Variance estimation of inequality indices in complex sampling designs

Antal, Erika
University of Neuchâtel
rue de la Pierre-à-Mazel 7
2000 Neuchâtel, Switzerland
E-mail: erika.antal@unine.ch

Langel, Matti
University of Neuchâtel
E-mail: matti.langel@unine.ch

Tillé, Yves
University of Neuchâtel
E-mail: yves.tille@unine.ch

1 Introduction

Variance estimation for non-linear functions of interest under a complex sampling framework can be conducted by different methods. We propose here some new results in bootstrap and linearization techniques for variance estimation of non-linear statistics. Different inequality measures are studied, with emphasis put on the well-known Gini index. A simulation study using a Poisson sampling design is conducted in order to compare the performance of the different methods when estimating the sampling variance of the Gini index.

In survey sampling, we are interested in a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N from which a random sample S of size n is selected by means of a sampling design $p(s) = \Pr(S = s)$, for all $s \in U$. Define the inclusion probabilities $\pi_k = \Pr(k \in S), k \in U$, and the joint inclusion probabilities $\pi_{k\ell} = \Pr(k \in S \text{ and } \ell \in S), k, \ell \in U$. The sampling weights w_k can be equal to the Horvitz and Thompson (1952) weights $d_k = 1/\pi_k$ or can have been improved by a calibration technique (Deville and Särndal, 1992) or a non-response adjustment. Let $y_1, \ldots, y_k, \ldots, y_N$ denote the characteristic of interest (here, the income) of the units in the population and $y_{(1)}, \ldots, y_{(k)}, \ldots, y_{(N)}$ the same incomes ordered in increasing order. In order to estimate totals

$$Y = \sum_{k \in U} y_k \text{ and } N = \sum_{k \in U} 1,$$

one can use weighted estimators

$$\widehat{Y} = \sum_{k \in S} w_k y_k \text{ and } \widehat{N} = \sum_{k \in S} w_k.$$

We now focus on performing finite population inference on a non-linear function of interest θ estimated by means of a complex sample. Specific variance estimation methods, such as bootstrap or linearization, are thus required.

In the next section, various approaches to linearization are presented. Section 3 introduces three inequality measures : the Gini index, the Quintile Share Ratio and the Zenga index. A finite population estimator and a linearized variable are proposed for each measure. The bootstrap approach is then presented in Section 4, while a new bootstrap method for a Poisson sampling design is suggested in Section 5. The paper ends with a comparative simulation study on the Gini index and final discussions.

2 Linearization techniques for variance estimation

Linearization combines a range of techniques for calculating an approximation of the variance of a non-linear statistic. It consists in approximating $\widehat{\theta}$ by a sum of terms, i.e. finding a linearized variable v_k such that

$$\widehat{\theta} - \theta \approx \sum_{k \in S} w_k v_k - \sum_{k \in U} v_k.$$

Next, the variance of $\widehat{\theta}$ is simply approximated by the variance of the estimated total $\sum_{k \in S} w_k v_k$. Nevertheless, the v_k's often depend on parameters of the population that must be estimated. By estimating these parameters, one can obtain $\widehat{v}_k$ an estimator of v_k and thus construct an estimator of the variance by plugging $\widehat{v}_k$ in the expression of the variance of a total corresponding to the given sampling design. For instance, for a Poisson sampling design, we have the estimator

$$\widehat{\text{var}}_{lin}(\widehat{\theta}) = \sum_{k \in S} \left(\frac{\widehat{v}_k}{\pi_k} \right)^2 (1 - \pi_k). \tag{1}$$

For details on the asymptotic framework validating linearization, one can relate to Isaki and Fuller (1982), Deville and Särndal (1992) and Deville (1999). If the function of interest is a smooth function of totals, the most straightforward way of deriving a linearized variable is by performing a Taylor series expansion with respect to these totals (Woodruff, 1971). However, most inequality measures are not smooth functions of totals and require other approaches. Three interesting approaches are presented hereafter.

