Straightforward variance estimation for the Gini coefficient

with stratified and clustered survey data

AHMED A. HOQUE\textsuperscript{+} AND JUDITH A. CLARKE\textsuperscript{++}

\textsuperscript{+}Department of Economics, University of Victoria,
PO Box 1700 STN CSC, Victoria, B.C., V8W 2Y2, Canada
Email: aahoque@uvic.ca

\textsuperscript{++}Department of Economics, University of Victoria,
PO Box 1700 STN CSC, Victoria, B.C., V8W 2Y2, Canada
Email: jaclarke@uvic.ca

Summary
Obtaining variances for the plug-in estimator of the Gini coefficient for inequality has preoccupied researchers for decades with proposed analytic formulae being cumbersome to apply, in addition to being obtained assuming an iid structure. Bhattacharya (2007, \textit{Journal of Econometrics}) provides an (asymptotic) variance when data arise from a multistage complex survey, a sampling design common with household data, often used in inequality studies. We show that Bhattacharya’s variance estimator is equivalent to an asymptotic version of that derived over a decade earlier by Binder and Kovačević (1995, \textit{Survey Methodology}). In addition, we show that Davidson’s (2009, \textit{Journal of Econometrics}) derived variance, for the iid case, is a simplification of that provided by Binder and Kovačević. Further, we consider practical ways to calculate the plug-in Gini estimator and its asymptotic variance, along with providing an empirical illustration of our results.

Keywords: Inequality; Asymptotic inference; Gini index; Complex survey

1. INTRODUCTION

Arguably the best known and most widely employed measure of inequality, the Gini coefficient (GC), proposed by Corrado Gini in 1914, has been the focus of many theoretical and empirical studies. For instance, the GC is reported extensively as a way to rank countries in terms of income inequality by, for example, the United Nations, the World Bank and the Central Intelligence Agency. Two recent empirical studies that adopt the GC, among the many that could be listed, are Nho (2006), who considers regional income inequality for Korean households, and Slater et al. (2009), who examine the prevalence of being overweight and obese in Canadian adults across a range of socio-economic and geographic groupings.
Both of these cited works form GCs from sample data obtained from multistage stratified cluster designs, so-called complex surveys, with stratification and clustering used to guarantee representation of groups of interest as well as to keep costs as low as possible\(^1\). The design of such a sample, which cannot be assumed to be a randomly drawn \textit{iid} sample, needs to be accounted for when forming both the GC estimator and an associated variance. Such applications motivate our work.

Obtaining a variance for a GC estimator has been investigated by many researchers including, but not limited to, Hoeffding (1948), Glasser (1962), Sendler (1979), Beach and Davidson (1983), Gastwirth and Gail (1985), Schechtman and Yitzhaki (1987), Sandström et al. (1985, 1988), Nygård and Sandström (1989), Yitzhaki (1991), Shao (1994), Binder and Kovačević (1995), Bishop et al. (1997), Karagiannis and Kovačević (2000), Giles (2004), Modarres and Gastwirth (2006), Bhattacharya (2007), Xu (2007), Davidson (2009) and Qin et al. (2010). Some of these studies propose analytic (asymptotic) variances while others adopt resampling methods, with the latter works often claiming that such tools are preferable, as they avoid the mathematical and coding complexities associated with the available analytic expressions\(^3,4\). The extent of studies, even for our incomplete list, suggests that there should be nothing more to say about a seemingly simple task of providing a variance for a sample statistic.

However, the research that provides analytic asymptotic variances has adopted several different methods, with links between them and the resulting formulae often unclear. For instance, some obtain asymptotic variances based on \textit{U}-statistics (e.g., Hoeffding, 1948; Yitzhaki, 1991; Bishop et al. (1997); Xu, 2007) whereas others use \textit{L}-statistic theory (e.g., Nygård and Sandström, 1989; Shao, 1994). Via the use of Taylor-series expansions, Binder and Kovačević (1995) and Davidson (2009) provide approximation expressions for GC estimators, from which they obtain variances; Binder and Kovačević (1995) allow for complex survey sample data whereas Davidson (2009) assumes an \textit{iid} sample. The use of estimating equations underlies the work of Binder and Kovačević (1995) (see also Kovačević and Binder, 1997), who, after generating an appropriate asymptotic approximation expression, appeal to standard survey theory to provide a so-called linearization variance. Coding for this variance is, in our

\(^{1}\) See, for instance, Cochran (1977), Skinner et al. (1989) and Cameron and Trivedi (2005, Chapter 24).

\(^{2}\) Nho (2006)’s work, for example, uses household data from the Korean National Survey of Household Incomes and Expenditures, which stratifies the country according to geographical regions and administrative districts and then forms clusters based on Census enumeration areas. Clusters are selected using probability sampling proportional to size and then two segments of five households are randomly selected from the sampled clusters.

\(^{3}\) Ogwong (2000, p123) provides one such example, stating, to justify his jackknife approach, that “standard errors of the Gini index that have been suggested so far ... are either mathematically very complicated or require heavy computation which cannot be conveniently undertaken using commonly available regression software packages”.

\(^{4}\) Empirical work has also often followed such recommendations; for instance, Nho (2006, p341) states “The standard error of the gini index is estimated by jackknife and bootstrap methods rather than the traditional delta method because of the complexity of the latter.”
view, not complicated, especially with access to software that accounts for survey design\textsuperscript{5}. Although Davidson (2009) limits attention to an iid random sample, he approaches variance estimation in a similar fashion by also deriving an approximation for the GC’s estimator, which he uses to suggest a variance estimator. Bhattacharya (2007), based on his earlier, more general, paper (Bhattacharya, 2005) frames estimation of the GC, with complex survey sample data, within generalized method of moments (GMM) theory, and appeals to available results for independent nonidentically distributed random variables to show the consistency and asymptotic normality of a plug-in estimator. Using sample empirical process theory and the functional delta method, Bhattacharya (2007) derives an expression for the asymptotic variance of the GC estimator, taking account of the sample design. One notable (of the many) contribution from Bhattacharya’s work is the breakdown of the variance into three components: the estimate of the variance without taking the sample design into account; the impact of clustering on the variance; and the stratum effect on the variance. The resulting variance formula, to quote Davidson (2009, p30), “is however not at all easy to implement”\textsuperscript{6}.

