UNIVERSITÀ
DI SIENA
1240

QUADERNI DEL DIPARTIMENTO
DI ECONOMIA POLITICA E STATISTICA

Riccardo De Santis
Lucio Barabesi
Gianni Betti

Variance estimation techniques for poverty and inequality measures from complex surveys: a simulation study

n. 829 – Maggio 2020

Abstract

The theme of variance estimation is central in sampling surveys, due to the necessity of furnishing a measure of accuracy for the estimates. In the ambit of social surveys, where we have to face with complex designs and complex statistics, it may be a major issue. To solve this matter, two main approaches can be found in the literature, and both have advantages and disadvantages. However, linearization methods can be safely used in a design-based approach. On the contrary, resampling methods are introduced only in a model-based approach, which means that the properties have to be assessed. Furthermore, some approximations are required. Therefore, we decide to conduce a simulation study by the use of a complete population available. We will focus on some poverty measures considered by the statistical office of the European Union.

1 Introduction

In statistical surveys we always focus on investigating about the value of one - or more - characteristics of a reference population, which represent the target variables. However, it is usually not possible to investigate the whole population, due to temporary and budgetary constraints. Social surveys provide an explanatory example, for instance when we are dealing with a population of a country. Therefore, there is the necessity of analyzing only a part of the population, and consequently of trying to generalize these results. The choice of the sample strategy is crucial, because it influences the estimation procedure, involving several steps. After that the information are collected, we compute the estimates of the target variables. However, it is necessary to calculate also a measure which reflects the accuracy of these estimates, to comprehend their validity. In complex surveys, when we are dealing with complex sample designs and complex statistics, it turns not to be an easy task. Hence, we will focus on this issue. In particular, we will consider the estimation of poverty and inequality measures in population-based surveys of households and persons.

Section 1 contains the introduction, which aims to present the general framework and the notation used. Consequently, we concentrate on the variance estimation techniques. A large amount of literature has been published on this issue, but we may find two main approaches, described in Section 2 and 3, respectively the resampling and the linearization methods. The main reference about the poverty and inequality measures is the European Union - Statistics on Income and Living Conditions (EU-SILC) survey, which involves several countries of the entire European continent. A brief description is contained in Section 4, followed by some inequality indices of interest. Later on, the results of a simulation study are shown in Section 5. Finally, Section 6 contains the conclusion.

To start, let us introduce some basic notions of sampling surveys. Let $\mathcal{U}$ be a fixed population indexed on the first N integers, *i.e.* $\mathcal{U} = \{1, \dots, N\}$, and let Y denote the target variable, while y_i be the value of Y for the i-th individual of the population. Furthermore, let $\theta := \theta(y_1, \dots, y_N)$ be the population parameter. If S is a random sample of fixed size n, θ may be estimated as

$$\widehat{\theta} := \widehat{\theta}(\{y_i : i \in S\}).$$

Therefore, the parameter estimator $\widehat{\theta}$ is computed through the selection of a random sample $S \subset \mathcal{U}$ of size n from the population $\mathcal{U}$. Finally, let π_i be the first-order inclusion probabilities for one unit i, and $\pi_{i,j}$ be the second-order inclusion probabilities for two units i and j, associated with the sample design.

The procedure for the selection of the sample is central, and it is chosen by the researcher by taking care of different motivations. A well-known sample design is the Simple Random Sampling (SRS), which assigns the same first-order inclusion probability to each unit i of the population, namely $\pi_i = n/N$ for each $i \in \mathcal{U}$. However, other sample designs are usually preferred in large sample surveys for many different reasons, mainly the desire to obtain a pre-determined level of accuracy for the estimation, the budget constraints, the availability of auxiliary variables, the difficulties to obtain the sampling frame, and the interest for the parameter estimation for sub-populations. Thompson (2012) and Arnab (2017) provide an overview of the main sample designs. In the next Sections we mainly consider stratification and two-stage designs.

Stratification design consists in partitioning the population $\mathcal{U}$ in L sub-populations, usually named as strata (*e.g.* the regions of a country, in a survey conducted at national level). A sample of size n_l, say S_l, is selected within each stratum l of size N_l, consequently the full sample is $S = \bigcup_{l=1}^{L} S_l$ of size $n = \sum_{l=1}^{L} n_l$. The first-order probabilities are normally unequal between units of different strata, the solely exception being the case of the proportional allocation, when the sampling fraction in each stratum is equal to the sampling fraction of the whole population.

