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

Yves G. Berger

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 generalized 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.

1. The Gini Coefficient

In this section, we introduce some notations, define the Gini coefficient (Gini 1914) and define its estimators. Consider a finite population denoted by \( U = \{1, \ldots, i, \ldots, N\} \), 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)

\[
\gamma = \frac{1}{\tau} \sum_{i \in U} (2F(y_i) - 1)y_i
\]

(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 reexpressed as

\[
\gamma = \gamma^* + \frac{1}{N}
\]

(2)

\(^{1}\) University of Southampton, Southampton Statistical Sciences Research Institute, Highfield, Southampton, SO17 1BJ, UK. Email: Y.G.Berger@soton.ac.uk

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.
where
\[ \gamma^* = \frac{2 \text{cov}(y, F(y))}{\mu} \] (3)
and \( \mu = \tau/N \). The quantity \( \gamma^* \) is an alternative expression for the Gini coefficient proposed by Anand (1983) and by Lerman and Yitzhaki (1984).

More generally, \( \gamma \) becomes \( \gamma^* \) when we replace \( F(y_i) \) in (1) with the smooth (or mid-interval) distribution function \( F^*(y_i) = \frac{1}{2} [F(y_i) + F(y_i - 0)] \), where \( F(y_i - 0) = \lim_{y \downarrow y_i} F(y) \). 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 \( I \) denotes 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 \( y_i (i \in s) \). Thus, it has to be estimated from the observed sampled values \( y_i (i \in s) \). A substitution estimator for \( \gamma \) is given by (Kovacevic and Binder 1997):
\[ \hat{\gamma} = \frac{1}{N} \sum_{i \in s} w_i (2\hat{F}(y_i) - 1)y_i \] (4)
where
\[ \hat{F}(y) = \frac{1}{N} \sum_{i \in s} w_i \delta\{y_i \leq y\} \] (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 and Thompson (1952) weights of individual \( i \). The quantity \( \pi_i \) is the first-order inclusion probability of \( i \); that is, the probability that individual \( i \) is in the sample. Using the Horvitz–Thompson weights guarantees that \( \hat{\gamma} \) is an approximately unbiased estimator for \( \gamma \).

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 \( \hat{F}^*(y_i) = \frac{1}{2} [\hat{F}(y_i) + \hat{F}(y_i - 0)] \), where \( \hat{F}(y_i - 0) = \lim_{y \downarrow y_i} \hat{F}(y) \). 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, \( \hat{\gamma} = \hat{\gamma}^* \) when \( |\nu| = 0 \) 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 $v$ 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 to be 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 article, 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 and Rao 1981; Robinson and Särndal 1983; Särndal et al. 1992, p. 175; Andersson and Nordberg 1994; Deville 1999) is to use "pseudo-values" $z_j$ such that \( \text{var}(\hat{\gamma}) \approx \text{var}(\hat{\tau}_c) \), where $\hat{\tau}_c = \sum_{s \in S} w_i z_{i}$. The approximation \( \approx \) is justified by some large-sample arguments (see Krewski and 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 $\hat{\tau}_c$.

This estimator is given by

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

(6)

where $\hat{\Delta}_{ij} = (\pi_{ij} - \pi_i \pi_j)^{-1}$, and $\pi_{ij}$ 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

\[ z_j = \frac{1}{\tau} \left[ 2y_j \hat{F}(y_j) - (\hat{y} + 1)y_j \right] \]  

(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 \( \hat{F}(y_j) \) has a nonnegligible contribution into the variance of \( \hat{y} \).

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

\[ z_j = \frac{1}{\tau} \left[ 2y_j \hat{F}(y_j) - (\hat{y} + 1)y_j + \frac{\hat{\tau}}{N} + \frac{2}{N} \sum_{i \neq j} w_i y_i \delta \{ y_j \leq y_i \} \right] \]  

(8)

In Sections 3 and 4, the linearization estimator in (6) with \( z_j \) given by (8) will be compared with the generalized 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; 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 generalized jackknife variance estimator that fully captures the effect 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 \( \hat{y} \) is not expressible as a function of means, we show in this section that the generalized jackknife variance estimator is a consistent estimator for the variance of \( \hat{y} \) provided that the linearization estimator in (6) is consistent.