2.1 Deville approach

The influence function proposed initially by Hampel (1974) and Hampel et al. (1985) is a tool first proposed to study the robustness of an estimator but can also be used to approximate the variance. Deville (1999) proposes a modified version of the influence function in order to compute a linearized variable for sampling from a finite population. In order to define the influence function, Deville uses a measure M with unit mass for each point of the population. According to Deville's definition, the measure M is positive, discrete, with a total mass N while the total mass is equal to 1 for the influence function proposed by Hampel (1974). A function of interest θ can be presented as a functional $T(M)$ that associates for each measure a real number or a vector. For instance, a total Y can be written

$$Y = \int y dM = \sum_{k \in U} y_k.$$

Besides, we also suppose that the considered functionals are linear and homogeneous in the sense that there always exists a real number α such that $T(tM) = t^{\alpha} T(M)$, for all $t \in \mathbb{R}$. Coefficient α is called the degree of the functional $T(M)$. The measure M is estimated by a measure $\widehat{M}$ that has a mass equal to w_k for each point x_k of sample S. The plug-in estimator of a functional $T(M)$ is simply $T(\widehat{M})$. For instance, the estimator of a total is given by

$$\int y d\widehat{M} = \sum_{k \in S} w_k y_k.$$

Deville's influence function is defined by

$$IT(M, x) = \lim_{t \to 0} \frac{T(M + t\delta_x) - T(M)}{t},$$

when this limit exists, where δ_x is the Dirac measure at point x. This influence function is the Gâteaux differential in the direction of the Dirac mass at point x. Deville (1999) shows that this

influence function $v_k = IT(M, x_k)$ is a linearized variable of $T(\widehat{M})$ in the sense that it allows for the approximation of the interest function:

$$\frac{T(\widehat{M}) - T(M)}{N^\alpha} \approx \frac{1}{N^\alpha}\left(\sum_{k\in S} w_k v_k - \sum_{k\in U} v_k\right).$$

Computation of influence functions follows the rules of differential calculus (Deville, 1999). We propose hereafter an additional result that enables us to compute directly the linearized variable of a double sum $S = \sum_{k\in U}\sum_{\ell\in U}\phi(y_k, y_\ell)$.

Result 1. *If*

$$S(M) = \int\int \phi(x,y)dM(x)dM(y),$$

where $\phi(.,.)$ is a function from $\mathbb{R}^2$ in $\mathbb{R}$, then

$$IS(M,\xi) = \int \phi(x,\xi)dM(x) + \int \phi(\xi,y)dM(y).$$

If $\phi(x,y) = \phi(y,x)$ for all x, y then the influence function can simply be written as

$$IS(M,\xi) = 2\int \phi(x,\xi)dM(x).$$

2.2 Demnati and Rao approach

A fast technique to obtain a direct linearized variable consists in computing the Deville influence function, not on the measure M but on the estimated measure $\widehat{M}$. We then obtain

$$IT(\widehat{M}, x_k) = \lim_{t\to 0}\frac{T(\widehat{M} + t\delta_x) - T(\widehat{M})}{t}.$$

Measure $\widehat{M}$ has a mass equal to w_k for each point x_k of the sample. If we refer to the definition of the derivative, we can notice that a simple way for obtaining the linearized variable is to simply differentiate the estimate with respect to w_k

$$IT(\widehat{M}, y_k) = \frac{\partial T(\widehat{M})}{\partial w_k}.$$

The computation of a simple derivative with respect to the weights is advocated by Demnati and Rao (2004) in order to compute the linearized variable of a function of totals. This method also allows for the computation of a linearized variable for any function of interest whose observations are weighted by w_k.

2.3 Graf approach

In a recent paper, Graf (2010) proposes another way of computing the linearized variable by applying a Taylor expansion with respect to the indicator variables I_k, where for all $k \in U$

$$I_k = \begin{cases} 1 & \text{if } k \in S, \\ 0 & \text{if } k \notin S, \end{cases}$$

determines the presence of unit k in the sample. The Graf method is coherent because the expansion is done with respect to the only source of randomness in the estimator.

3 Inequality measures

Three inequality measures are studied here, the Gini index, the Quintile Share Ratio and the Zenga index. All these measures can be defined in the continuous case by means of the Lorenz (1905) curve given by

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)dy}{\int_0^{\infty} yf(y)dy} = \frac{1}{\mu}\int_0^{\alpha} F^{-1}(u)du,$$

where $f(y)$ is a probability density function of a positive continuous random variable Y that represents the income, $F(y)$ is its cumulative distribution function, $F^{-1}(.)$, the inverse function of $F(.)$ and

$$\mu = \int_0^{\infty} yf(y)dy.$$

Thus the Gini index, Quintile Share Ratio and Zenga index are respectively defined by

$$G = 1 - 2\int_0^1 L(\alpha)d\alpha,$$

$$QSR = \frac{1 - L(0.8)}{L(0.2)},$$

$$Z = 1 - \int_0^1 \frac{L(\alpha)}{\alpha} \cdot \frac{1-\alpha}{1-L(\alpha)} d\alpha.$$

In the following a finite population estimator for each measure is presented as well as an estimated linearized variable, allowing for variance estimation.