Two-stage design consists in partitioning the population $\mathcal{U}$ in M distinct clusters, called PSUs (Primary Selection Units). At the first stage, $m < M$ PSUs are selected. Let G be the sample of the m PSUs. At the second stage, a sample S_g - of size n_g - is drawn from each PSU, *i.e.* $g \in G$, previously selected. The full sample is $S = \bigcup_{g \in G} S_g$ of size $n = \sum_{g \in G} n_g$, where the first-order probabilities are normally unequal between units located in different PSUs. The procedure can be generalized with more than two stage (multi-stage designs) and by including stratification.

Once the sample has been selected and the variable of interest is collected for each sample unit, the following step focuses on computing the estimate of the parameter of interest, *i.e.* $\widehat{\theta}$, which should be accompanied by a measure of its precision, *e.g.* by an estimate of its variance.

The expected value of $\widehat{\theta}$ is indicated with $E[\widehat{\theta}]$, and - as is well known - the estimator is unbiased if the relation $E[\widehat{\theta}] = \theta$ holds. Otherwise,

$$B[\widehat{\theta}] = E[\widehat{\theta}] - \theta \tag{1}$$

represents the bias of the estimator. In addition, the variance of $\widehat{\theta}$ is defined as

$$Var[\widehat{\theta}] = E\big[(\widehat{\theta} - E[\widehat{\theta}])^2\big]. \tag{2}$$

A common target parameter is the total of a variable T_Y, given by

$$T_Y = \sum_{i \in \mathcal{U}} y_i. \tag{3}$$

In such a case the Horvitz-Thompson (HT) estimator of T_Y (Horvitz and Thompson, 1952) turns out to be

$$\widehat{T}_{Y,HT} = \sum_{i \in S} w_i y_i, \tag{4}$$

where w_i represents the sample weight for unit i, *i.e.* $w_i = \pi_i^{-1}$. The variance of $\widehat{T}_{Y,HT}$ (Arnab, 2017) can be written as

$$Var[\widehat{T}_{Y,HT}] = \sum_{i \in \mathcal{U}} \frac{(1 - \pi_i) y_i^2}{\pi_i} + 2 \sum_{i \geq j \in \mathcal{U}} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j, \tag{5}$$

for $i \neq j$, while an unbiased estimator for the variance (Arnab, 2017) is

$$\widehat{Var}[\widehat{T}_{Y,HT}] = \sum_{i \in S} \frac{(1-\pi_i)y_i^2}{\pi_i^2} + 2 \sum_{i \geq j \in S} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_i \pi_j} \frac{y_i y_j}{\pi_{i,j}}, \tag{6}$$

for $i \neq j$, and $\pi_{i,j} > 0$ for each $(i, j) \in S$.

Actually, variance estimation for a general parameter may often be cumbersome. In complex population-based surveys (*e.g.* EU-SILC) the variance estimation of the estimator of a general parameter θ can be tricky for two different reasons: the use of complex designs, which does not allow to know the second-order probabilities, and θ may be non-linear, as many poverty measures. The procedures to estimate the variance of complex statistics can be subdivided into two main approaches: the methods based on resampling techniques, and the methods based on linearization techniques. Both the methods present advantages and disadvantages: resampling methods may need a massive computational burden, even if the same procedure can be applied for θ of any complexity, and standardized routines for the common statistical software are available or may be easily implemented, without the necessity of computing specific quantities for each statistic, a point which can be helpful for researchers. Linearization techniques need a smaller computational burden, even if they require to compute the linear form for each statistic, which may be a difficult task for researchers, and it may not be unique.

In the case of complex surveys, with large sample size and large population size, it is possible to use the "ultimate cluster approach" (Särndal *et al.*, 1992), which consists in a simplification in computing the variance estimation by taking account solely of the variation among PSU totals. This method requires first-stage sampling fractions to be small, a condition which is usually met in large surveys. It allows easier computation for variance estimation and a great flexibility.

2 Resampling methods

There are many resampling methods presented in the literature (Efron, 1982, Davison and Hinkley, 1997), mainly the Jackknife Repeated Replication (JRR), the Bootstrap and the Grouped Balanced Method. The concept is to estimate the variance through comparisons among replications generated by repeated re-sampling of the same parent sample.