We focus on three works: Binder and Kovačević (1995), Bhattacharya (2007) and Davidson (2009). First, we show that the approximation obtained by Bhattacharya (2007) is algebraically equivalent to that provided by Binder and Kovačević (1995), over a decade earlier. Secondly, we show that Davidson’s (2009) derived approximation is a special example of Binder and Kovačević’s (1995) expression. That this results is not surprising, as Davidson notes (p32) that the random variable in his approximation “can with some effort be shown to be the same as that used by Bhattacharya (2007)”; the difference here is that Davidson’s approximation follows from that of Binder and Kovačević (1995) with very little effort. We believe that it is important to note the connections of the recent econometrics works (Bhattacharya, 2007; Davidson, 2009) with an earlier work, from the survey literature, which seems to have slipped their attention.

We then consider variance estimators, showing that Bhattacharya’s and Binder and Kovačević’s variances for the GC estimator are equivalent, at least asymptotically, with the asymptotic analysis referring to the number of sampled clusters in each stratum going to infinity at the same rate. This outcome is expected given the equivalence of the approximations for the estimator. That the two works lead to the same variance is most useful, as calculating Binder and Kovačević’s variance is straightforward, especially packages designed to account for sampling design. We also show that Davidson’s variance estimator is a special case of that provided by Binder and Kovačević.

\textsuperscript{5} For econometricians, Stata is likely the best known such package. Jenkins (2006) provides a Stata add-on entitled \texttt{svylorenz} that readily produces Binder and Kovačević’s variance estimator.

\textsuperscript{6} In addition, our reading of Bhattacharya’s work suggests some typographical errors that further complicate use of the results. We have attempted to correct these in our presentation.
Our final contribution is to show how auxiliary regressions can be used to obtain variances without the need for specialized survey software. Although using artificial regressions for estimating the GC has been considered in the literature (Ogwong, 2000; Giles, 2004; Davidson, 2009), this has been limited to data presumed to arise under an iid assumption, whereas our regressions allow for the complex survey design. We also suggest ways to compute variance estimators, often regarded as burdensome for the GC.

This paper is organized as follows. Section 2 briefly considers sampling designs and weighting. We consider estimation of the GC and approximations for the estimator in Section 3. Our main results are presented in Section 4, where we examine (asymptotic) variance (hereafter, variance) estimation of the usual plug-in estimator of the GC using data obtained under a complex survey design. In this Section 4, we also examine variance estimation with a random iid sample. Section 5 considers practically calculating estimators and we provide a brief empirical application in Section 6. We conclude in Section 7 and provide algebraic proofs in an Appendix.

2. SAMPLING DESIGNS AND WEIGHTS

There are many sampling designs with effects on estimation of a population parameter discussed in standard statistics texts; e.g., Cochran (1977) and Wolter (2007). Our focus is on multistage complex sampling, often adopted when obtaining household data, a sampling design that may involve one or more combinations of sampling techniques, with the key outcome being that the sample cannot be regarded as iid. Ignoring the sample design (such as behaving as if the sample is iid) can lead to inconsistent population parameter estimators and inconsistent variances for these estimators. Some sampling designs include stratification, clustering, double sampling, multiple frames, poststratification and so on; see, e.g., Wolter (2007). Of particular interest, is first stage stratification and second stage clustering, as it turns out that further sampling stages do not affect the variance estimator, which is computed from quantities formed from the ultimate clusters (see, e.g., Skinner et al., 1989, p47).

Stratifying divides a population into relatively homogenous subgroups before sampling (e.g., area of residence, gender and race) with sample selection then proceeding separately for each stratum. Such a design typically breaks down the identical part of an iid assumption. The independent component, on the other hand, is usually violated with clustering, which splits the population into contiguous groupings; e.g., villages in rural areas and blocks or enumeration areas in urban areas. Given this contiguity, units within clusters are usually correlated.

Our notation with respect to the design follows. Let \( \{U[hci]: h = 1, \ldots, H; c = 1, \ldots, N_h; i = 1, \ldots, M_{hc}\} \) be a finite population stratified into \( H \) strata, with \( N_h \) clusters or primary sampling units within
each stratum so that the population consists of $N = \sum_{h=1}^{H} N_h$ clusters. In cluster $c$, within stratum $h$, there are $M_{hc}$ units, which we assume to be households, leading to $M_h = \sum_{c=1}^{N_h} M_{hc}$ households in stratum $h$; the population number of households is then $M = \sum_{h=1}^{H} \sum_{c=1}^{N_h} M_{hc}$. Usually, a sample of clusters is chosen using probability proportional to cluster size\(^7\), with a fixed number of households then selected with replacement\(^8\) from each cluster using simple random sampling.

We suppose that a sample of $n_h$ clusters is drawn from the $h$’th stratum, with the total number of sampled clusters being $N = \sum_{h=1}^{H} n_h$. From cluster $c$, within stratum $h$, we suppose that $m$ households are selected\(^9\), such that the total number of sampled households is $M = m \sum_{h=1}^{H} n_h = mN$. If $s_{hci}$ is the number of members in the $i$’th household in the $c$’th cluster within the $h$’th stratum, then the total number of individuals in the sample is $n_0 = \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} s_{hci}$.

Given the complex survey’s design, along with the common practice of oversampling particular subgroups to ensure stable estimates, households and individuals in the population likely will not have the same probability of being included in the sample stages, a feature that is accounted for by a sampling weight for each observation that denotes how many individuals this observation represents in the population. The weight, which is inversely proportional to selection probabilities, is

$$W_{hci} = \frac{M_{hc} N_h}{mn_h s_{hci}},$$

which is often normalized as:

$$w_{hci} = \frac{W_{hci}}{\sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} W_{hci}},$$

such that $\sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} = 1$\(^{10}\).