3.1 Notation and Definition

The total income of the αN poorest units is defined by $\widetilde{Y}(\alpha) = \sum_{k\in U} y_k \mathbb{1}[y_k \leq Q_\alpha]$, where Q_α is the α-quantile and $\mathbb{1}[A]$ is an indicator function equal to 1 if A is true and 0 otherwise. This definition is however not very accurate because the quantiles can be defined in several different ways when the cumulative distribution function is a step function (see Hyndman and Fan, 1996). We thus prefer to use the less ambiguous definition of the total income of the αN poorest units proposed in Langel and Tillé (2011b) and given by

$$Y(\alpha) = \sum_{k\in U} y_{(k)} H\left[\alpha N - (k-1)\right],$$

where $H(.)$ is the cumulative distribution function of a uniform [0,1] random variable

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

The function of interest $Y(\alpha)$ is then strictly increasing in α in $(0,1)$, which is not the case of $\widetilde{Y}(\alpha)$. In order to estimate $Y(\alpha)$, we can use

$$\widehat{Y}(\alpha) = \sum_{k\in S} w_k y_{(k)} H\left(\frac{\alpha\widehat{N} - \widehat{N}_{k-1}}{w_k}\right),$$

where $\widehat{N}_k$ are the cumulative weights according to the y_k ordered non-decreasingly in the sample, i.e.

$$\widehat{N}_k = \sum_{\ell\in S} w_\ell \mathbb{1}[y_\ell \leq y_k],$$

and $\widehat{N}_0 = 0$. In a finite population, the Lorenz curve can then be defined by $L(\alpha) = Y(\alpha)/Y$ and estimated by $\widehat{L}(\alpha) = \widehat{Y}(\alpha)/\widehat{Y}$. Accordingly, functions $L(\alpha)$ and $\widehat{L}(\alpha)$ are also strictly increasing in $(0,1)$.

3.2 The Gini index

A finite population estimator of the Gini index is

$$\widehat{G} = \frac{2}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k \widehat{N}_k y_k - \left(1+\frac{1}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k^2 y_k\right). \tag{2}$$

An estimated linearized variable for the Gini index is

$$\widehat{v}_k^G = \frac{1}{\widehat{N}\widehat{Y}}\left[2\widehat{N}_k(y_k-\widehat{\overline{Y}}_k)+\widehat{Y}-\widehat{N}y_k-\widehat{G}(\widehat{Y}+y_k\widehat{N})\right]. \tag{3}$$

where

$$\widehat{\overline{Y}}_k = \frac{\sum_{\ell\in S} w_\ell y_\ell \mathbb{1}(y_\ell \le y_k)}{\widehat{N}_k}.$$

This result is not new (Monti, 1991; Cowell and Victoria-Feser, 1996, 2003; Deville, 1999; Binder and Kovacevic, 1995) and can be obtained using, among others, any of the three methods proposed above. Note also that Result 1 allows for a very fast derivation of (3).

3.3 Quintile Share Ratio

The QSR is defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. It can be estimated from a sample by

$$\widehat{\mathrm{QSR}} = \frac{\widehat{Y}-\widehat{Y}(0.8)}{\widehat{Y}(0.2)}.$$

An estimated linearized variable for the QSR is (Langel and Tillé, 2011b)

$$\begin{aligned}\widehat{v}_k^{QSR} = & \frac{y_k - \left\{y_k H\left(\frac{0.8\widehat{N}-\widehat{N}_{k-1}}{w_k}\right)+\widehat{\widetilde{Q}}_{0.8}\left[0.8-\mathbb{1}\left(y_k<\widehat{\widetilde{Q}}_{0.8}\right)\right]\right\}}{\widehat{Y}(0.2)} \\ & -\frac{\left(\widehat{Y}-\widehat{Y}(0.8)\right)\left\{y_k H\left(\frac{0.2\widehat{N}-\widehat{N}_{k-1}}{w_k}\right)+\widehat{\widetilde{Q}}_{0.2}\left[0.2-\mathbb{1}\left(y_k<\widehat{\widetilde{Q}}_{0.2}\right)\right]\right\}}{\widehat{Y}(0.2)^2},\end{aligned}$$

where $\widehat{\widetilde{Q}}_\alpha = y_i$, with $\widehat{N}_{i-1} < \alpha\widehat{N} \le \widehat{N}_i$.

