UNIVERSITY OF Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M08/04
Methodology

A Note On The Asymptotic Equivalence Of Jackknife And Linearization Variance Estimation For The Gini Coefficient

Yves G. Berger

Abstract

The Gini coefficient has proved valuable as a measure of income inequality. In cross-sectional studies of the Gini coefficient, information about the accuracy of its estimate is crucial. We show how to use jackknife and linearization to estimate the variance of the Gini coefficient, allowing for the effect of the sampling design. The aim is to show the asymptotic equivalence (or consistency) of the generalised jackknife estimator and the Taylor linearization estimator for the variance of the Gini coefficient. A brief simulation study supports our findings.

A Note on the Asymptotic Equivalence of Jackknife and Linearization Variance Estimation for the Gini Coefficient

Yves G. Berger
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton, SO17 1BJ, U.K.
Email: Y.G.Berger@soton.ac.uk

Abstract

The Gini coefficient (Gini, 1914) has proved valuable as a measure of income inequality. In cross-sectional studies of the Gini coefficient, information about the accuracy of its estimates is crucial. We show how to use jackknife and linearization to estimate the variance of the Gini coefficient, allowing for the effect of the sampling design. The aim is to show the asymptotic equivalence (or consistency) of the generalised jackknife estimator (Campbell, 1980) and the Taylor linearization estimator (Kovačević and Binder, 1997) for the variance of the Gini coefficient. A brief simulation study supports our findings.

Key words: Inclusion probability, linearization, survey weight, sampling design.

Acknowledgements

The author is grateful to Marinela Kustura (Central Bureau of Statistics, Croatia) for her help with preliminary simulations. The author is also grateful to four anonymous referees for helpful comments.

1. The Gini coefficient

In this section, we introduce some notations, define the Gini coefficient and define its estimators.

Consider a finite population denoted by $U=\{1,...,i,...,N\}$, where N is the number of individuals in this population. Let $y_i \geq 0$ denote the income of an individual labelled i. The finite population Gini coefficient is defined by (Glasser, 1962)

$$\gamma=\frac{1}{\tau}\sum_{i\in U}(2F(y_i)-1)\,y_i\ , \tag{1}$$

where $\tau=\sum_{i\in U} y_i$. The function $F(y)$ denotes the income distribution function defined by

$$F(y)=\frac{1}{N}\sum_{i\in U}\delta\{y_i \leq y\},$$

where $\delta\{y_i \leq y\}$ takes the value 1 if $y_i \leq y$ and the value 0 otherwise.

When $y_i \neq y_j$ for all $i \neq j$, equation (1) can be re-expressed as

$$\gamma=\gamma^{*}+\frac{1}{N}, \tag{2}$$

where

$$\gamma^* = \frac{2\operatorname{cov}(y, F(y))}{\mu}, \tag{3}$$

$$\operatorname{cov}(y, F(y)) = \frac{1}{N}\sum_{i\in U} y_i F(y_i) - \frac{\tau}{N^2}\sum_{i\in U} F(y_i),$$

and $\mu = \tau / N$. The quantity γ^* is an alternative expression for the Gini coefficient proposed by Anand (1983) and by Lerman and Yitzhaki (1984).

More generally, γ becomes γ^* when we replace $F(y_i)$ in (1) with the smooth (or mid-interval) distribution function $F^*(y_i) = [F(y_i) + F(y_i - 0)]/2$, where $F(y_i - 0) = \lim_{y \uparrow y_i} F(y)$. Note that $F^*(y_i)$ is not a cumulative discrete distribution, as $F^*(y_i)$ is not the fraction of observation less or equal to y_i. This adjustment to the cumulative distribution allows the Gini coefficient to be computed using (3) (see Lerman and Yitzhaki, 1989). In other words, using the smooth distribution function effectively takes into account the correction $1/N$ in (2). For simplicity, we will ignore this correction in what follows.

Suppose that y_i is known only for the sampled individuals $i \in s$, where s denote a sample or subset of the population U. Hence, the Gini coefficient in (1) is an unknown population parameter, as it depends on unobserved quantities y_i ($i \notin s$). Thus, it has to be estimated from the observed sampled values y_i ($i \in s$). A substitution estimator for γ is given by (Kovačević and Binder, 1997):

$$\hat{\gamma} = \frac{1}{\hat{\tau}}\sum_{i\in s} w_i (2\hat{F}(y_i) - 1) y_i, \tag{4}$$

where

$$\hat{F}(y) = \frac{1}{\hat{N}}\sum_{i\in s} w_i \delta\{y_i \le y\}, \tag{5}$$

with $\hat{\tau}=\sum_{i\in s} w_i y_i$, $\hat{N}=\sum_{i\in s} w_i$, and $w_i=\pi_i^{-1}$ denotes the Horvitz-Thompson (1952) weights of individual i. The quantity π_i is the first-order inclusion probability of i; that is, the probability that individual i is in the sample. Using the Horvitz-Thompson weights guarantee that $\hat{\gamma}$ is an approximately unbiased estimator for γ.