At first, we introduce JRR in its original general version, while afterwards the "Jackknife Delete One PSU" is presented, a procedure which has been selected by Verma and Betti (2011) in many works concerning complex surveys (EU-SILC) for its simplicity, that allows to build standardized routines useful for non-statisticians. It also permits to take account of some aspects that can affect the variance as non-response, calibration, composite estimation, stratification and multi-stage sampling.

2.1 Jackknife repeated replication

Given a sample S of size n, JRR consists in sequentially deleting points y_i, and computing $\widetilde{\theta}_i$, for each $i \in S$, where $\widetilde{\theta}_i$ represents the parameter estimator obtained deleting the i-th observations, namely

$$\widetilde{\theta}_i := \widetilde{\theta}_i(\{y_j : j \in S, j \neq i\}).$$

Consequently, n different estimators are obtained, each one by the use of $(n-1)$ observations. Thus, an estimator for $Var[\widehat{\theta}]$ may be given by

$$\widehat{Var}[\widehat{\theta}] = \frac{n-1}{n} \sum_{i \in S} (\widetilde{\theta}_i - \widetilde{\widetilde{\theta}})^2, \tag{7}$$

where $\widetilde{\widetilde{\theta}} = \sum_{i \in S} \widetilde{\theta}_i / n$.

When n is large, the computational burden to apply the method can be very huge, even with the use of statistical software. The most common choice to avoid this problem is to remove a block of observations at each replication. Obviously, an excessive short number of replications gives unreliable estimates. Some arrangements are required with more complex design than Simple Random Sampling: the procedure described in next Subsection considers stratification and multi-stage sampling and has been developed for complex surveys.

2.2 Jackknife delete one PSU

The application of the "Jackknife Delete One PSU" needs a situation where two or more Primary Selection Units (PSUs) are selected from each stratum of the population independently, at the first stage, while subsampling of any complexity is allowed within each PSU. Each JRR replication consists in deleting one sample PSU from one particular stratum, increasing the weights of the remaining primary units in that stratum appropriately, and computing the parameter estimate. Consequently, there are as many replications as the amount of PSUs which are present in the sample.

Consider a population $\mathcal{U}$ divided in L different strata, where each stratum l contains M_l PSUs: at the first stage, a number $m_l \leq M_l$ of PSUs, with $m_l \geq 2$, is drawn within each stratum. Let G_l be the sample of the m_l PSUs: at the second stage, a sample $S_{gl} \subseteq \mathcal{U}_{gl}$ of size $n_{gl} \leq N_{gl}$ is drawn within each PSU, *i.e.* $g_l \in G_l$, previously selected. Subsampling of any complexity is allowed within each sample PSU, and may differ between different PSUs.

The procedure to compute the modified weights w'_{igl} is the following: let w_{igl} indicate the weight of the i-th unit in g-th PSU and l-th stratum, and let k_{gl} be a quantity required for the computation of the weights, defined as

$$k_{gl} = \frac{w_l}{w_l - w_{gl}}, \tag{8}$$

where

$$w_l = \sum_{g \in G_l} w_{gl} \text{ and } w_{gl} = \sum_{i \in S_{gl}} w_{igl}.$$

Therefore, the following weights are defined for each individual unit i, with reference to the replication (gl), *i.e.* the replication which deletes the observations located in the PSU g_l,

$$w'_{igl} = \begin{cases} w_{igl} & \text{if} \quad i \notin l \\ k_{gl} w_{igl} & \text{if} \quad i \in l, \notin g_l \\ 0 & \text{if} \quad i \in g_l \end{cases}. \tag{9}$$

As before, let $\widehat{\theta}$ be the full-sample estimator of the population parameter θ, let

$$\widetilde{\theta}_{gl} := \widetilde{\theta}_{gl}(\{y_i : i \notin S_{gl}\})$$

be the estimator obtained at the gl-th replication, with the weights defined above, and $\widetilde{\widetilde{\theta}}_l$ be the simple average of the m_l values $\widetilde{\theta}_{gl}$ obtained by deleting each PSU which is located in stratum l. The variance of $\widehat{\theta}$ can be estimated by (Verma and Betti, 2011)

$$\widehat{Var}[\widehat{\theta}] = \sum_{l=1}^{L} \left[\left(1 - \frac{n_l}{N_l}\right) \frac{m_l - 1}{m_l} \sum_{g \in G_l} \left(\widetilde{\theta}_{gl} - \widetilde{\widetilde{\theta}}_l\right)^2 \right], \tag{10}$$

where $(1 - n_l N_l^{-1})$ is the finite population correction.