### 3. GC ESTIMATION AND ASYMPTOTIC APPROXIMATIONS

In subsection 3.1, we define the GC, providing its natural plug-in estimator when sample data are obtained from a multistage complex survey. We also give the formula for the estimator considered when

---

\(^{7}\) That is, the probability of selecting a sampling unit is proportional to the number of units in the cluster.

\(^{8}\) Sampling with or without replacement may not be relevant when the sample size is small relative to the population. Moreover, with Bhattacharya’s (2005, 2007) asymptotic analysis, this distinction is not of import; see, e.g., Bhattacharya (2005, p148).

\(^{9}\) The assumption that the same number of households is sampled from each cluster can be relaxed.

\(^{10}\) Sampling weights also usually account for other features such as different nonresponse rates of household interviews.
the sample is regarded as \(iid\), randomly drawn from an underlying population. We follow in subsection 3.2 with a consideration of approximations for the estimator, showing that Binder and Kovačević (1995) and Bhattacharya (2007) obtain equivalent expressions. We also show that the approximation of Davidson (2009) is a special example of that derived by Binder and Kovačević (1995).

### 3.1. Gini coefficient and plug-in estimators

The Gini coefficient, bounded by zero and one, is typically defined as twice the area between the 45° line and the Lorenz (1905) curve. The Lorenz curve graphically illustrates the distribution of the well-being variable (e.g., household income, education attainment) by displaying the cumulative share of the well-being variable against its recipient share. Specifically, for a random well-being variable \(y \in [0, \infty)\) with cumulative distribution function \(F(y)\) and finite non-zero mean \(\mu = \int_0^\infty y dF(y)\), the Lorenz curve is

\[
L(p) = \frac{1}{\mu} \int_0^{z_p} y dF(y),
\]

where \(z_p = F^{-1}(p) = \inf\{y | F(y) \geq p\}\) is the \(p\)th quantile or fractile of the distribution function with \(p = F(z_p) = \int_0^{z_p} dF(y), 0 \leq p \leq 1\). On the 45° line, the line of equality, \(p = L(p)\), whereas there is inequality when \(p > L(p)\). Given expression (3.1), the GC is then commonly defined as

\[
G = 1 - 2 \int_0^1 L(p) dp = \frac{2}{\mu} \int_0^\infty y F(y) dF(y) - 1.
\]

The GC, a summary measure of the degree of inequality, is zero (one) for a perfectly equal (unequal) distribution. To proceed, let \(\hat{F}(y_{hc\ell})\) be the empirical distribution function for \(y_{hc\ell}\), which given the sampling design, is

\[
\hat{F}(y_{hc\ell}) = \frac{1}{\sum_{r=1}^H \sum_{s=1}^{n_r} \sum_{t=1}^m w_{rst}} \sum_{r=1}^H \sum_{s=1}^{n_r} \sum_{t=1}^m \hat{w}_{rst} I[y_{rst} \leq y_{hc\ell}] = \sum_{d=1}^M w_d I[y(d) \leq y_{hc\ell}],
\]

where \(y(d)\) is the \(d\)th order statistic in the full sample and \(w_d\) its associated sampling weight. Although not necessary, it is sometimes helpful to write expressions in terms of order statistics, as it provides consistency with some of the related research. Then, let \(\hat{z}(p)\) be the estimated sample quantile (i.e., \(\hat{z}(p) = \inf\{y_{hc\ell} | \hat{F}(y_{hc\ell}) \geq p\}\)) and \(\hat{\mu} = \frac{1}{\sum_{h=1}^H \sum_{c=1}^{n_h} \sum_{t=1}^m w_{hc\ell} y_{hc\ell}}\) be the estimator of \(\mu\), such that we obtain the following estimator for \(L(p)\), allowing for the sampling design:

---

Hoque and Clarke (2012)  
Page 6
Variance estimation for the Gini coefficient

\[ \hat{L}(p) = \frac{1}{\tilde{\mu}} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} I(y_{hci} \leq \hat{z}(p)) = \frac{\hat{\alpha}(p)}{\tilde{\mu}} , \]

with \( \hat{\alpha}(p) \) being an estimator of \( \alpha(p) = E \left( Y I(Y \leq z(p)) \right) \). Using this, we have

\[ \hat{\alpha} = 1 - 2 \int_{0}^{1} \hat{L}(p) dp = 1 - \frac{2}{\tilde{\mu}} \int_{0}^{1} \left[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} I(y_{hci} \leq \hat{z}(p)) \right] dp \]

\[ = 1 - \frac{2}{\tilde{\mu}} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} \left[ \int_{0}^{1} I(y_{hci} \leq \hat{z}(p)) dp \right] , \]

which can be estimated by

\[ \hat{\alpha} = 1 - \frac{2}{\tilde{\mu}} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} \left[ \int_{0}^{1} I(y_{hci} \geq \hat{z}(p)) dp \right] = 1 - \frac{2}{\tilde{\mu}} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} \left( 1 - \hat{F}(y_{hci}) \right) . \]

This is the estimator adopted by Bhattacharya (2007). Re-arranging, we arrive at Binder and Kovačević’s (1995) estimator:

\[ \hat{\alpha} = 2 \tilde{\mu} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} y_{hci} \hat{F}(y_{hci}) - 1 . \]

(3.4)

For an iid randomly drawn sample of size \( M \), this result suggests using

\[ \hat{\alpha} = \frac{2}{M \tilde{\mu}} \sum_{d=1}^{M} y_d \hat{F}(y_d) - 1 , \]

(3.5)

where \( \hat{F}(y_d) = \frac{1}{M} \sum_{j=1}^{M} I(y_j \leq y_d) \). However, noting that the second expression in (3.2) can be equivalently written as \( \hat{G} = \frac{1}{\mu} \int_{0}^{\infty} y d(F(y))^2 - 1 \), and using order statistics, Davidson (2009) suggests the following estimator for \( \hat{G} \), which overcomes the problem of defining the empirical distribution function as being either right- or left-continuous:

\[ \hat{G} = \frac{2}{M \tilde{\mu}^2} \sum_{d=1}^{M} y_{d} \hat{F}(y_d) \left( d - 1 \right) - 1 . \]

(3.6)

In this formula, an average of the lower and upper limits is used for \( \hat{F} \). Expression (3.6) is equivalent to that considered by, amongst others, Sendler (1979), Nygård and Sandström (1989), Ogwong (2000) and Giles (2004).