Nygård and Sandström (1985) proposed an alternative estimator. Their $\hat{\gamma}^*$ is given by (4) after replacing $\hat{F}(y_i)$ with the sample smooth (or mid-interval) distribution function $\hat{F}^*(y_i)=[\hat{F}(y_i)+\hat{F}(y_i-0)]/2$, where $\hat{F}(y_i-0)=\lim_{y\uparrow y_i}\hat{F}(y_i)$. Taking a slightly different approach, Lerman and Yitzhaki (1989) proposed substituting $\hat{F}^*(y_i)$ into (3). Using $\sum_{i\in s}\hat{F}^*(y_i)/\pi_i=\hat{N}/2$, it can be shown that their estimator reduces to $\hat{\gamma}^*$. Deville (1997) likewise proposed an estimator algebraically equivalent to $\hat{\gamma}^*$.

The estimator $\hat{\gamma}^*$ is asymptotically identical to $\hat{\gamma}$ under mild conditions, as $\hat{\gamma}=\hat{\gamma}^*+\nu$, where $|\nu|<\max\{w_i : i\in s\}/\hat{N}$. Thus, $\hat{\gamma} \doteq \hat{\gamma}^*$ when $|\nu| \doteq 0$ or when $w_i/\hat{N}=O_p(1/n)$ uniformly; that is, when none of the weights is disproportionately large (Krewski and Rao, 1981). In this situation, the quantity ν is of probability order $1/n$, which implies that the difference between the variances of $\hat{\gamma}$ and $\hat{\gamma}^*$ is of probability order $1/n^2$ (Deville, 1997). This difference can be ignored in the estimation of the variance. We will assume that the sample size is large enough that the same expression can be used to estimate the variance of both $\hat{\gamma}$ and $\hat{\gamma}^*$.

In what follows, we investigate the jackknife and the linearization variance of the estimator $\hat{\gamma}$ in (4) based on the estimate of the distribution function (5).

Lerman and Yitzhaki (1984) and Ogwang (2000) showed that the Gini coefficient can be easily estimated using the regression coefficient of an ordinary least squares regression. By assuming this regression model true, the variance of the regression coefficient can be used to estimate the variance of the Gini coefficient (Ogwang, 2004, Giles, 2004). Unfortunately, this model-driven approach can give biased estimates for the variance in practice, as the residuals of the regression model are rarely independent (Ogwang, 2004). For example, Modarres and Castwirth (2006) showed that the regression technique can significantly overestimate the true variance. An additional problem with this approach is that it ignores the sampling design.

In this paper, we do not assume a model. Instead, we propose variance estimators based on a design-based approach in which the variability of $\hat{\gamma}$ comes from the random selection of the sample. This allows us to account for the complexity of the sampling design. For further details about the model-based approach see Sandström (1983) and Nygård and Sandström (1985).

2. Variance estimation by linearization

We now consider estimating the variance of $\hat{\gamma}$ in (4). The basic idea of the linearization method (e.g., Krewski & Rao, 1981; Robinson and Särndal, 1983; Särndal *et al.*, 1992, p.175; Andersson and Nordberg, 1994; Deville, 1999) is to use 'pseudo-values' z_i such that $\text{var}(\hat{\gamma}) \doteq \text{var}(\hat{\tau}_z)$, where $\hat{\tau}_z = \sum_{i \in s} w_i z_i$. The approximation $\doteq$ is justified by some large-sample arguments (see Krewski & Rao, 1981). The variance is defined with respect to the sampling design; that is, with respect to the probability distribution $p(s)$ of the randomly-selected sample s. The linearization variance estimator (Robinson and Särndal, 1983; Särndal *et al.*, 1992, p.175) is then the design-based estimator for the variance of $\hat{\tau}_z$. This estimator is given by

$$\hat{\text{var}}(\hat{\gamma})_L = \sum_{i\in s}\sum_{j\in s}\breve{\Delta}_{ij} w_i w_j z_i z_j \tag{6}$$

where $\breve{\Delta}_{ij} = (\pi_{ij} - \pi_i \pi_j)\pi_{ij}^{-1}$, and π_{ij} denotes the joint inclusion probability of individuals i and j; that is, the probability that both i and j are in the sample. Unfortunately, the estimator in (6) can take negative values (Cochran 1977, p.261). This issue will be discussed briefly in Section 3.