2.3 Possible variations

JRR may be also applied if a whole group of PSUs is deleted at each replication, and even if only a subset of the PSU groups is used, reducing the computational burden. If we focus on one stratum l, the m_l PSUs could be grouped in b_l clusters, where each PSU belongs only to one cluster, and $c_l < b_l$ clusters are eliminated at each replication. In such a case the factor $(m_l - 1)/m_l$ in the equation (10) is replaced by $(b_l - 1)/c_l$. Besides, it may be also possible to replace the quantities k_{gl} with $k_l = m_l/(m_l - 1)$, consequently the weights would turn out to be the same for each PSU located in the same stratum l. This modification results more appropriate when the PSUs have similar size.

2.4 Further details

JRR can be used for any statistic θ through the equation (10). However, it performs better for certain types. For simpler statistics such as means, ratios and functions of ratios, it has been widely used. About more complex statistics, Efron (1982) shows in a model-based approach that the JRR variance estimator is generally upward biased, and the bias decreases when the sample size and the number of replications increase. He also shows some classes of statistics for which the method can be used safely, which include the Gini Coefficient and other inequality indices. Unfortunately, JRR does not perform well for unsmoothed functions of sample aggregates: Efron (1982) shows the inconsistency for the median. Furthermore, even if it is consistent under a model-based approach, its properties in the design-based approach have to be assessed (Arnab, 2017).

The method presented above requires at least two PSUs for each stratum. It can be possible that someone is dealing with a survey where one - or more - strata have only one PSU. Furthermore, Kott (2001) puts the minimum number of PSUs in each stratum at five to guarantee the near unbiasedness of the variance estimator, but notes that the bias may be acceptable in some situations. Rust and Kalton (1987) suggest methods for grouping or collapsing strata or PSUs, to make possible the application of the "Jackknife Delete One PSU": they show that appropriate collapsing usually does not introduce additional bias or variability in the variance estimation. However, it is an operation which requires particular attention, because reducing the number of PSUs is a potential source of bias and instability for the variance estimator, due to the reduced number of replications. The major motivation for grouping units is to reduce the computational burden required, as a consequence of the reduced number of replications. The approach suggested by Verma and Betti (2011) is to define computational PSUs reasonably uniform and with sufficiently large size: through experiences in EU-SILC and similar applications, Verma *et al.* (2010) find 200 "computational PSUs" to be a safe choice in all cases, and even 100 in almost all cases. Further details of practical utility can be found in Verma *et al.* (2010).

The method may also be implemented with a (Stratified) Simple Random Sampling, through the construction of random "computational PSUs" within each stratum, following the instructions explained in this Section: it allows to greatly reduce the number of replications.

3 Linearization methods

About the linearization methods, the most known method is the Taylor linearization. However, it requires that the statistic is a regular function of estimated totals, continuously differentiable up to order two. As explained by Osier (2009) and Langel and Tillé (2013), for many complex statistics - as many poverty measures - this request is not satisfied, therefore a different way to derive the variance estimator has to be found. The main concept is to obtain a linearized variable z_i for each observation y_i such that

$$\widehat{\theta} - \theta = \sum_{i \in S} w_i z_i - \sum_{i \in \mathcal{U}} z_i + R, \tag{11}$$

where R represents a remainder term which is stochastically negligible respect to the design. The consequence is that the variance of the estimator may be approximated by the variance of the linearized variable. In practice, z_i is unknown and has to be replaced by its sample counterpart $\widehat{z}_i$.

There are several methods about the linearization approach (for a full overview see Langel and Tillé, 2013), even if two are more relevant in the literature: the estimating equation's and the influence function's approach.

3.1 Estimating equation

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^\mathsf{T}$ be nuisance parameters, which are not of immediate interest but are involved in the estimation of θ, we therefore have to take account of them. For finite population, most parameters of interest can be expressed as the solution of the equation

$$U(\theta, \boldsymbol{\lambda}) = 0, \tag{12}$$

called estimating equation, where

$$U(\theta, \boldsymbol{\lambda}) = \sum_{i \in \mathcal{U}} u(y_i, \boldsymbol{\lambda}, \theta) \tag{13}$$

is a suitable function, whose sampling counterpart is given by

$$\widehat{U}(\theta, \boldsymbol{\lambda}) = \sum_{i \in S} w_i u(y_i, \boldsymbol{\lambda}, \theta), \tag{14}$$

where w_i is a weight such that $\sum_{i \in S} w_i = \widehat{N}$ - *i.e.* the HT estimator of N.