Hoque and Clarke (2012)
3.2 Approximations for $\hat{G}$

We now turn to obtaining approximate expressions for $\hat{G}$ from which we can obtain variance estimators. Bhattacharya (2007) frames estimation as a method of moment problem, showing that an approximate expression for the Lorenz share at a fixed percentile $p$, with $\theta = (z(p), \alpha(p), \mu)$ is given by

$$ L(p) - L(p) \approx \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} m(y_{hci}, \theta) , \quad (3.7) $$

where

$$ m(y_{hci}, \theta) = \frac{1}{\mu} (y_{hci} l_{hci} (y_{hci} \leq z(p)) - \alpha(p)) + \frac{1}{\mu} (z(p) - \alpha(p)) (y_{hci} \leq z(p)) + \frac{\alpha(p)}{\mu^2} (\mu - y_{hci}) $$

$$ = \frac{1}{\mu} [y_{hci} l_{hci} (y_{hci} \leq z(p)) + z(p) (y_{hci} \leq z(p)) - \frac{\alpha(p)}{\mu} y_{hci}] . $$

Then, the approximate expression for $\hat{G}$ is

$$ \hat{G} - G \approx -2 \int_{0}^{1} dp \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} m(y_{hci}, \theta) $$

$$ = -2 \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} \left[ \int_{0}^{1} m(y_{hci}, \theta) dp \right] $$

$$ = \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} \Psi_{hci} . \quad (3.8) $$

An estimator of $\Psi_{hci}$ is:

$$ \hat{\Psi}_{hci} = -2 \sum_{d=1}^{M} w_d \left\{ \frac{1}{\mu} \left[ y_{hci} l_{hci} (y_{hci} \leq y(d)) + y(d) \left( \hat{f}(y(d)) - I(y_{hci} \leq y(d)) \right) - \frac{\hat{a}(y(d))}{\mu} y_{hci} \right] \right\} , $$

where $\hat{a}(y(d)) = \sum_{r=1}^{H} \sum_{s=1}^{n_r} \sum_{t=1}^{m} w_{rst} y_{rst} I(y_{rst} \leq y(d))$ is used to estimate $\hat{a}(p)$. We then have the approximation presented by Bhattacharya (2007):

$$ \hat{G} - G \approx $$
Variance estimation for the Gini coefficient

\[
\sum_{h=1}^{H} \sum_{c=1}^{n_{h}} \sum_{i=1}^{m} w_{hci} \left( -2 \sum_{d=1}^{M} w_{d} \left[ \frac{1}{\hat{\mu}} \left( y_{hcl}(y_{hcl} \leq y_{(d)}) + y_{hcl}I(y_{hcl} \leq y_{(d)}) - \frac{\hat{\alpha}(y_{(d)})}{\hat{\mu}} \right) \right] \right). 
\] (3.9)

It should be noted that we have written Bhattacharya’s approximate expression using the traditional summing over stratum, clusters and households, in contrast to Bhattacharya who rearranges the summing to take account of his asymptotic analysis, which is with respect to the number of clusters\(^{11}\). Specifically, let \( a_{h} = \frac{n_{h}}{n} \) and \( I(c \in h) = 1 \) when cluster \( c \) is in stratum \( h \), 0 otherwise. Then, let \( W_{hci}^{*} = NW_{hci} \)

\[
eq (I(c \in h)) \times \frac{m_{h} n_{h}}{m_{a_{h}}} s_{hci} \quad \text{and} \quad W_{hci}^{*} = W_{hci} / \sum_{h=1}^{H} \sum_{c=1}^{n_{h}} \sum_{i=1}^{m} W_{hci} \quad \text{so that we write (3.8) as}
\]

\[
\hat{G} - G \approx
\]

\[
- \frac{2}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \sum_{i=1}^{m} W_{hci}^{*} \left[ \int_{0}^{1} \frac{dp}{\mu} \left\{ \frac{1}{\hat{\mu}} \left[ y_{hcl}(y_{hcl} \leq z(p)) + z(p) \left( p - I(y_{hcl} \leq z(p)) \right) - \frac{\hat{\alpha}(p)}{\hat{\mu}} y_{hcl} \right] \right\} \right].
\]

Re-arranging, we obtain Bhattacharya’s (2007, p685) approximation for \( \hat{G} \):

\[
\sqrt{N}(\hat{G} - G) \approx - \frac{2}{\sqrt{N}} \sum_{c=1}^{N} \Psi_{c}
\] (3.10)

where

\[
\Psi_{c} = \int_{0}^{1} \frac{dp}{\mu} \left\{ \sum_{h=1}^{H} \sum_{i=1}^{m} w_{hci}^{*} \left( \frac{1}{\mu} \left[ y_{hcl}(y_{hcl} \leq z(p)) + z(p) \left( p - I(y_{hcl} \leq z(p)) \right) - \frac{\hat{\alpha}(p)}{\mu} y_{hcl} \right] \right) \right\}.
\]

We now consider Binder and Kovačević’s (1995) approximation for \( \hat{G} \). Binder and Kovačević (1995) approach estimation using estimating equations\(^{12}\), a general way to estimate population parameters. Some examples of methods that lead naturally to estimating equations are maximum likelihood, method of moments and least squares. To illustrate, following Binder and Kovačević (1995), suppose interest lies in estimation of a finite population parameter \( \lambda \) that can be written as the solution to

\(^{11}\) As many household surveys have many more clusters sampled per stratum than the number of strata or the number of households sampled per cluster, it seems sensible to consider an asymptotic framework with the number of clusters going to infinity, holding the number of strata and households per cluster fixed and finite. This is Bhattacharya’s framework. In contrast, Krewski and Rao (1981), Rao and Wu (1985), Binder and Kovačević (1995) and Kovačević and Binder (1997), amongst others, consider asymptotic analysis with the number of strata tending to infinity, assuming the number of clusters per stratum is fixed.