The form of the pseudo-values z_j can be illustrated in the simplest case when the sampling variation of $\hat{F}(y_i)$ in $\hat{\gamma}$ is ignored. In this case, $\hat{\gamma}$ is a ratio of two sums and the Taylor linearization of this ratio gives naïve pseudo-values given by

$$z_j = \frac{1}{\hat{\tau}}\left[2y_j\hat{F}(y_j) - (\hat{\gamma}+1)y_j\right]. \tag{7}$$

This method was cautiously suggested by Nygård and Sandström (1985) who reported that it over-estimates the variance significantly (see also Sandström *et al.*, 1985, 1988). In Section 4, we empirically confirm that using the pseudo-value in (7) does not result in accurate estimates for the variance. This is because the sampling variation in $\hat{F}(y_i)$ has a nonnegligible contribution into the variance of $\hat{\gamma}$.

Kovačević and Binder (1997) (see also Deville, 1997, 1999) showed that additional terms were needed in the pseudo-values. They set

$$z_j = \frac{1}{\hat{\tau}}\left[2y_j\hat{F}(y_j) - (\hat{\gamma}+1)\left(y_j + \frac{\hat{\tau}}{\hat{N}}\right) + \frac{2}{\hat{N}}\sum_{i\in s} w_i y_i \delta\{y_j \le y_i\}\right]. \tag{8}$$

In Section 3 and 4, the linearization estimator in (6) with z_j given by (8) will be compared with the generalised jackknife estimator to be defined in Section 3.

3. The jackknife estimator for the variance

The jackknife is a numerical method which can be used to estimate a variance (Miller, 1974). In particular, the jackknife technique is commonly employed to estimate the variance of the Gini coefficient (Yitzhaki, 1991; Karoly,1992; Karagiannis and Kovačević, 2000; Newson, 2006 and Frick *et al*., 2006). In this section, we compare the jackknife estimator with the linearization estimator. We show that these estimators are asymptotically equivalent and consistent under mild conditions.

Campbell (1980) proposed a generalised jackknife variance estimator that fully captures the impact of the sampling design. Berger and Skinner (2005) showed that, under mild conditions, this estimator is consistent for a parameter expressible as a function of means. Although, $\hat{\gamma}$ is not expressible as a function of means, we show in this section that the generalised jackknife variance estimator is a consistent estimator for the variance of $\hat{\gamma}$ provided that the linearization estimator in (6) is consistent.

Campbell's generalised jackknife variance estimator (see also Berger and Skinner, 2005) is given by

$$\hat{\mathrm{var}}(\hat{\gamma})_{GJ} = \sum_{i\in s}\sum_{j\in s}\breve{\Delta}_{ij} w_i w_j \tilde{z}_i \tilde{z}_j \,, \tag{9}$$

where the quantities $\tilde{z}_j$ are pseudo-values:

$$\tilde{z}_j = w_j^{-1}(1 - w_j \hat{N}^{-1})(\hat{\gamma} - \hat{\gamma}_{(j)}) \,, \tag{10}$$

with

$$\hat{\gamma}_{(j)} = \frac{1}{\hat{\tau}_{(j)}} \sum_{i\in s_{(j)}} w_i (2\hat{F}(y_i)_{(j)} - 1) y_i \,,$$

$$\hat{F}(y)_{(j)} = \frac{1}{\hat{N}_{(j)}} \sum\nolimits_{i\in s_{(j)}} w_i \delta\{y_i \le y\} \,,$$

$\hat{\tau}_{(j)} = \sum_{i \in s_{(j)}} w_i y_i$,

$\hat{N}_{(j)} = \sum_{i \in s_{(j)}} w_i$, and $s_{(j)} = s \setminus \{j\}$, the last being s with the j-th individual deleted.

Berger and Skinner (2005) showed that under simple random sampling without replacement, the variance estimator (9) reduces to the customary jackknife estimator with finite population correction (e.g. Miller, 1974) given by

$$\hat{\text{var}}(\hat{\gamma})_{CJ} = \left(1 - \frac{n}{N}\right) \frac{1}{n(n-1)} \sum_{i \in s} (\hat{\gamma}_j - \bar{\gamma})^2 \, , \tag{11}$$

where $\hat{\gamma}_j = n\hat{\gamma} - (n-1)\hat{\gamma}_{(j)}$ and $\bar{\gamma} = (1/n)\sum_{i \in s} \hat{\gamma}_j$. Moreover, the generalised jackknife estimator in (9) remains consistent under unequal probabilities sampling (Berger and Skinner); whereas the customary jackknife estimator in (11), does not, because the finite population correction factor 1 – *n/N* is *ad hoc*.

