stage survey designs including cases without stratification (i.e., $S = 1$), and those that have independent observations within clusters (interclass correlation is zero).

It is with no doubt that the two-stage procedure is practically more feasible under this survey design than the purely sequential procedure. The confidence intervals of both procedures yielded a coverage probability closer to the desired confidence coefficient, however, the purely sequential procedure produces confidence intervals whose width are always less than the desired bound ω. The two-stage procedure is also known to over-estimate the optimal cluster size as compared to the purely sequential procedure Mukhopadhyay and De Silva (2009) and this property can be seen in results from the simulation (in the supplementary material) and the application to the NSS data. Furthermore, the estimated optimal cluster sizes have smaller standard errors under purely sequential procedure as compared to two-stage procedure.

Supplementary Materials: The following are available online at http://www.mdpi.com/2225-1146/8/2/26/s1.

Author Contributions: Conceptualization, B.C.; methodology, F.B.D. and B.C.; software, F.B.D.; validation, F.B.D., F.K. and B.C.; formal analysis, F.B.D., F.K. and B.C.; investigation, F.B.D. and B.C.; writing–original draft preparation, F.B.D., F.K. and B.C.; writing–review and editing, F.B.D., F.K. and B.C.; visualization, F.B.D., F.K. and B.C.; supervision, B.C.; project administration, B.C.; funding acquisition, B.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research work of Bhargab Chattopadhyay was part of the project sanctioned by Science and Engineering Research Board, Government of India (ECR/2017/001213).

Acknowledgments: This author is also grateful to the Ministry of Statistics and Program Implementation, Government Of India for permitting the use of the household data related to the consumer expenditure for the year 2007–2008 (Round 64, Schedule 1.0).

Conflicts of Interest: The authors declare no conflict of interest. The funding agency had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Aguirregabiria, Victor, and Pedro Mira. 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75: 1–53. [CrossRef]

Arcidiacono, Peter, and John Bailey Jones. 2003. Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica* 71: 933–46. [CrossRef]

Bhattacharya, Debopam. 2005. Asymptotic inference from multi-stage samples. *Journal of Econometrics* 126: 145–71. [CrossRef]

Bhattacharya, Debopam. 2007. Inference on inequality from household survey data. *Journal of Econometrics* 137: 674–707. [CrossRef]

Binder, David A., and Milorad S. Kovacevic. 1995. Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology* 21: 137–46.

Chattopadhyay, Bhargab, and Shyamal Krishna De. 2016. Estimation of Gini index within pre-specified error bound. *Econometrics* 4: 30. [CrossRef]

Chattopadhyay, Bhargab, and Ken Kelley. 2017. Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation. *Psychological Methods* 22: 94–113. [CrossRef]

Chattopadhyay, Bhargab, and Nitis Mukhopadhyay. 2013. Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Sequential Analysis* 32: 134–57. [CrossRef]

Cochran, William G. 1997. *Sampling Techniques*, 3rd ed. Hoboken: John Wiley & Sons. [CrossRef]

De, Shyamal K., and Bhargab Chattopadhyay. 2017. Minimum risk point estimation of Gini index. *Sankhya B 79*: 247–277. [CrossRef]

Fuller, Wayne A. 2009. *Sampling Statistics*. Hoboken: Wiley. [CrossRef]

Ghosh, Bhaskar Kumar, and Pranab Kumar Sen. 1991. *Handbook of Sequential Analysis*. New York: CRC Press, vol. 118.

Ghosh, Malay, Nitis Mukhopadhyay, and Pranab K. Sen. 1997. *Sequential Estimation*. New York: John Wiley & Sons, Inc. [CrossRef]

Greene, William H. 1998. Gender economics courses in liberal arts colleges: Further results. *The Journal of Economic Education* 29: 291–300. [CrossRef]

Gut, Allan. 2009. *Stopped Random Walks*. New York: Springer. [CrossRef]

Hoque, Ahmed Anisul, and Judith Anne Clarke. 2015. On variance estimation for a Gini coefficient estimator obtained from complex survey data. *Communications in Statistics: Case Studies, Data Analysis and Applications* 1: 39–58. [CrossRef]

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663–85. [CrossRef]

Kanninen, Barbara J. 1993. Design of sequential experiments for contingent valuation studies. *Journal of Environmental Economics and Management* 25: S1–S11. [CrossRef]

Kelley, Ken, Francis Bilson Darku, and Bhargab Chattopadhyay. 2018. Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods* 23: 226–43. [CrossRef]

Kelley, Ken, Francis Bilson Darku, and Bhargab Chattopadhyay. 2019. Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods* 24: 492–515. [CrossRef]

Langel, Matti, and Yves Tillé. 2013. Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176: 521–40. [CrossRef]

Lee, Eun, and Ronald Forthofer. 2006. *Analyzing Complex Survey Data*. New York: SAGE Publications, Inc. [CrossRef]

Mahalanobis, Prasanta Chandra. 1940. A sample survey of the acreage under jute in Bengal. *Sankhyā: The Indian Journal of Statistics* 4: 511–530.

Mukhopadhyay, Nitis, and Basil M. De Silva. 2009. *Sequential Methods and Their Applications*. Boca Raton: CRC Press.

National Sample Survey Office. 2007. Note on Estimation Procedure of NSS 64th Round. Available online: http://catalog.ihsn.org/index.php/catalog/1906/download/35538 (accessed on 21 July 2019).

National Sample Survey Office. 2015. India—Household Consumer Expenditure Survey: 64th Round, Schedule 1.0, July 2007–June 2008. Available online: http://www.icssrdataservice.in/datarepository/index.php/catalog/4/study-description (accessed on 21 July 2019).

Organization for Economic Cooperation and Development. 2017. Income Inequality. Available online: https://data.oecd.org/inequality/income-inequality.htm (accessed on 21 July 2019).

Peng, Liang. 2011. Emperical likelihood methods for the Gini index. *Australian & New Zealand Journal of Statistics* 53: 131–39. [CrossRef]

RStudio Team. 2018. *RStudio: Integrated Development Environment for R*. Boston: RStudio, Inc.

Sen, Pranab Kumar. 1988. Functional jackknifing: Rationality and general asymptotics. *The Annals of Statistics* 16: 450–69. [CrossRef]

Stein, Charles. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics* 16: 243–58. [CrossRef]

Stein, Ch. 1949. Some problems in sequential estimation. *Econometrica* 17: 77–78.

Wells, J. 1998. Applications: Oversampling through households or other clusters: Comparisons of methods for weighting the oversampled elements. *Australian & New Zealand Journal of Statistics* 40: 269–78. [CrossRef]

Zitikis, Ričardas, and Joseph L. Gastwirth. 2002. The asymptotic distribution of the S-Gini index. *Australian & New Zealand Journal of Statistics* 44: 439–46. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).