Campbell’s generalized jackknife variance estimator (see also Berger and Skinner 2005) is given by

\[ \text{vár}(\hat{y})_{GJ} = \sum_{i \neq j} \sum_{j \neq k} \hat{\lambda}_{ij} w_i w_j z_j \]  

(9)

where the quantities \( z_j \) are pseudo-values:

\[ z_j = w_j^{-1}(1 - w_j \hat{N}^{-1})(\hat{y} - \hat{y}_{(j)}) \]  

(10)
with
\[ \hat{g}_{(j)} = \frac{1}{\hat{\pi}_{(j)}} \sum_{i \in S_{(j)}} w_i (2 \hat{F}(y_i) \lambda_{(j)} - 1) y_i \]
\[ \hat{F}(y_{(j)}) = \frac{1}{\hat{N}_{(j)}} \sum_{i \in S_{(j)}} w_i \delta[y_i \leq y] \]
\[ \hat{\pi}_{(j)} = \sum_{i \in S_{(j)}} w_i y_i \]
\[ \hat{N}_{(j)} = \sum_{i \in S_{(j)}} w_i \]
and \( s_{(j)} = s \{ j \} \), 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
\[ \text{var}(\hat{g})_{CJ} = \left(1 - \frac{n}{N}\right) \frac{1}{n(n-1)} \sum_{i \in S} (\hat{\gamma}_i - \hat{\gamma})^2 \] (11)
where \( \hat{\gamma} = n \hat{\gamma} - (n - 1) \hat{\gamma}_{(j)} \) and \( \hat{\gamma} = (1/n) \sum_{i \in S} \hat{\gamma}_i \). Moreover, the generalized jackknife estimator in (9) remains consistent under unequal probabilities sampling (Berger and Skinner 2005), 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 rewritten as
\[ \tilde{z}_j = \frac{\hat{\pi}_{(j)} \hat{z}_j}{\hat{\pi}_{(j)} \hat{\pi}_{(j)}} - 2 \frac{w_i y_i}{\hat{N}_{(j)} \hat{\pi}_{(j)}} \] (12)
where \( \hat{z}_j \) is given by (8). This means that \( \tilde{z}_j \) is approximately equal to \( z_j \) given by (8), provided that \( \hat{\pi}_{(j)} \hat{\pi}_{(j)} = 1 + O_p(1/n) \) and \( w_i y_i / (\hat{N}_{(j)} \hat{\pi}_{(j)}) = O_p(1/n) \). 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 naive 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 = .8, scale = 1). We focus on these distributions as they are a 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 \( \pi_i \). 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 \( \sigma^2_e = (1 - \rho^2)(N - 1)^{-1} \sum_{i \in U} (y_i - \mu)^2 \), \( \alpha = 5 + \rho \mu \), \( \rho = 0.7 \), and \( \mu = \tau/N \) is
<table>
<thead>
<tr>
<th></th>
<th>Gamma $\gamma = 0.34$ and $\gamma^* = 0.34$</th>
<th>Lognormal $\gamma = 0.28$ and $\gamma^* = 0.27$</th>
<th>Weibull $\gamma = 0.60$ and $\gamma^* = 0.60$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E(\hat{\gamma})$</td>
<td>$E(\hat{\gamma}^*)$</td>
<td>$\text{var}(\hat{\gamma})/\text{var}(\hat{\gamma}^*)$</td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.47</td>
<td>0.28</td>
<td>0.83</td>
</tr>
<tr>
<td>25</td>
<td>0.37</td>
<td>0.33</td>
<td>0.96</td>
</tr>
<tr>
<td>50</td>
<td>0.35</td>
<td>0.33</td>
<td>0.98</td>
</tr>
<tr>
<td>100</td>
<td>0.34</td>
<td>0.34</td>
<td>0.99</td>
</tr>
</tbody>
</table>
the population mean of the \( y_i \). The \( x_i \) are treated as fixed after they are generated. The \( \pi_{ij} \) 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

\[
RB = \frac{\text{Bias}(\text{var}(\hat{y}))}{\text{MSE}(\hat{y})}
\]