3.4 Zenga index

Let $\widehat{Y}_k = \sum_{\ell\in S} w_\ell y_\ell \mathbb{1}[\ell \le k]$ and $\widehat{A}_k = \widehat{N}_{k-1}y_k - \widehat{Y}_{k-1}$ for $k = 2, ..., n$. A finite population estimator for the Zenga index can then be written $\widehat{Z} = \sum_{k\in S}\widehat{Z}_k$, where

$$\widehat{Z}_k = \begin{cases} \left(\frac{\widehat{Y}}{\widehat{N}y_1}-1\right)\log\left(\frac{\widehat{Y}}{\widehat{Y}-\widehat{Y}_1}\right), & \text{if } k=1, \\ \frac{\widehat{A}_k}{\widehat{Y}+\widehat{A}_k}\log\left(\frac{\widehat{N}_k}{\widehat{N}_{k-1}}\right)+\left[\frac{\widehat{Y}}{\widehat{N}y_k}-\frac{\widehat{Y}}{\widehat{Y}+\widehat{A}_k}\right]\log\left(\frac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right), & \text{if } k=2,...,n-1, \\ \left(1-\frac{\widehat{Y}}{\widehat{N}y_n}\right)\log\left(\frac{\widehat{N}}{\widehat{N}_{n-1}}\right), & \text{if } k=n. \end{cases}$$

The Demnati and Rao approach has been applied by Langel and Tillé (2011a) to derive an estimated linearized variable $\widehat{v}_\ell^Z$ for the Zenga Index:

$$\widehat{v}_\ell^Z = \frac{\partial\widehat{Z}}{\partial w_\ell} = \sum_{k\in S}\frac{\partial\widehat{Z}_k}{\partial w_\ell},$$

with

$$\frac{\partial \widehat{Z}_k}{\partial w_\ell} = \begin{cases} \dfrac{\widehat{N}y_\ell - \widehat{Y}}{\widehat{N}^2 y_1}\log\left(\dfrac{\widehat{Y}}{\widehat{Y}-\widehat{Y}_1}\right) + y_\ell\left(\dfrac{\widehat{Y}}{\widehat{N}y_1}-1\right)\left[\dfrac{1}{\widehat{Y}} - \dfrac{\mathbb{1}(\ell>1)}{\widehat{Y}-\widehat{Y}_1}\right], & \text{if } k=1, \\ \dfrac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell<k) - \widehat{A}_k y_\ell}{\left(\widehat{Y}+\widehat{A}_k\right)^2}\log\left[\dfrac{\widehat{N}_k\left(\widehat{Y}-\widehat{Y}_{k-1}\right)}{\widehat{N}_{k-1}\left(\widehat{Y}-\widehat{Y}_k\right)}\right] \\ \quad + \dfrac{\widehat{A}_k}{\widehat{Y}+\widehat{A}_k}\left[\dfrac{\mathbb{1}(\ell\le k)}{\widehat{N}_k} - \dfrac{\mathbb{1}(\ell<k)}{\widehat{N}_{k-1}}\right] + \dfrac{\widehat{N}y_\ell - \widehat{Y}}{\widehat{N}^2 y_k}\log\left(\dfrac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right) \\ \quad + \dfrac{\widehat{Y}y_\ell}{\widehat{Y}-\widehat{Y}_{k-1}}\left[\mathbb{1}(\ell=k) - \dfrac{y_k w_k}{\widehat{Y}-\widehat{Y}_k}\mathbb{1}(\ell>k)\right]\left(\dfrac{1}{\widehat{N}y_k} - \dfrac{1}{\widehat{Y}+\widehat{A}_k}\right), & \text{if } k=2,...,n-1, \\ \dfrac{\widehat{Y}-\widehat{N}y_\ell}{\widehat{N}^2 y_n}\log\left(\dfrac{\widehat{N}}{\widehat{N}_{n-1}}\right) + \left(1-\dfrac{\widehat{Y}}{\widehat{N}y_n}\right)\left[\dfrac{1}{\widehat{N}} - \dfrac{\mathbb{1}(\ell<n)}{\widehat{N}_{n-1}}\right], & \text{if } k=n. \end{cases}$$