In the Appendix, we demonstrate that $\tilde{z}_j$ defined by (10) can be re-written as

$$\tilde{z}_j = \frac{\hat{\tau}}{\hat{\tau}_{(j)}} z_j - 2 \frac{w_j y_j}{\hat{N} \hat{\tau}_{(j)}} , \tag{12}$$

where z_j is given by (8). This means that $\tilde{z}_j$ is approximately equal to z_j given by (8), provided that $\hat{\tau} / \hat{\tau}_{(j)} = 1 + O_p(1/n)$ and $w_j y_j / (\hat{N} \hat{\tau}_{(j)}) = O_p(1/n)$. Hence, the jackknife estimator in (9) and the linearization estimator in (6) are approximately equal when the z_j are given by (8). As a consequence, the generalized jackknife estimator is consistent provided that the linearization estimator is.

4. Simulation study

In this section, the jackknife estimators in (9) and (11) are compared numerically with two linearization estimators (see (6)): the naïve linearization estimator that uses the pseudo values in (7) and the linearization estimator that uses the pseudo values in (8).

We evaluate three populations each of $N = 500$ y_i values, generated by the following probability distributions: a Gamma distribution (shape parameter = 2.5, rate = 1), a Lognormal distribution (mean = 1.119, standard deviation = 0.602) and a Weibull distribution (shape= 0.8, scale = 1). We focus on these distributions as they are good approximation of income distributions (Salem and Mount, 1974; McDonald, 1984).

We use the Chao (1982) sampling design for selecting units with unequal inclusion probabilities π_i. These are set proportional to a size variable x_i generated from the model $x_i = \alpha + \rho\, y_i + e_i$, where the e_i come from a normal distribution with mean zero and variance $\sigma_e^2 = (1-\rho^2)(N-1)^{-1}\sum_{i \in U}(y_i - \mu)^2$, $\alpha = 5 + \rho\, \mu$, $\rho = 0.7$, and $\mu = \tau / N$ is the population mean of the y_i. The x_i are treated as fixed after they are generated. The π_{ij} are computed exactly using the recursive formula proposed by Chao (1982).

For each population, $B = 10\,000$ samples are selected. The empirical relative bias is defined here as

$$\text{RB} = \frac{\text{Bias}(\hat{\text{var}}(\hat{\gamma}))}{\text{MSE}(\hat{\gamma})},$$

where $\text{Bias}(\hat{\text{var}}(\hat{\gamma}))$ and $\text{MSE}(\hat{\gamma})$ denote respectively the empirical bias and the empirical mean square error of $\hat{\gamma}$. Furthermore,

$$\text{Bias}(\hat{\text{var}}(\hat{\gamma})) = \frac{1}{B}\sum_{\text{b=1}}^{\text{B}} \hat{\text{var}}(\hat{\gamma})_b - \text{var}(\hat{\gamma}), \text{ and}$$

$$\text{MSE}(\hat{\gamma}) = \frac{1}{B-1}\sum_{\text{b=1}}^{\text{B}} (\hat{\gamma}_b - \gamma)^2 \ ,$$

where $\hat{\gamma}_b$ is the estimate for the b-th sample, whereas $\hat{\text{var}}(\hat{\gamma})_b$ is an estimate of its variance.

The quantity $\text{var}(\hat{\gamma})$ denotes the empirical variance of $\hat{\gamma}$, which is

$$\text{var}(\hat{\gamma}) = \frac{1}{B-1}\sum_{\text{b=1}}^{\text{B}} [\hat{\gamma}_b - \text{E}(\hat{\gamma})]^2 \ ,$$

where

$$\text{E}(\hat{\gamma}) = \frac{1}{B}\sum_{\text{b=1}}^{\text{B}} \hat{\gamma}_b \ .$$

The empirical relative root mean square error of $\hat{\text{var}}(\hat{\gamma})$ is

$$\text{RRMSE}(\hat{\text{var}}(\hat{\gamma})) = \frac{\text{MSE}(\hat{\text{var}}(\hat{\gamma}))^{1/2}}{\text{MSE}(\hat{\gamma})} \ ,$$

where

$$\text{MSE}(\hat{\text{var}}(\hat{\gamma})) = \frac{1}{B-1}\sum_{\text{b=1}}^{\text{B}} [\hat{\text{var}}(\hat{\gamma})_b - \text{var}(\hat{\gamma})]^2 \ .$$

Table 1 displays the empirical expectation of $\hat{\gamma}$ and $\hat{\gamma}^*$ and the ratio of their empirical variances under the distributions for several sample sizes. Table 1 shows that both $\hat{\gamma}$ and $\hat{\gamma}^*$ can have large absolute biases when the sample size is small. The ratio of the variances is close to one when the sample size is sufficiently large. This is a result we expect, as the difference between the variances of $\hat{\gamma}$ and $\hat{\gamma}^*$ is of order $1/n^2$ (see Section 1). Thus, the variance estimators developed here for estimating the variance of $\hat{\gamma}$

can also be used to estimate the variance of $\hat{\gamma}^*$ provided that the sample size is sufficiently large. For small sample sizes, $\hat{\gamma}$ and $\hat{\gamma}^*$ may be biased, and the linearization technique and the jackknife are not recommended for variance estimation.