\(^{12}\) First proposed by Godambe (1960); see also Godambe and Thompson (1978, 1984), Binder (1991) and Binder and Patak (1994).
Variance estimation for the Gini coefficient

\[ \int u(y, \lambda) dF(y) = 0 \]

where \( F(y) \) is the finite population distribution function. Then, with \( \hat{F}(y) \) being a consistent empirical distribution function, the estimating equations estimator of \( \lambda \) is that value of \( \hat{\lambda} \) such that

\[ \int \hat{u}(y, \hat{\lambda}) d\hat{F}(y) = 0 \]

where \( \hat{u}(y, \lambda) \), the estimating equation, is an estimate or approximation of \( u(y, \lambda) \), which may or may not be needed in practise. Given these definitions, it is straightforward to see how moment equations and likelihood equations are examples of estimating equations. Not surprisingly, sometimes more than one estimating equation is required. Using this method, Binder and Kovačević (1995) show that the following two estimating equations are required for estimation of the Lorenz curve ordinate and the 100th percentile of the distribution:

\[ u_1(y, L(p)) = I(y \leq z(p)) y - L(p) y \quad \& \quad u_2(y) = I(y \leq z(p)) - p \]

Using these estimating equations and approximations, based on theorems from Francisco and Fuller (1991), Binder and Kovačević (p141) obtain an approximation for the Lorenz share at a fixed percentile \( p \):

\[ \hat{L}(p) - L(p) \approx \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} u(y_{hci}, \theta) \]

where

\[ u(y_{hci}, \theta) = \frac{1}{\mu} (y_{hci} - z(p)) I(y_{hci} \leq z(p)) + p z(p) - y_{hci} \alpha(p)/\mu \]

By simple inspection, \( u(y_{hci}, \theta) = m(y_{hci}, \theta) \). Given the equivalence of these approximations, the approximate expressions of Binder and Kovačević (1995) and Bhattacharya (2007) for \( \hat{G} \) are also the same. It is, nevertheless, useful to show this result, as we find Binder and Kovačević’s (1995) expression for the approximation to be more convenient in practice. Binder and Kovačević (1995, p143) provide the sample estimating equation for the GC:

\[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci}[(2\hat{F}(y_{hci}) - 1)y_{hci} - \hat{G} y_{hci}] = 0 \]

and estimated expression for their approximation for \( \hat{G} \)
\[ \hat{G} - G \approx \]
\[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} \left\{ \frac{2}{\hat{\mu}} y_{hci} \left( \hat{F}(y_{hci}) - \frac{(\hat{G} + 1)}{2} \right) + B(y_{hci}) - \frac{\hat{\mu}}{2} (\hat{G} + 1) \right\} \]
\[ = \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} u_{hci}^* \], (3.12)
where
\[ B(y_{hci}) = \sum_{r=1}^{H} \sum_{s=1}^{n_r} \sum_{t=1}^{m} w_{rst} y_{rst} I(y_{rst} \geq y_{hci}) \cdot \]

The equivalence of expressions (3.9) and (3.11) is shown in the Appendix under the Proof of Result 1.

When calculating variances, considered in the next section, practitioners need to generate either \( u_{hci}^* \) or \( \Psi \). Although equivalent terms, having already formed \( \hat{\mu}, \hat{F}(y_{hci}) \) and \( \hat{G} \), it is our belief that it is computationally easier to code \( u_{hci}^* \) rather than \( \Psi \), as \( \hat{\mu} \) is the only required additional term.

To end this subsection, we simplify Binder and Kovačević’s (1995) expression (3.11) for the case of a randomly drawn iid sample, showing that it straightforwardly gives that obtained by Davidson (2009). In the iid case, expression (3.11) becomes:
\[ G - G \approx \frac{1}{M} \sum_{j=1}^{M} 2 \left\{ y_j \left( \hat{F}(y_j) - \frac{(\hat{G} + 1)}{2} \right) + B(y_j) - \frac{\hat{\mu}}{2} (\hat{G} + 1) \right\} \], (3.13)
where \( B(y_j) = \frac{1}{M} \sum_{i=1}^{M} y_i I(y_i \geq y_j) \). Using Davidson’s notation, \( B(y_j) = \hat{\mu} - \hat{m}(y_j), \frac{1}{\hat{\mu}} = \frac{(\hat{G} + 1)}{2} \) and \( \hat{l} = \frac{\hat{\mu}}{2} (\hat{G} + 1) \), where \( I = \int_0^\infty y F(y) dy \). Making these substitutions, with some minor algebraic manipulations, we obtain
\[ \hat{G} - G \approx \frac{1}{M} \sum_{d=1}^{M} \left\{ - \hat{l} (y_d - \hat{\mu}) + y_d \hat{F}(y_d) - \hat{m}(y_d) - (2 \hat{l} - \hat{\mu}) \right\} \], (3.14)
which is Davidson’s (p32) estimated approximation for \( \hat{G} \). Note that we can equivalently write (3.13) as:
\[ \hat{G} - G \approx \frac{1}{M} \sum_{d=1}^{M} u_d^* \],
where

Hoque and Clarke (2012)
\[ u_d^* = \frac{2}{\hat{\mu}} \left[ y_d \left( \hat{\bar{y}}(y_d) - \left( \frac{\hat{G} + 1}{2} \right) \right) + B(y_d) - \frac{\hat{\mu}}{2} (\hat{G} + 1) \right] \]  
\[ = \frac{(\hat{Z}_d - \bar{Z})}{\hat{\mu}}, \]

in Davidson’s adopted notation, with \( \hat{Z}_d = -(\hat{G} + 1)y_d + 2 \left( y_d \hat{\bar{y}}(y_d) - \hat{\bar{m}}(y_d) \right) \) and \( \bar{Z} = \hat{\mu}(\hat{G} - 1) \).