4 Bootstrap

A bootstrap sample is a random sample with replacement selected from S. Let S_k^* be the number of times unit k is repeated in the bootstrap sample. For a general estimator $\widehat{\theta}$ the bootstrap estimator is given by $\widehat{\theta}^*$. For the case of the total Y, which is unbiasedly estimated by the Horvitz-Thompson estimator (HT) (Horvitz and Thompson, 1952), $\widehat{Y}_\pi = \sum_{k\in S}\frac{y_k}{\pi_k}$, the bootstrap estimator can be written as $\widehat{Y}^* = \sum_{k\in S}\frac{y_k}{\pi_k}S_k^*$.

Let $\Pr^*(.) = \Pr(.|S)$, $\mathrm{E}^*(.) = \mathrm{E}(.|S)$ and $\mathrm{var}^*(.) = \mathrm{var}(.|S)$ respectively denote the probability, the expectation and the variance of the bootstrap sample conditionally to the original sample. This gives us

$$\mathrm{E}^*(\widehat{Y}^*) = \sum_{k\in S}\frac{y_k}{\pi_k}\mathrm{E}(S_k^*),$$

and

$$\mathrm{var}^*(\widehat{Y}^*) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\mathrm{cov}(S_k^*, S_\ell^*|S).$$

To calculate the variance of the estimator of the total and its estimator let define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell, k\neq\ell\in U$, and $\breve{\Delta}_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$. When $k=\ell$, we obtain $\Delta_{kk} = \pi_k(1-\pi_k), k\in U$, and $\breve{\Delta}_{kk} = 1-\pi_k$. With this notation, the variance of the Horvitz-Thomson estimator is given by

$$\mathrm{var}(\widehat{Y}_\pi) = \sum_{k\in U}\sum_{\ell\in U}\frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell},$$

and can be unbiasedly estimated by using the Horvitz-Thompson (HT) variance estimator:

$$\widehat{\mathrm{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S}\frac{y_k y_\ell}{\pi_k\pi_\ell}\breve{\Delta}_{k\ell}.$$

Thus a necessary and sufficient condition for the bootstrap estimator of the total to be unbiased is given by

$$\mathrm{E}^*(S_k^*) = 1, k\in S. \tag{4}$$

Moreover, in order to have an unbiased variance estimator of the total, a first condition is that

$$\mathrm{var}^*(S_k^*) = \breve{\Delta}_{kk} = 1-\pi_k, k\in S. \tag{5}$$

Ideally, another condition for a bootstrap method to unbiasedly estimate the variance of the HT-estimator is that

$$\text{cov}(S_k^*, S_\ell^*|S) = \check{\Delta}_{k\ell}, k \neq \ell \in S. \quad (6)$$

These conditions on the covariances are however difficult to meet when the sample is selected with fixed sample size and unequal inclusion probabilities. In this particular case, it is difficult to exactly satisfy more than conditions (4) and (5). Condition (6) can however be approximately satisfied.

The bootstrap estimator of the variance of a statistic of interest $\widehat{\text{var}}_{boot}(\widehat{\theta})$ is computed by generating a set of bootstrap samples and then computing the $\text{var}(\widehat{\theta}^*)$, the variance of the outcomes of $\widehat{\theta}^*$. Moreover, if a bootstrap method provides an approximately unbiased estimator for the variance of totals, it will also provide approximately unbiased variance estimators for smooth functions of totals.

5 Bootstrap for Poisson design

In a Poisson design with inclusion probabilities π_k,

$$p(s) = \pi_k^{\mathbb{1}(k\in s)}(1-\pi_k)^{\mathbb{1}(k\notin s)} \text{ for all } s \subset U,$$

where $\mathbb{1}(A) = 1$ is equal to 1 if A is true and 0 otherwise. The inclusion probability is $\Pr(k \in S) = \pi_k$. Moreover, $\pi_{k\ell} = \pi_k\pi_\ell$ when $k \neq \ell \in U$ and $\pi_{kk} = \pi_k$. Thus $\Delta_{k\ell} = 0$, when $k \neq \ell \in U$ and $\Delta_{kk} = \pi_k(1-\pi_k)$. We thus have, $\check{\Delta}_{k\ell} = 0$, when $k \neq \ell \in U$ and $\check{\Delta}_{kk} = 1-\pi_k$. With Poisson sampling design the sample size n is random thus the estimator of variance is calculated by $\widehat{\text{var}}_{HT}(\widehat{Y}_\pi)$.