By solving the equation $\widehat{U}(\theta, \boldsymbol{\lambda}) = 0$ is possible to obtain an estimator $\widehat{\theta}$ - the so-called "estimating equation" estimator - of the population parameter θ, while analogously an estimator of the nuisance parameters - *i.e.* $\widehat{\boldsymbol{\lambda}}$ - is obtained. For instance, the population mean of y, *i.e.* μ_y, can be estimated by

$$\sum_{i \in S} w_i (y_i - \widehat{\theta}), \tag{15}$$

and the resulting estimator is

$$\widehat{\theta} = \left(\sum_{i \in S} w_i y_i \right) \left(\sum_{i \in S} w_i \right)^{-1}. \tag{16}$$

It is important to point out that the choice of an estimating function for a certain parameter θ may not be unique in general, but a linearized variable is specific to the parameter concerned, irrespective of the particular estimating equation adopted, which is a result of considerable practical utility.

In the general situation where one - or more - nuisance parameters are present, the first step to achieve an expression for the variance estimation of $\widehat{\theta}$ is to subdivide the term $\widehat{U}(\widehat{\theta}, \widehat{\boldsymbol{\lambda}})$ as

$$\widehat{U}(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) = \begin{pmatrix} \widehat{U}_1(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \\ \widehat{U}_2(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \end{pmatrix}, \tag{17}$$

where the first component $\widehat{U}_1(\widehat{\theta}, \widehat{\boldsymbol{\lambda}})$ is a scalar dealing with θ, while $\widehat{U}_2(\widehat{\theta}, \widehat{\boldsymbol{\lambda}})$ is a vector of dimension M dealing with the nuisance parameters $\boldsymbol{\lambda}$, such that the following equality holds,

$$\begin{pmatrix} \widehat{U}_1(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \\ \widehat{U}_2(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \end{pmatrix} = \begin{pmatrix} \sum\limits_{i \in S} w_i u_1(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \\ \sum\limits_{i \in S} w_i u_2(\widehat{\theta}, \widehat{\boldsymbol{\lambda}}) \end{pmatrix} = 0. \tag{18}$$

The following steps, omitted here, consists in a decomposition of the two terms and an approximation through Taylor linearization: the entire procedure and the corresponding proof can be found in Binder and Patak (1994), Binder and Kovacevic (1997).

Through two assumptions made by the authors which are met in most cases of practical importance, namely that $\widehat{U}_2(\widehat{\theta}, \widehat{\boldsymbol{\lambda}})$ does not depend on θ and that the derivative of the estimating functions with respect to θ does not depend on $\boldsymbol{\lambda}$, the following expression is obtained,

$$\widehat{\theta} - \theta = \sum_{i \in \mathcal{S}} w_i z_i - \sum_{i \in \mathcal{U}} z_i + R, \tag{19}$$

where R is a remainder term stochastically negligible (see Binder and Patak, 1994), and z_i is defined as

$$z_i = \Big[-\widehat{U}_1(\theta, \boldsymbol{\lambda}) + J_{1,\boldsymbol{\lambda}}^{\mathsf{T}} J_{2,\boldsymbol{\lambda}}^{-1} \widehat{U}_2(\theta, \boldsymbol{\lambda}) \Big] J_{1,\theta}^{-1}, \tag{20}$$

where

$$J_{1,\theta} = \frac{\partial U_1(\theta, \boldsymbol{\lambda})}{\partial \theta}$$

is a scalar,

$$J_{1,\boldsymbol{\lambda}} = \frac{\partial U_1(\theta, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$$

is a vector of order M, while

$$J_{2,\boldsymbol{\lambda}} = \frac{\partial U_2(\theta, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$$

is a square matrix of order M.

Once the linearized variable z_i is achieved, it can be possible to approximate the variance of $\widehat{\theta}$ by

$$Var[\widehat{\theta}] = Var\Big[\sum_{i \in S} w_i z_i\Big] + R, \tag{21}$$

where R is a remainder term of order n^{-1} (see Binder and Patak, 1994). To approximate the variance of the parameter, it is important to replace θ and $\boldsymbol{\lambda}$ with $\widehat{\theta}$ and $\widehat{\boldsymbol{\lambda}}$ only in the final expression of the variance, obtaining the estimated linearized variable $\widehat{z}_i$. Formulae to compute the variance estimation of linear variables are well-known, taking care of some aspects as the sample design.