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

\[
\text{Bias}(\text{var}(\hat{y})) = \frac{1}{B} \sum_{b=1}^{B} \text{var}(\hat{y})_b - \text{var}(\hat{y})
\]

and

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

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

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

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

where

\[
E(\hat{y}) = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b
\]

The empirical relative root mean squared error of \( \text{var}(\hat{y}) \) is

\[
\text{RRMSE}(\text{var}(\hat{y})) = \frac{\text{MSE}(\text{var}(\hat{y}))^{1/2}}{\text{MSE}(\hat{y})}
\]

where

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

Table 1 displays the empirical expectation of \( \hat{y} \) and \( \hat{y}^* \) and the ratio of their empirical variances under the distributions for several sample sizes. Table 1 shows that both \( \hat{y} \) and \( \hat{y}^* \) 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 \( \hat{y} \) and \( \hat{y}^* \) is of order \( 1/n^2 \) (see Section 1). Thus, the variance estimators developed here for estimating the variance of \( \hat{y} \) can also be used to estimate the variance of \( \hat{y}^* \) provided that the sample size is sufficiently large. For small sample sizes, \( \hat{y} \) and \( \hat{y}^* \) may be biased, and the linearization technique and the jackknife are not recommended for variance estimation.
Tables 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 \(z_b = (\hat{y}_b - \gamma) \text{vár}(\hat{y}_b)^{-1/2}\) and \(\delta(|z_b| \leq 1.96) = 1\) when \(|z_b| \leq 1.96, 0\) otherwise.

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 generalized jackknife (9).

The linearization and jackknife estimators also produce more reasonable confidence 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.

The two 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 generalized jackknife (9).

5. Discussion

This article has shown that the linearization technique proposed by Kováčević and Binder (1997) and the generalized 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 \(\pi_{ij}\) 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 \(\pi_{ij}\) employing only the first-order inclusion probabilities (see Berger 2004). Berger (2007) proposed a \(\pi_{ij}\)-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.
|       | Gamma $\gamma = 0.34$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|-------|----------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5     | 209%                | -6.3%  | 7.1% | 5.2% | 254% | -5.7%  | 4.5% | 5.1% | 127% | -30.1% | 24.9% | 26.8% | 104% | 10.5% | 10.2% | 10.0% | 102% | -3.9% | 11.0% | 4.6% |
| 25    | 366                 | -4.0%  | 4.4% | 2.8% | 522  | -5.2%  | 3.0% | 2.9% | 104  | 10.5%  | 9.2%  | 6.4%  | 102  | -3.9% | 11.0% | 4.6% | 93   | -0.9% | 18.2% | 3.0% |
| 50    | 391                 | -4.9%  | -0.8% | -0.9% | 598  | -4.9%  | 1.5% | 0.1% | 102  | -3.9% | 11.0% | 4.6% | 93   | -0.9% | 18.2% | 3.0% | 73   | -0.2% | 29.0% | 2.1% |
| 100   | 394                 | -2.8%  | -2.8% | -0.6% | 694  | 0.8%   | 8.4% | 3.7% | 93   | -0.9% | 18.2% | 3.0% | 73   | -0.2% | 29.0% | 2.1% | 73   | -0.2% | 29.0% | 2.1% |
| 150   | 369                 | -2.7%  | -5.3% | -1.2% | 692  | -3.0%  | 7.5% | -1.1% | 73   | -0.2% | 29.0% | 2.1% | 73   | -0.2% | 29.0% | 2.1% | 73   | -0.2% | 29.0% | 2.1% |
Table 3. Empirical RRMSE (%) of the variance estimator based upon (7), (8), (9) and (11) for the three distributions and several sample sizes