	Gamma			Lognormal			Weibull		
	$\gamma = 0.34$ and $\gamma^* = 0.34$			$\gamma = 0.28$ and $\gamma^* = 0.27$			$\gamma = 0.60$ and $\gamma^* = 0.60$		
n	E($\hat{\gamma}$)	E($\hat{\gamma}$*)	$\frac{\text{var}(\hat{\gamma})}{\text{var}(\hat{\gamma}^*)}$	E($\hat{\gamma}$)	E($\hat{\gamma}$*)	$\frac{\text{var}(\hat{\gamma})}{\text{var}(\hat{\gamma}^*)}$	E($\hat{\gamma}$)	E($\hat{\gamma}$*)	$\frac{\text{var}(\hat{\gamma})}{\text{var}(\hat{\gamma}^*)}$
	0.47	0.28	0.83	0.42	0.22	0.90	0.67	0.49	0.75
5	0.37	0.33	0.96	0.30	0.26	0.98	0.62	0.58	0.93
25	0.35	0.33	0.98	0.29	0.27	0.99	0.61	0.60	0.97
50	0.35	0.34	0.99	0.28	0.27	0.99	0.61	0.60	0.98
100	0.34	0.34	0.99	0.28	0.27	1.00	0.61	0.60	0.99

Table 1: Empirical expectation and ratio of variance of $\hat{\gamma}$ and $\hat{\gamma}^*$, for the three distributions and several sample sizes.

	Gamma $\gamma = 0.34$				Lognormal $\gamma = 0.28$				Weibull $\gamma = 0.60$			
	Linearization		Jackknife		Linearization		Jackknife		Linearization		Jackknife	
n	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)
5	209%	-6.3%	7.1%	5.2%	254%	-5.7%	4.5%	5.1%	127%	-30.1%	24.9%	26.8%
25	366	-4.0	4.4	2.8	522	-5.2	3.0	2.9	104	-10.5	9.2	6.4
50	391	-4.9	-0.8	-0.9	598	-4.9	1.5	0.1	102	-3.9	11.0	4.6
100	394	-2.8	-2.8	-0.6	694	0.8	8.4	3.7	93	-0.9	18.2	3.0
150	369	-2.7	-5.3	-1.2	692	-3.0	7.5	-1.1	73	-0.2	29.0	2.1

Table 2: Empirical RB (%) of the variance estimator based upon (7), (8), (9) and (11) for the three distributions and several sample sizes.

Table 2 and 3 display the RB and the RRMSE of the linearization and jackknife variance estimators for several sample sizes. Table 4 provides the empirical coverages of 95% confidence intervals computed in the following manner:

$$\text{Coverage} = \frac{1}{B}\sum_{b=1}^{B}\delta(|\, z_b \,| \leq 1.96),$$

with $z_b = (\hat{\gamma}_b - \gamma)\,\hat{\text{var}}(\hat{\gamma})_b^{-1/2}$ and $\delta(|\, z_b \,| \leq 1.96)$ equal to 1 when $|\, z_b \,| \leq 1.96$, 0 otherwise.

	Gamma $\gamma = 0.34$				Lognormal $\gamma = 0.28$				Weibull $\gamma = 0.60$			
	Linearization		Jackknife		Linearization		Jackknife		Linearization		Jackknife	
n	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)
5	217%	19.9%	34.6%	31.1%	258%	17.1%	29.1%	31.3%	159%	41.4%	86.3%	96.5%
25	369	27.4	30.7	29.1	524	32.0	38.0	37.8	114	30.0	42.1	39.3
50	394	23.0	23.0	23.1	599	26.7	30.4	28.8	108	19.8	28.1	22.6
100	395	18.2	17.0	18.2	694	19.1	24.6	20.3	96	12.9	24.4	13.7
150	370	15.5	14.4	15.5	693	13.2	18.8	13.3	75	10.7	31.4	11.1

Table 3: Empirical RRMSE (%) of the variance estimator based upon (7), (8), (9) and (11) for the three distributions and several sample sizes.

	Gamma $\gamma = 0.34$				Lognormal $\gamma = 0.28$				Weibull $\gamma = 0.60$			
	Linearization		Jackknife		Linearization		Jackknife		Linearization.		Jackknife	
n	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)	(7)	(8)	(11)	(9)
5	99%	55%	69%	68%	100%	38%	56%	56%	93%	73%	87%	87%
25	100	89	91	91	100	89	91	91	98	90	93	93
50	100	92	93	93	100	92	93	93	99	93	95	94
100	100	94	94	94	100	94	95	94	99	94	96	95
150	100	94	94	94	100	94	95	94	99	95	97	95

Table 4: Empirical Coverage (%) of the confidence interval based on the variance estimator based upon (7), (8), (9) and (11) for the three distributions and several sample sizes.