For Multistage Stratified designs, an estimator of the variance (Osier *et al.*, 2013) is

$$\widehat{Var}(\widehat{\theta}) = \sum_{l=1}^{L}\left[(1-\frac{n_l}{N_l})\frac{m_l}{m_l-1}\sum_{g\in G_l}\left(\widehat{z}_{gl}-\frac{\widehat{z}_l}{m_l}\right)^2\right], \tag{22}$$

where $\widehat{z}_{igl}$ is the linearized variable for unit i in PSU g and stratum l, w_{igl} is the corresponding sample weight,

$$\widehat{z}_{gl} = \sum_{i\in S_{gl}} w_{igl}\widehat{z}_{igl},$$

$$\widehat{z}_l = \sum_{g\in G_l} \widehat{z}_{gl},$$

and $(1-n_l N_l^{-1})$ is the finite population correction.

As an example, the method is applied to one poverty measure, the Poverty Rate index (PR) - see Subsection 4.1. Following the definition of the European statistical office, it is defined as the proportion of units whose income is below the 60% of the median $(\widetilde{Y})$ of the income distribution, which is formally called Poverty Threshold (PT), *i.e.*

$$PT = 0.6\widetilde{Y} \tag{23}$$

and

$$PR = \frac{1}{N}\sum_{i\in\mathcal{U}}\mathcal{I}_{]-\infty,PT]}(y_i), \tag{24}$$

where $\mathcal{I}_A(y)$ is the indicator function, which takes value 1 if $y\in A$ and 0 otherwise. Therefore, in such a case the variable of interest y_i represents the equivalized income of unit i (see Section 4 for the definition of the equivalized income).

The estimating equations for PR turn out to be

$$\begin{pmatrix} \widehat{U}_1(PR,\widetilde{Y}) = \frac{1}{N}\sum_{i\in S} w_i\Big(\mathcal{I}_{]-\infty,PT]}(y_i) - PR\Big) \\ \widehat{U}_2(PR,\widetilde{Y}) = \frac{1}{N}\sum_{i\in S} w_i\Big(\mathcal{I}_{]-\infty,\widetilde{Y}]}(y_i) - \frac{1}{2}\Big) \end{pmatrix}. \tag{25}$$

By the use of (20), the linearized form of the Poverty Rate is given by

$$z_i = \frac{1}{N}\Big[\mathcal{I}_{]-\infty,PT]}(y_i) - PR\Big] - \frac{1}{N}\left[0.6\frac{f_K(PT)}{f_K(\widetilde{Y})}\Big(\mathcal{I}_{]-\infty,\widetilde{Y}]}(y_i)-\frac{1}{2}\Big)\right], \tag{26}$$

where $J_{1,PR}=1$, $J_{1,\widetilde{Y}} = 0.6 f_K(PT)$, $J_{2,\widetilde{Y}} = f_K(\widetilde{Y})$.

Finally, the estimated linearized variable $\widehat{z}_i$ is computed by replacing the population parameters with their sample counterparts. It should be remarked that $J_{1,\widetilde{Y}}$ and $J_{2,\widetilde{Y}}$ are representations of sample quantiles given by Francisco and Fuller (1991), which are valid under some assumptions that do not hold for their sample estimation from finite population (see Subsection 2.3). Consequently, the variance estimator is not robust for statistics which are function of sample quantiles.

3.2 Influence function

The second method presented is based on the concept of influence function, which was first introduced in robust statistics by Hampel (1974). The purpose was to grasp the effect of an infinitesimal contamination on a parameter of interest.

The first step is to express the population parameter θ as a functional $\theta = T(M)$, where M is a measure which allocates a mass of 1, $M(k) = 1$, only at point k, such that its total mass is equal to N, namely the size of the population. The specialization of M into a discrete measure turns T into a discrete functional. The influence function of T is defined as

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \tag{27}$$

for all $k \in \mathcal{U}$, where δ_k is the Dirac measure for the unit k ($\delta_k = 1$ if and only if $i = k$), and z_k is the linearized variable. This influence function is the Gâteaux-differential in the direction of the Dirac mass at point k.