In this section, we have shown that recent derivations of approximate expressions for \( \hat{G} \) by Bhattacharya (2007) and Davidson (2009) are either equivalent to or a special example of the expression obtained over a decade earlier by Binder and Kovačević (1995). We now turn to variance estimators for \( \hat{G} \).

4. VARIANCE ESTIMATORS

We first present the variance estimator of Binder and Kovačević (1995), which is based on standard survey theory, and then consider its limiting form as the number of clusters goes to infinity. We then move to Bhattacharya’s (2007) variance estimator, showing that his computationally tedious formula, although informative from a sampling point of view, is equivalent to the limiting form of Binder and Kovačević’s (1995) variance, a formula that is far easier to use in practise. Finally, we show that Davidson’s (2009) variance formula, suggested for the \( \text{iid} \) case, is an example of the complex survey variance of Binder and Kovačević (1995). That this holds is expected from our previous results of the equivalence of respective approximations for \( \hat{G} \).

Using expression (3.12), an estimator of the variance of \( \hat{G} \) is

\[ \text{Var}(\hat{G}) = \text{Var} \left( \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hc} u_{hc,i} \right), \]

which is merely the variance of a survey weighted total, a well-discussed estimation problem in the survey literature. For instance, following Skinner et al. (1989, p47)\(^{13}\), the standard nonparametric estimator is:

\[ \text{Var}(\hat{G}) = \frac{1}{\hat{n}_h} \left( \frac{n_h}{n_h - 1} \right) \sum_{c=1}^{n_h} (u_{hc}^* - \bar{u}_h^*)^2, \]  
\[ (4.1) \]

where \( u_{hc}^* = \sum_{i=1}^{m} w_{hc} u_{hc,i} \) and \( \bar{u}_h^* = (1/n_h) \sum_{c=1}^{n_h} u_{hc}^* \). As \( (n_h/(n_h - 1)) = (a_h/(a_h - (1/n))) \to 1 \) as \( n_h, n \to \infty \), an asymptotically\(^{14}\) equivalent estimator is

\(^{13}\) This assumes that primary sampling units, the clusters, are selected with replacement.
Variance estimation for the Gini coefficient

\[ V\text{ar}(\hat{g}) = \sum_{h=1}^{H} \sum_{c=1}^{n_h} (u_{hc}^* - \bar{u}_h)^2. \]  

(4.2)

Bhattacharya (2007), on the other hand, based on his general paper Bhattacharya (2005), provides the asymptotic variance estimator:

\[
V\text{ar}(\hat{g}) = \text{Var}\left( \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} \Phi_{hci} \right) \\
= \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci}^2 \Phi_{hci}^2 + \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} w_{hci'} \Phi_{hci} \Phi_{hci'} \\
- \sum_{h=1}^{H} \frac{1}{n_h} \left( \sum_{c=1}^{n_h} \sum_{h=1}^{m} w_{hci} \bar{\Phi}_{hci} \right)^2
\]

(4.3)

where, from Section 3,

\[ \bar{\Phi}_{hci} = -2 \sum_{d=1}^{M} w_d \left\{ \frac{1}{\hat{\mu}} y_{hci} I(y_{hci} \leq y_{(d)}) + y_{(d)} \left( \hat{f}(y_{(d)}) - I(y_{hci} \leq y_{(d)}) \right) - \frac{\hat{\alpha}(y_{(d)})}{\hat{\mu}} y_{hci} \right\}. \]

This is equivalent to \( u_{hci}^* \) (recall Result 1 in the Appendix). Bhattacharya’s expression (4.3) is useful for understanding the impact of the sampling design on variance estimation; the first term is the estimator of the variance under a simple random sampling design (or iid assumption) with weights, the second term is the effect on the variance from clustering and the third term is the impact of stratification on the variance. As we would expect a positive covariance between values obtained from the same cluster, the cluster effect is expected to be positive, raising the variance over that which would arise under an iid with weights assumption. Stratification reduces the variance; the more homogeneous are the units within a stratum and the more heterogeneous are units across strata, the higher would be this stratification effect.

Despite the benefits of writing the variance as expression (4.3), it is not especially friendly for practitioners to code, whereas the form of expression (4.1) (or (4.2)) is simpler to practically calculate.

Not surprisingly, given our earlier findings, Bhattacharya’s variance (4.3) is equivalent to the limiting form of Binder and Kovačević’s variance, expression (4.2). We show this as Proof of Result 2 in the Appendix. Aside from showing the equivalence of Bhattacharya’s recent result with that of one derived over a decade ago, this outcome is beneficial for practitioners as it turns out to be relatively easy to

\[ ^4 \text{As per Bhattacharya (2005, 2007), the asymptotic behaviour is with respect to the number of sampled clusters for each stratum going to infinity at the same rate, leading to } \alpha_h \text{ remaining fixed.} \]
estimate expression (4.2) (and (4.1) should a researcher wish to adopt the standard survey sampling variance). Estimation is discussed in the next section.

Prior to doing so, we consider the case of an iid sample. The natural estimator from expression (4.2) is

$$\widehat{\text{Var}}(\hat{G}) = \frac{1}{M^2} \sum_{d=1}^{M} (u_d^* - \bar{u})^2 = \frac{1}{M^2} \sum_{d=1}^{M} u_d^*$$

(4.4)

where $u_d^*$ is as defined in expression (3.15) and using that $\bar{u}^* = \frac{1}{M} \sum_{d=1}^{M} u_d^* = 0$. In terms of Davidson’s (2009) notation, this is equivalent to $\widehat{\text{Var}}(\hat{G}) = \frac{1}{(M\bar{y})^2} \sum_{d=1}^{M} (\hat{Z}_j - \bar{Z})^2$, which is explicitly his equation (19) on p32. That is, as expected, Davidson’s proposed variance estimator is a special case of that suggested by Binder and Kovačević (1995).