The naïve variance estimator based upon (7) is not recommended, as it clearly over-estimates the variance significantly (see Table 2). However, the linearization variance estimator based upon (8) and the jackknife estimator in (9) have small RB and RRMSE. The jackknife estimators may slightly over-estimate the variance, and the linearization estimator may slightly under-estimate the variance. We observe that the RRMSE of the linearization estimator based upon (8) is smaller than the RRMSE of the generalised jackknife (9).

The linearization and jackknife estimators also produce more reasonable coverage intervals than the naïve estimator based on (7). Between the two, we have a slightly better coverage with the jackknife estimators. It is natural to have a poor coverage with small sample sizes, as the normal assumption is not suitable when the sample size is too small.

Both jackknife estimators have roughly the same RB for the Gamma and the Lognormal distribution. However with the Weibull distribution which has the largest Gini coefficient, the RB of the customary jackknife (11) is larger than the RB of the generalised jackknife (9).

6. Discussion

This paper showed that linearization technique proposed by Kovačević and Binder (1997) and the generalised jackknife are asymptotically equivalent and consistent under mild conditions. This finding is supported by a simulation study.

We assumed here that the survey weights were the Horvitz-Thompson weights. Our methodology can be easily extended to more complex weighting schemes. For example, under calibration the pseudo-values in (8) or (12) could be replaced by linear-regression residuals treating the pseudo-values themselves as the dependent variables and the calibration variables as the explanatory variables (Deville, 1999; Berger and Skinner, 2003).

The variance estimators in (6) and (9) depend on joint inclusion probabilities π_{ij} which can be cumbersome to compute under an unequal probability sampling scheme. Furthermore, both the linearization and generalized jackknife estimator can be negative. Under a single stage, stratified,sampling design featuring unequal inclusion probabilities within strata, it is tempting to use the simplified Hájek (1964) variance estimator. This estimator approximates the π_{ij} employing only the first-order inclusion probabilities (see Berger, 2004). Berger (2007) proposed a π_{ij}-free jackknife estimator which is consistent for a class of high-entropy stratified designs using Rao-Sampford unequal-probability sampling within strata (Rao, 1965; Sampford, 1967). This estimator also uses the pseudo-values in (10) and could be employed to estimate the variance of the Gini coefficient. The estimator proposed by Berger (2007) is always nonnegative.

Large national household surveys often employ two stage or multistage sampling. For such surveys, the joint inclusion probabilities π_{ij} will often not be known, and stage-wise approximations to them may be necessary. For that reason the generalised jackknife has more promise for single-stage business surveys.

Many surveys use single imputation to handle item nonresponse. In this situation, one can use the Rao and Shao (1992) method, which consists of adjusting the imputed values whenever a responding unit is deleted. Berger and Rao (2006) showed how to implement the Rao and Shao (1992) method to accommodate imputed values with the generalized jackknife. They also showed that the resulting jackknife variance estimator is consistent under mild conditions.

The computation of pseudo-values in (10) can be computationally intensive. Yitzhaki (1991), Karoly (1992), Karagiannis and Kovačević (2000) and Newson (2006) proposed simple methods to compute the customary jackknife with finite population correction in (11). Generalising these methods to Campbell's jackknife in (9) would be a fruitful direction for future research.

Appendix – Proof of (12)

Using

$$\hat{\gamma} = \frac{2}{\hat{\tau}} \sum_{i \in s} w_i y_i \hat{F}(y_i) - 1,$$

it can be shown that

$$\begin{aligned}
\hat{\gamma}_{(j)} &= \frac{2}{\hat{\tau}_{(j)}} \left[\sum_{i \in s} \frac{w_i y_i}{\hat{N}_{(j)}} \left(\sum_{k \in s} w_k \delta_{ki} - w_j \delta_{ji} \right) - \frac{w_j y_j}{\hat{N}_{(j)}} \left(\sum_{k \in s} w_k \delta_{kj} - w_j \delta_{jj} \right) \right] - 1, \\
&= \frac{2}{\hat{N}_{(j)} \hat{\tau}_{(j)}} \left[\hat{N} \sum_{i \in s} w_i y_i \hat{F}(y_i) - w_j \sum_{i \in s} w_i y_i \delta_{ji} - w_j y_j \hat{N} \hat{F}(y_j) + w_j^2 y_j \delta_{jj} \right] - 1, \\
&= \frac{2}{\hat{N}_{(j)} \hat{\tau}_{(j)}} \left[(\hat{\gamma}+1) \frac{\hat{N}\hat{\tau}}{2} - w_j \sum_{i \in s} w_i y_i \delta_{ji} - w_j y_j \hat{N} \hat{F}(y_j) + w_j^2 y_j \right] - 1,
\end{aligned}$$