The measure M is estimated by the empirical measure $\widehat{M}_k = w_k$, for each unit $k \in S$, where w_k is a weight (*e.g.* sample weight), thus an estimator of θ is $\widehat{\theta} = T(\widehat{M})$. Deville (1999) justifies this procedure showing that

$$T(\widehat{M}) - T(M) = \sum_{k \in S} w_k z_k - \sum_{k \in \mathcal{U}} z_k + R, \tag{28}$$

where again R is a remainder term stochastically negligible. Under some asymptotic conditions expressed by Deville (1999), which are theoretically satisfied for large samples, the variance of the linearized variable z_k is an approximation of the variance of $\widehat{\theta}$, *i.e.*

$$Var[\widehat{\theta}] = Var\Big[\sum_{i \in S} w_i z_i\Big] + R, \tag{29}$$

where R is a remainder of order n^{-1} (see Deville, 1999). In practice, only the sample data are available, thus an estimated linearized variable $\widehat{z}_k$ is obtained, replacing the unknown values with the corresponding quantities estimated from the sample.

Actually this approach starts from the population parameter and not from the estimator. Demnati and Rao (2004) proposed to use not the discrete measure defined in $\mathcal{U}$, but directly the following measure defined for S, *i.e.* $\widehat{M}(k) = w_k$, for $k \in S$. The consequence is that now the starting point is not the parameter, but the estimator, and the linearized variable based on that functional is

$$I[T(\widehat{M})]_k = \widehat{z}_k = \lim_{t \to 0} \frac{T(\widehat{M} + t\delta_k) - T(\widehat{M})}{t}, \tag{30}$$

for all $k \in S$. The result obtained turns out to be exactly the same of the result given by Deville's approach.

As explained in the Subsection above, the variance of the linearized variable can be calculated by standard methods and well-known formulae. Osier (2009) shows the procedures to compute the influence functions, which are similar to the derivative rules, without the necessity of computing limits which may result to be a difficult task, and he also shows examples for some poverty measures. Langel and Tillé (2013) show an additional result to compute the linearized form of a double sum (*e.g.* quadratic form) directly, while Barabesi *et al.* (2016) derive a rule for computing the influence function of a general family of complex population functionals, which includes many poverty measures, and provide examples for some inequality indices of interest.

As an example, the steps to obtain the influence function and the linearized form of one poverty measure, the Poverty Rate index (PR) defined in the Subsection above, are shown.

At first, the index is expressed as a functional of M, *i.e.*

$$PR = F[M, PT(M)] = T(M), \tag{31}$$

where F is the cumulative income distribution function,

$$F(M, y) = \frac{1}{N} \sum_{i \in \mathcal{U}} \mathcal{I}_{]-\infty, y]}(y_i). \tag{32}$$

The influence function of the Poverty Rate can be written as a sum of two terms: the first one is the influence function of T with respect to M, holding the parameter $PT(M)$ constant (c), while the second one accounts for the influence function of $PT(M)$, *i.e.*

$$IPR_k(M) = z_k = z_k^0 + z_k^v, \tag{33}$$

where

$$\begin{cases} z_k^0 = & IF_k[M, PT(M)|PT(M) = c] \\ z_k^v = & \left[\dfrac{dF(M,y)}{dy}|y = PT(M)\right] IPT(M) \end{cases}. \tag{34}$$

We firstly introduce some results needed: the influence function of the ratio

$$R = \frac{\theta_1}{\theta_2} = \frac{U(M)}{V(M)}$$

is defined as

$$IR_k(M) = \frac{V(M)IU_k(M) - U(M)IV_k(M)}{V(M)^2}, \tag{35}$$

while the influence function of the population total of a variable

$$T_Y = \sum_{k \in \mathcal{U}} y_k = T_Y(M)$$

is defined as

$$IT_{Y,k}(M) = y_k. \tag{36}$$

Consequently, the influence function of F, the cumulative income distribution function (32), is

$$IF_k(M) = \frac{1}{N} \Big[\mathcal{I}_{]-\infty, y]}(y_k) - F(y)\Big]. \tag{37}$$

Given (37), the influence function of F with respect to M, holding $PT(M)$ constant, is

$$z_k^0 = \frac{1}{N}\Big[\mathcal{I}_{]-\infty,PT(M)]}(y_k) - PR(M)\Big]. \tag{38}$$

The next step is the computation of $IPT_k(M)$: by definition, the median income $\widetilde{Y}(M)$ satisfies the identity

$$F[M, \widetilde{Y}(M)] = 1/2,$$

thus its influence function is

$$IF_k[M, \widetilde{Y}(M)] = 0. \tag{39}$$

Through the same rules used to get (34), the functional can be rewritten as

$$IF_k[M, \widetilde{Y}(M)|\widetilde{Y}(M) = c] + \Big[\frac{dF(M,y)}{dy}|y = \widetilde{Y}(M)\Big] I\widetilde{Y}_k(M) = 0. \tag{40}$$