Patak and Beaumont (2009) have proposed a bootstrap method for Poisson design that uses normal independent variables with expectation equal to 1 and variances equal to $1-\pi_k$ thus

$$S_k^* = N(1, 1-\pi_k).$$

Unfortunately, this method requires the use of non-integer weights that can be negative. Instead we recommend the use of a discrete random variable for S_k^*.

Antal and Tillé (2011) have proposed a simple bootstrap method that uses n independent Bernoulli random variables X_k with parameter π_k and n independent Poisson random variables Z_k with parameter $\lambda = 1$. For this method, the bootstrap sample is given by

$$S_k^* = X_k + (1-X_k)Z_k, k \in S.$$

Thus, the probability mass function of S_k^* is given by:

$$\Pr^*(S_k^* = r) = \pi_k\mathbb{1}[r=1] + \frac{(1-\pi_k)}{e\cdot r!}, r = 0, 1, 2, \ldots$$

where $e \approx 2.71$ is the Euler constant. The bootstrap variable S_k^* satisfies conditions (4), (5), and (6).

An even simpler method is to consider n independent Bernoulli random variables $X_k, k \in S$ with parameter π_k and n independent Bernoulli random variables Y_k with parameter $1/2$. Define the bootstrap sample by

$$S_k^* = X_k + 2(1-X_k)Y_k, k \in S.$$

The probability distribution of S_k^* is thus

$$S_k^* = \begin{cases} 0 & \text{with a probability } (1-\pi_k)/2 \\ 1 & \text{with a probability } \pi_k \\ 2 & \text{with a probability } (1-\pi_k)/2. \end{cases}$$

Again, the bootstrap variable S_k^* meets conditions (4), (5), and (6). Here, the bootstrap sample does not contain more than twice the same unit.

6 Simulation study

In order to compare the performance of different variance estimation methods for a nonlinear function of interest like the Gini index, we run simulations. The variance under the simulations, say the Monte Carlo variance, was considered as the true variance of the estimator. We generated a population of $N = 1500$ units from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, with $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0, 7)$, $\varepsilon_k \sim \mathcal{N}(0, 1)$ and $\sigma = 15$ with regression parameters $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. The model and its parameters were chosen intentionally to have a distribution for y similar to a lognormal - as it is often used for income distributions - with a correlated and positive explanatory variable x in the regression model. From this population, $sim = 1000$ samples were drawn using Poisson sampling design. Concerning the inclusion probabilities, they were calculated proportional to the values of a variable z, which was generated from equation $z = y^{0.2}p$ where $p \sim \ln\mathcal{N}(0, 0.25)$. In this manner, the correlation between y and z is about 0.5. We knowingly used a large sample rate 1/3 in expectation and a skewed population in order to better illustrate the performance of the tested methods. From each of these samples, we calculated the estimator of the Gini index as in Expression (2), and its variance estimator by linearization by plugging Expression (3) in (1).

From each of the 1000 initial samples, 1000 bootstrap samples were selected using four different bootstrap methods. Besides the new bootstrap method proposed (Method NEW), three other resampling methods were tested. The first one was the method proposed by Antal and Tillé (2011) (Method AT), the second one was the method of Patak and Beaumont (2009) (Method Patak-Beaumont) and the third one was the generalization of the bootstrap method without replacement proposed by Booth et al. (1994) for unequal inclusion probabilities (Method WOR). This last method is a variant of the initial bootstrap with replacement method (Gross, 1980; Chao and Lo, 1985) that consists of creating an artificial population from the sample and then drawing bootstrap samples from it with the same design as the initial one. After drawing the bootstrap samples, the bootstrap estimators and their variance were computed for each of the initial samples. The means of these variances were then compared with the approximations of the true variance. As the Gini index is not a smooth function of the total, estimating its variance can be difficult, but the simulations show that the methods perform well.