where $\delta_{ji} = \delta\{y_j \le y_i\}$. Thus,

$$\begin{aligned}
\hat{\gamma} - \hat{\gamma}_{(j)} &= \frac{2}{\hat{N}_{(j)} \hat{\tau}_{(j)}} \Bigg[w_j y_j \hat{N} \hat{F}(y_j) + (\hat{\gamma}+1) \frac{\hat{N}_{(j)} \hat{\tau}_{(j)}}{2} - (\hat{\gamma}+1) \frac{\hat{N}\hat{\tau}}{2} \\
&\qquad + w_j \sum_{i \in s} w_i y_i \delta_{ji} - w_j^2 y_j \Bigg] \\
&= \frac{w_j}{\hat{N}_{(j)} \hat{\tau}_{(j)}} \Bigg[2 y_j \hat{N} \hat{F}(y_j) + (\hat{\gamma}+1) \frac{\hat{N}_{(j)} \hat{\tau}_{(j)} - \hat{N}\hat{\tau}}{w_j} \\
&\qquad + 2 \sum_{i \in s} w_i y_i \delta_{ji} - 2 w_j y_j \Bigg]. \qquad (13)
\end{aligned}$$

We have $\hat{N}_{(j)} \hat{\tau}_{(j)} - \hat{N}\hat{\tau} = (\hat{N} - w_j)(\hat{\tau} - w_j y_j) - \hat{N}\hat{\tau} = -w_j (y_j \hat{N} + \hat{\tau})$ which substituted into (13) gives

$$\hat{\gamma} - \hat{\gamma}_{(j)} = \frac{w_j}{\hat{N}_{(j)} \hat{\tau}_{(j)}} \left[2 y_j \hat{N} \hat{F}(y_j) - (\hat{\gamma}+1)(y_j \hat{N} + \hat{\tau}) + 2 \sum_{i \in s} w_i y_i \delta_{ji} - 2 w_j y_j \right].$$

Now, as $\tilde{z}_j = w_j^{-1} (1 - w_j \hat{N}^{-1})(\hat{\gamma} - \hat{\gamma}_{(j)}) = w_j^{-1} \hat{N}^{-1} \hat{N}_{(j)} (\hat{\gamma} - \hat{\gamma}_{(j)})$, we obtain

$$\tilde{z}_j = \frac{1}{\hat{N}\hat{\tau}_{(j)}}\left[2y_j\hat{N}\hat{F}(y_j) - (\hat{\gamma}+1)(y_j\hat{N}+\hat{\tau}) + 2\sum_{i\in s} w_i y_i \delta_{ji} - 2w_j y_j\right],$$

which implies (12). This completes the proof.

References

Anand, S. (1983) Inequality and Poverty in Malaysia: Measurement and Decomposition, Oxford University Press.

Andersson, C. and Nordberg, L. (1994). A method for variance estimation of non-linear function of totals in surveys - Theory and a software implementation, Journal of Official Statistics, Vol. 10, pp. 395-405.

Berger, Y.G. (2004). A Simple Variance Estimator for Unequal Probability Sampling Without Replacement, J. App. Stat., Vol. 3, pp. 305-315.

Berger (2007). A Jackknife Variance Estimator for Unistage Stratified Samples With Unequal Probabilities. Biometrika. Vol 94, pp. 953-964.

Berger, Y.G. & Rao, J.N.K. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. J. R. Statist. Soc. B, Vol. 68, pp.531-47.

Berger, Y.G. and Skinner, C.J. (2003), Variance Estimation of a Low-Income Proportion. J. Roy. Statist. Soc. Ser. C., Vol. 52, pp. 457-468.

Berger, Y.G. and Skinner, C.J. (2005). A Jackknife Variance Estimator for Unequal Probability Sampling, J. Roy. Statist. Soc. Ser. B., Vol. 67, pp. 1-11.

Campbell, C. (1980). A different view of finite population estimation, Proceeding Survey Research Methods Section ASA, pp. 319-24.

Chao, M. T. (1982). A General Purpose Unequal Probability Sampling Plan, Biometrika, Vol. 69, pp. 653-656.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. John Wiley, New York.

Deville J.C. (1997). Estimation de la variance du coefficient de Gini mesuré par sondage (In french), Actes des Journées de Méthodologie Statistiques, Insee méthodes.

Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques, Survey Methodology, Vol. 25, pp. 193-203.