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

<table>
<thead>
<tr>
<th>n</th>
<th>Gamma $\gamma = 0.34$</th>
<th>Lognormal $\gamma = 0.28$</th>
<th>Weibull $\gamma = 0.60$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linearization (6)</td>
<td>Linearization (6)</td>
<td>Linearization (6)</td>
</tr>
<tr>
<td></td>
<td>Jackknife</td>
<td>Linearization (6)</td>
<td>Jackknife</td>
</tr>
<tr>
<td></td>
<td>(7)</td>
<td>(8)</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>(11)</td>
<td>(9)</td>
</tr>
<tr>
<td>5</td>
<td>99%</td>
<td>100%</td>
<td>93%</td>
</tr>
<tr>
<td></td>
<td>55%</td>
<td>38%</td>
<td>73%</td>
</tr>
<tr>
<td></td>
<td>69%</td>
<td>56%</td>
<td>87%</td>
</tr>
<tr>
<td></td>
<td>68%</td>
<td>56%</td>
<td>87%</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>100</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>89</td>
<td>89</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>91</td>
<td>93</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>93</td>
<td>93</td>
<td>95</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td>150</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>94</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>94</td>
<td>94</td>
<td>95</td>
</tr>
</tbody>
</table>
Large national household surveys often employ two-stage or multistage sampling. For such surveys, the joint inclusion probabilities \( p_{ij} \) will often not be known, and stage-wise approximations to them may be necessary. For that reason the generalized 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). Generalizing these methods to Campbell’s jackknife in (9) would be a fruitful direction for future research.

Appendix – Proof of (12)

Using

\[
\hat{\gamma} = 2 \sum_{i \in s} w_{yi} \hat{F}(y_i) - 1
\]

it can be shown that

\[
\hat{\eta}_{(j)} = \frac{2}{N_{(j)} \hat{\eta}_{(j)}} \left[ \sum_{i \in s} w_{yi} \left( \sum_{i \in s} w_{yi} \delta_{ji} - w_{j} \delta_{ji} \right) \right] - 1,
\]

\[
= \frac{2}{N_{(j)} \hat{\eta}_{(j)}} \left[ \hat{N} \sum_{i \in s} w_{yi} \hat{F}(y_i) - w_{j} \sum_{i \in s} w_{yi} \delta_{ji} - w_{j} \sum_{i \in s} w_{yi} \hat{F}(y_i) + w_{j}^2 \hat{F}(y_j) - 1, \right.
\]

\[
= \frac{2}{N_{(j)} \hat{\eta}_{(j)}} \left[ (\hat{\gamma} + 1) \frac{\hat{N}^2}{2} - w_{j} \sum_{i \in s} w_{yi} \delta_{ji} - w_{j} \sum_{i \in s} w_{yi} \hat{F}(y_i) + w_{j}^2 \hat{F}(y_j) - 1, \right.
\]

where \( \delta_{ji} = \delta(y_j \leq y_i) \). Thus,

\[
\hat{\gamma} - \hat{\gamma}_{(j)} = \frac{2}{N_{(j)} \hat{\eta}_{(j)}} \left[ w_{j} \hat{F}(y_j) -(\hat{\gamma} + 1) \frac{\hat{N}^2}{2} - w_{j} \sum_{i \in s} w_{yi} \delta_{ji} - w_{j}^2 \hat{F}(y_j) \right]
\]

\[
= \frac{w_{j}}{N_{(j)} \hat{\eta}_{(j)}} \left[ 2 w_{j} \hat{F}(y_j) + (\hat{\gamma} + 1) \frac{\hat{N}^2}{w_{j}} + 2 \sum_{i \in s} w_{yi} \delta_{ji} \right].
\]
We have $\hat{N}(j) = \hat{N}/\hat{N}(j) - N/\hat{N}(j) - w_j/N = -w_j(y_j\hat{N} + \hat{\sigma})$ which substituted into (13) gives

$$\hat{\gamma} - \hat{\gamma}(j) = \frac{w_j}{N(j)\hat{N}(j)} \left[ 2y_j\hat{N}\hat{F}(y_j) - (\hat{\gamma} + 1)(y_j\hat{N} + \hat{\sigma}) + 2\sum_{i\in j} w_{ji}\delta_{ji} - 2w_jy_j \right]$$

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

$$\hat{z}_j = \frac{1}{N\hat{N}(j)} \left[ 2y_j\hat{N}\hat{F}(y_j) - (\hat{\gamma} + 1)(y_j\hat{N} + \hat{\sigma}) + 2\sum_{i\in j} w_{ji}\delta_{ji} - 2w_jy_j \right]$$

which implies (12). This completes the proof.

6. References


Received June 2006
Revised March 2008