In order to measure the performance of the different methods, the following four indicators were used:

- Lower error rate (L) in %

$$\mathrm{L} = \frac{100}{sim}\sum_{i=1}^{sim} \mathbb{1}\left[\widehat{\theta} - 1.96 \times \sqrt{\widehat{\mathrm{var}}(\widehat{\theta})} > \theta\right].$$

- Upper error rate (U) in %

$$\mathrm{U} = \frac{100}{sim}\sum_{i=1}^{sim} \mathbb{1}\left[\widehat{\theta} + 1.96 \times \sqrt{\widehat{\mathrm{var}}(\widehat{\theta})} < \theta\right].$$

- Relative Bias

$$\mathrm{RB} = 100 \times \frac{\widehat{\mathrm{var}}(\widehat{\theta}) - \mathrm{var}_{sim}(\widehat{\theta})}{\mathrm{var}_{sim}(\widehat{\theta})} = 100 \times \frac{B}{\mathrm{var}_{sim}(\widehat{\theta})},$$

  where B is the Bias of $\widehat{\mathrm{var}}(\widehat{\theta})$.

- Relative Root Mean Squared Error

$$\mathrm{RRMSE} = 100 \times \frac{\sqrt{B^2 + \mathrm{var}[\widehat{\mathrm{var}}(\widehat{\theta})]}}{\mathrm{var}_{sim}(\widehat{\theta})}.$$

The RB gives a measure of the bias of the estimator of variance. The RRMSE measures its accuracy and in the case of unbiasedness of the variance estimator it is equal to the variation coefficients. The Error Rates which is the sum of the Lower error rate and the Upper error rate allows us to evaluate the capacity of the methods to provide a valid inference. The lower and the upper error rates give us an idea of how skewed the distribution of the estimator $\widehat{\theta}$ is.

Table 1: Performance of different methods to estimate the variance of the estimator of the Gini index in Poisson sampling

POISSON	L	U	Relative bias(%)	RRMSE(%)
GINI index				
Method AT	1.5	4.0	-1.8751	12.3487
Method NEW	1.5	3.6	-1.2133	12.3172
Method Patak-Beaumont	1.6	3.7	-1.1179	12.5330
Method WOR	1.6	3.1	-0.9928	12.5137
Linearization	1.5	3.8	-1.4180	11.4409

Table 1 presents the outcomes achieved using the Poisson sampling design with inclusion probabilities proportional to variable z. As the relative biases show, each method slightly underestimate the variance of the estimator of Gini, but these biases are really negligible, around 1%. The error rates show a slightly positively skewed distribution, with coverage rates around 95% and the RRMSE have also the same order for each of these methods. We can conclude that the methods provide essentially the same results, their performance are equivalent. An advantage of the linearization approach is that it is not computationally intensive. The differences between the bootstrap methods are in their applicabilities, their implementations and the after-treatments, as the calibration or the imputation for nonresponse. The bootstrap with replacement method uses artificial populations whose creation can be cumbersome because of the rounding problems. The Patak and Beaumont (2009) method apply non-integer weight that can even be negative. The method of Antal and Tillé (2011) and the method proposed uses non-integer positive weights, thus the bootstrap samples can be directly used to compute the variance of the functions of interest. Moreover with the new method, extreme samples are also avoided because the units can be repeated at most twice.

References

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite. *Accepted in Journal of the American Statistical Association*.

Binder, D. A. and Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equation approach. *Survey Methodology*, 21:137–145.

Booth, J. G., Butler, R. W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89:1282–1289.

Chao, M.-T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhyā*, A47:399–405.

Cowell, F. and Victoria-Feser, M.-P. (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality*, 1:191–219.

Cowell, F. A. and Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica*, 64:77–101.

Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30:17–34.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–204.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Graf, M. (2010). Use of survey weights for the analysis of compositional data. Working paper, Swiss federal statistical office.

Gross, S. T. (1980). Median estimation in sample surveys. In *ASA Proceedings of the Section on Survey Research Methods*, pages 181–184. American Statistical Association.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393.

Hampel, F. R., Ronchetti, E., Rousseuw, P. J., and Stahel, W. (1985). *Robust Statistics: The Approach Based on the Influence Function*. Wiley, New-York.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50:361–365.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77:89–96.

Langel, M. and Tillé, Y. (2011a). Inference by linearization for the Zenga inequality index: A comparison with the Gini index. *Technical report, University of Neuchatel*.

Langel, M. and Tillé, Y. (2011b). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141:2676–2985.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9:209–219.

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica*, 51:561–577.

Patak, Z. and Beaumont, J.-F. (2009). Generalized bootstrap for prices surveys. In *paper presented at the 57th Session of the International Statistical Institute, Durban, South-Africa*.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66:411–414.