Frick, R. J., Goebel, J. , Schechtman, E., Wagner, G.G. and Yitzhaki, S.(2006). Using Analysis of Gini (ANOGI) for Detecting Whether Two Sub-Samples Represent the Same Universe: The German Socio- Economic Panel Study (SOEP) Experience, Sociological Methods and Research, Vol 34, 427-468.

Giles, D. E. A. (2004). Calculating a Standard Error for the Gini Coefficient: Some Further Results, Oxford Bulletin of Economics and Statistics, Vol. 66, pp. 425-433

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri, Atti del R. Istituto Veneto di Scienze Lettere ed Arti.

Glasser, G.J. (1962). Variance Formulas for the Mean Difference and Coefficient of Concentration. Journal of the American Statistical Association, 57, 648-654.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population, Annals of Mathematical Statistics Vol. 35, pp. 1491-1523.

Horvitz, G.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc., Vol. 47, pp. 663-685.

Karagiannis, E. and Kovačević, M.S. (2000). A method to calculate the jackknife variance estimator for the Gini coefficient, Oxford Bulletin of Economics and statistics Vol. 62, pp. 119-122.

Karoly, L. (1992). Change in the distribution of individual earnings in the United states: 1967-1986, The Review of Economics and Statistics, Vol 74, pp. 107-115

Kovačević, M.S., and Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization – the estimating equations approach, Journal of Official Statistics, Vol. 13, pp. 41-58.

Kovacevic, M.S. and Yung W. (1997). Variance estimation for measures of income inequality and polarization - an empirical study. Survey Methodology, Vol. 23, pp. 41-52.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified sample: properties of linearization jackknife, and balanced repeated replication methods, Ann. Statist., Vol. 9, pp. 1010-1019.

Lerman, R.I. and Yitzhaki, S. (1984). A Note on the Calculation and Interpretation of the Gini Index, Economics Letters, Vol. 15, pp. 636-368.

Lerman, R.I. and Yitzhaki, S. (1989). Improving the Accuracy of Estimates of Gini Coefficients, Journal of Econometrics, Vol. 42, pp. 43-47.

Mcdonald, J.B. (1984). Some generalized functions for the size distribution of income, Econometrica, Vol. 52, pp. 647-664.

Miller, R.G. (1974). The jackknife: a review, Biometrika, Vol. 61, pp. 1-15.

Modarres, R. and Castwirth, J. L. (2006), A Cautionary Note on Estimating the Standard Error of the Gini Index of Inequality, Oxford Bulletin of Economics and Statistics, Vol. 68, pp. 385-390

Newson, R. (2006) Efficient Calculation of Jackknife Confidence Intervals for Rank Statistics, Journal of Statistical Software, Vol. 15, pp. 1-10.

Nygård, F. and Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations, Journal of Official Statistics, Vol. 1, pp. 399-412

Ogwan, T. (2000). A Convenient Method of Computing the Gini Index and its Standard Error, Oxford Bulletin of Economics and Statistics, Vol. 62, pp. 123-129

Ogwan, T. (2004). Calculating a Standard Error for the Gini Coefficient: Some Further Results: reply, Oxford Bulletin of Economics and Statistics, Vol. 66, pp. 435-437

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. J. Indian Statist. Assoc. Vol 3, pp. 173-80.

Rao, J.N.K. and Shao, A.J. (1992) Jackknife variance estimation with survey data under hotdeck imputation. Biometrika, Vol 79, pp. 811-822.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992) Some recent work on resampling methods for complex surveys. Surv. Methodol., Vol 18, pp. 209-217.

Robinson, P.M. and Särndal, C.E. (1983). Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling, Sankhyā Ser. B Vol. 45, pp. 240-248.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. Biometrika. Vol 54, pp. 499-513.

Sandström, A. (1983). Estimating Income Inequality, Large Sample Inference in Finite Population, Dept. of Statist., University of Stockholm, Research Report.

Sandström, A., Wretman, J. H. and Waldén, B. (1985). Variance Estimators of the Gini Coefficient, Simple Random Sampling, Metron, Vol. 43, pp. 41-70

Sandström, A., Wretman, J. H. and Waldén, B. (1988). Variance Estimators of the Gini Coefficient – Probability Sampling, Journal of Business and Economic Statistics, Vol. 6, pp. 113-119

Salem, A.B.Z., AND mount, T.D. (1974). A Convenient Descriptive Model of Income Distribution: The Gamma Density, Econometrica, Vol. 42, pp. 1115-1127.

Särndal, C. E., Swenson, B. and Wretman, J. H. (1992). Model Assisted Survey Sampling, Springer-Verlag, New York.

Yitzhaki, S. (1991). Calculating Jackknife Variance Estimators for Parameters of the Gini Method, Journal of Business & Economic Statistics, Vol. 9, pp. 235-239.