5. CALCULATING ESTIMATES IN PRACTICE

In this section we discuss straightforward ways of practically obtaining the GC estimator and its associated variance in practice. In subsection 5.1 we examine obtaining $\hat{G}$ for both the complex survey and iid sample and in subsection 5.2 we consider estimating $\widehat{\text{Var}}(\hat{G})$ with a complex survey sample. We do not discuss calculating $\widehat{\text{Var}}(\hat{G})$ for an iid sample as this is just a scaled sum of squares.

5.1 Computing $\hat{G}$

When data are from an iid random sample, it has been show that $\hat{G}$ can be easily obtained from an artificial ordinary least squares (OLS) regression; see Ogwong (2000), Giles (2004) and Davidson (2009). Specifically, with unordered data, estimating the artificial regression:

$$(2\hat{F}(y_d) - 1)\sqrt{y_d} = \theta \sqrt{y_d} + v_d, \quad d = 1, \ldots, M$$

by OLS results in $\hat{\theta} = \hat{G}$. If data are ordered with $\hat{F}$ computed using the average of the lower and upper limits, as advocated by Davidson (2009), the artificial regression:

$$\left(\frac{2d}{M} - \frac{1}{M} - 1\right)\sqrt{y_d} = \theta \sqrt{y_d} + v_d, \quad d = 1, \ldots, M$$

estimated by OLS results in $\hat{\theta} = \hat{G}$ as defined by expression (3.6).

---

This also holds for the complex survey; i.e., $\sum_{h=1}^{H} \sum_{c=1}^{N_h} \sum_{i=1}^{m} w_{hec} u_{hec} = 0$.
A similar approach can be adopted with a complex survey. In these auxiliary regressions with complex survey data the sampling weights are assumed to be normalized such that
\[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} = 1. \]
Should a researcher be using a software package not explicitly aimed at handling sampling design, the OLS estimator of \( \theta \) in the artificial regression:
\[ (2\hat{F}(y_{hci}) - 1)\sqrt{w_{hci}y_{hci}} = \theta \sqrt{w_{hci}y_{hci}} + u_{hci}, \quad h = 1, \ldots, H; c = 1, \ldots, n_h; i = 1, \ldots, m \]
leads to \( \hat{G} \) given in expression (3.4); the data need not be ordered for this regression. That is, we simply estimate the OLS regression over all data ignoring the sampling design. If access is available to software that accounts for survey design, regression of:
\[ (2\hat{F}(y_{hci}) - 1)\sqrt{y_{hci}} = \theta \sqrt{y_{hci}} + u_{hci}, \quad h = 1, \ldots, H; c = 1, \ldots, n_h; i = 1, \ldots, m \]
having declared appropriate elements of the sampling design yields \( \hat{G} = \hat{\theta} \).

5.2 Computing \( \text{Var}(\hat{G}) \) with a complex survey sample

If a researcher is using a package designed for surveys\(^{16}\), then it is easy to calculate Binder and Kovačević’s (1995) variance estimator
\[ \text{Var}(\hat{G}) = \sum_{h=1}^{H} \left( \frac{n_h}{n_h - 1} \right) \sum_{c=1}^{n_h} (u_{hci}^* - \bar{u}_h^*)^2, \]
as this is the variance for a survey weighted total based on the estimated approximation
\[ \hat{G} - G \approx \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hci} u_{hci}^*. \]

For such software packages, the variance (often called the linearization variance) is easily generated using an appropriate “total” command after forming the series \( u_{hci}^* \), along with specifying the weight series and declaring strata and cluster identification variables. When each stratum contains a large number of clusters (as would be the case for many household surveys) there will be little difference between this estimator and its asymptotic version (i.e., that associated with the number of sampled clusters going to infinity at the same rate):

\(^{16}\) Examples include Stata (StataCorp), SPSS (SPSS Inc.), SUDAAN (Research Triangle Institute) and the “survey” package developed by Lumley for R (see Lumley, 2010).
Variance estimation for the Gini coefficient

\[ \text{Var}(\hat{G}) = \sum_{h=1}^{H} \sum_{c=1}^{n_{hc}} (u_{hc} - \bar{u}_{h})^2, \]

which, we recall, is the same as the variance derived by Bhattacharya (2007) (as shown via the Proof of Result 2 in the Appendix). However, should stratum contain few clusters, then it may be preferable to use Binder and Kovačević’s (1995) variance estimator, as the scaling factor \( (n_h/(n_h - 1)) \) may be useful for such a dataset.

When access is not available to survey software, one way to obtain Bhattacharya’s variance estimator is to estimate the artificial regression:

\[ u_{hc}^* = \sum_{h=1}^{H} \beta_h D_{hc} + v_{hc}, \]

where \( D_{hc} = 1 \) if cluster \( c \) is in stratum \( h \), 0 otherwise. Let \( SSR \) be the sum of squared residuals from this regression. It follows that \( \text{Var}(\hat{G}) = \frac{\sum_{h=1}^{H} \sum_{c=1}^{n_{hc}} (u_{hc}^* - \bar{u}_{h}^*)^2}{\text{SSR}} = \text{SSR}. \) To undertake this regression, a researcher needs to initially generate \( u_{hc}^* = \sum_{i=1}^{m} w_{hc_i} u_{hc_i}^* \), which only requires a few lines of code.

Finally, Binder and Kovačević’s (1995) variance estimator, \( \text{Var}(\hat{G}) = \frac{\sum_{h=1}^{H} \sum_{c=1}^{n_{hc}} (u_{hc}^* - \bar{u}_{h}^*)^2}{\text{SSR}} \), can be generated without survey software as the \( \text{SSR} \) from the artificial regression:

\[ \left( \frac{n_{hc}}{n_h - 1} \right)^{0.5} u_{hc}^* = \sum_{h=1}^{H} \beta_h D_{hc}^* + v_{hc}, \]

where \( D_{hc}^* = \sqrt{(n_h/(n_h - 1))} \) if cluster \( c \) is in stratum \( h \), 0 otherwise.

6. AN APPLICATION

To be completed.

7. CONCLUSION

In this paper, we examine variance estimation for Gini indices calculated from complex survey samples. We show that a relatively recently proposed variance estimator (Bhattacharya, 2007) is equivalent to an estimator derived over ten years earlier by Binder and Kovačević (1995), in an article published in the survey literature. A key advantage of this equivalence result is that the variance formula provided by Binder and Kovačević, along with the approximation for the Gini estimator, is, we believe, far easier to practically calculate, of importance to practitioners who often resort to resampling methods for variance
estimation under the belief that it was too computationally burdensome to estimate a variance obtained from asymptotic approximations. This is indeed not the case; asymptotic variances can be readily calculated, even for researchers without access to specialized complex survey software, so providing an alternative to resampling methods such as the bootstrap.

As an iid sample can be regarded as a special case of a complex survey sample, we also link recent work of Davidson (2009) to the earlier research undertaken by Binder and Kovačević, showing that Davidson’s derived approximation for the Gini estimator and his proposed variance estimator also follow directly from Binder and Kovačević’s work.

In addition to linking relatively recent econometric research with past survey literature research, we believe that our work dismisses the folklore that asymptotic variances for Gini indices, especially with complex survey data, are computationally burdensome to calculate. We provide applied researchers with easily implementable ways to calculate both a Gini coefficient estimator and an estimator of its associated variance.

ACKNOWLEDGMENTS

We thank David Giles for his helpful comments and suggestions, and we are grateful for support from Nilanjana Roy. This paper is based on work in Ahmed Hoque’s Ph.D. dissertation.

REFERENCES


APPENDIX: PROOFS OF RESULTS

**Proof of Result 1:** Here we show equivalence of $\hat{\Phi}_{hci}$ and $u_{hci}^*$, implying that Bhattacharya’s (2007) and Binder and Kovačević’s (1995) approximate expressions for $G$ are the same. We have
\[
\Phi_{hci} = -2 \sum_{d=1}^{M} w_d \left\{ \frac{1}{\hat{\mu}} \left[ y_{hci} I(y_{hci} \leq y(d)) + y(d) \left( \hat{F}(y(d)) - I(y_{hci} \leq y(d)) \right) - \frac{\hat{\alpha}(y(d))}{\hat{\mu}} y_{hci} \right]\right\}
\]
\[
= -2 \frac{1}{\hat{\mu}} \left[ \sum_{d=1}^{M} w_d y_{hci} I(y_{hci} \leq y(d)) - \sum_{d=1}^{M} w_d y(d) I(y_{hci} \leq y(d)) + \sum_{d=1}^{M} w_d y(d) \hat{F}(y(d)) \right] - \sum_{d=1}^{M} \left( \frac{\hat{\alpha}(y(d))}{\hat{\mu}} \right) y_{hci}.
\]

Now:

- \[\sum_{d=1}^{M} w_d y(d) \hat{F}(y(d)) = \frac{\hat{\mu}}{2}(\hat{G} + 1)\]
- \[\sum_{d=1}^{M} w_d y(d) I(y_{hci} \leq y(d)) = B(y_{hci})\]
- \[\sum_{d=1}^{M} w_d y_{hci} I(y_{hci} \leq y(d)) = y_{hci} - y_{hci} \hat{F}(y_{hci})\]
- \[\sum_{d=1}^{M} \frac{w_d y_{hci} \hat{\alpha}(y(d))}{\hat{\mu}} = \frac{y_{hci}}{\hat{\mu}} \sum_{r=1}^{H} \sum_{s=1}^{n_r} \sum_{t=1}^{m_r} w_{rst} y_{rst} \sum_{d=1}^{M} w_d I(y_{rst} \leq y(d)) = y_{hci} - \frac{\hat{G} + 1}{2} y_{hci}.
\]

Using these results, we have

\[
\Phi_{hci} = -2 \frac{1}{\hat{\mu}} \left[ y_{hci} - y_{hci} \hat{F}(y_{hci}) - B(y_{hci}) + \frac{\hat{\mu}}{2}(\hat{G} + 1) - y_{hci} + \frac{\hat{G} + 1}{2} y_{hci} \right]
\]
\[
= u^*_h \ #
\]

**Proof of Result 2:** Here we show that \(\bar{V}a\bar{r}(\hat{G})\) as given by expression (4.2) is equivalent to \(\bar{V}a\bar{r}(\hat{G})\) as defined in expression (4.3). From (4.2)

\[
\bar{V}a\bar{r}(\hat{G}) = \sum_{h=1}^{H} \sum_{c=1}^{n_h} (u^*_{hc} - \bar{u}_h)^2
\]
\[
= \sum_{h=1}^{H} \sum_{c=1}^{n_h} (u^2_{hc} - 2u^*_{hc} \bar{u}_h + \bar{u}_h^2) \ . \quad (A.1)
\]

We have,
\[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} \left( \sum_{i=1}^{m} w_{hcl} u_{hcl}^* \right)^2 \]

\[ = \sum_{h=1}^{H} \sum_{c=1}^{n_h} \left( \sum_{i=1}^{m} w_{hcl}^2 u_{hcl}^{*2} \right) + \sum_{i=1}^{m} \sum_{i' \neq i} w_{hcl} w_{hcl'} u_{hcl}^* u_{hcl'}^* , \tag{A.2} \]

and,

\[ \sum_{h=1}^{H} \sum_{c=1}^{n_h} (-2 \bar{u}_h^* u_{hc}^* + \bar{u}_h^{*2}) = \sum_{h=1}^{H} \left( -2 \bar{u}_h^* \sum_{c=1}^{n_h} u_{hc}^* + \sum_{c=1}^{n_h} \bar{u}_h^{*2} \right) \]

\[ = - \sum_{h=1}^{H} n_h \bar{u}_h^{*2} \]

\[ = - \sum_{h=1}^{H} \frac{1}{n_h} \left( \sum_{c=1}^{n_h} \sum_{i=1}^{m} w_{hcl} u_{hcl}^* \right)^2 , \tag{A.3} \]

Substituting (A.2) and (A.3) into (A.1) and using that \( u_{hcl}^* = \tilde{\Phi}_{hcl} \) gives the desired result that

\[ \text{Var}(\mathcal{G}) = \text{Var}(\mathcal{G}). \quad \# \]