The influence function of F has already been defined in (37): thus, holding $\widetilde{Y}(M)$ constant, we obtain

$$IF_k[M, \widetilde{Y}(M)|\widetilde{Y}(M) = c] = \frac{1}{N}\Big[\mathcal{I}_{]-\infty,\widetilde{Y}(M)]}(y_k) - \frac{1}{2}\Big]. \tag{41}$$

Besides, let $f = F'$ denote the derivative of the cumulative distribution function. As can be seen below, it is necessary that f exists, and it must be strictly non negative for each y. Unfortunately, for finite populations - and therefore also for samples - the cumulative distribution function is a step function, it means that its derivative is always 0 or not defined. Methods to solve this problem are shown in Subsection 2.3, and the choice is not unique: let f_K be the differentiable function that has been chosen (note again that in neither case it is a consistent estimator in a design-based approach). Thus, we can rewrite (40) as

$$\frac{1}{N}\Big[\mathcal{I}_{]-\infty,\widetilde{Y}(M)]}(y_k) - \frac{1}{2}\Big] + f_K(\widetilde{Y}(M)) I\widetilde{Y}_k(M) = 0. \tag{42}$$

Now, after having obtained $I\widetilde{Y}_k(M)$, the influence function of the Poverty Threshold ($IPT_k(M)$) can be easily written through the relation (23), which gives

$$IPT_k(M) = 0.6 I\widetilde{Y}_k(M) = -\frac{0.6}{f_K(\widetilde{Y}(M))}\frac{1}{N}\Big[\mathcal{I}_{]-\infty,\widetilde{Y}(M)]}(y_k) - \frac{1}{2}\Big]. \tag{43}$$

Going back to the equation (34), it is possible to substitute $IPT_k(M)$ to obtain the influence function of the Poverty Rate, namely

$$\begin{cases} z_k^0 = & \frac{1}{N}\Big[\mathcal{I}_{]-\infty,PT(M)]}(y_k) - PR(M)\Big] \\ z_k^v = & f_K(PT(M))\Big[-\frac{0.6}{f_K(\widetilde{Y}(M))}\frac{1}{N}\Big[\mathcal{I}_{]-\infty,\widetilde{Y}(M)]}(y_k) - \frac{1}{2}\Big]\Big] \end{cases}, \tag{44}$$

which can be rewritten as

$$\begin{cases} z_k^0 = & \frac{1}{N}\Big[\mathcal{I}_{]-\infty,PT(M)]}(y_k) - PR(M)\Big] \\ z_k^v = & -\frac{0.6}{N}\frac{f_K(PT(M))}{f_K(\widetilde{Y}(M))}\Big[\mathcal{I}_{]-\infty,\widetilde{Y}(M)]}(y_k) - \frac{1}{2}\Big] \end{cases}. \tag{45}$$

Finally,

$$IPR_k(M) = z_k = z_k^0 + z_k^v$$

is the influence function of the Poverty Rate, namely its linearized form. To obtain the estimated linearized variable, *i.e.* $\widehat{z}_k$, it is necessary to substitute the population values with their corresponding sample estimators.

3.3 Estimate the density function

When the linearized form of any poverty measure is computed, references to the density function at various points of the income distribution are usually involved (*e.g.* median). The cumulative distribution function is a step function, it implies that the density function is 0 nearly everywhere: as explained by Verma and Betti (2011), to solve this problem there are many density estimation techniques which provide a differentiable (artificially) smoothed function. The most common choice is to use the Kernel estimator, where there is the need to choice the kernel function K, which is a probability density function, and the bandwidth parameter h, which controls the degree of smoothing applied to the data in such a way that the density function is more smoothed when h increases and vice-versa. The Kernel method has been originally introduced in a model-based approach (Silverman, 1986), later an extension to finite populations has been developed (Jones and Bradbury, 1993): the density function estimated from the data is defined as

$$\widehat{f}_K(y) = \frac{1}{\widehat{N}} \frac{1}{h} \sum_{i \in S} w_i K\left(\frac{y - y_i}{h}\right). \tag{46}$$

Silverman (1986) shows that the efficiency between the most common functions is similar, while the choice of the bandwidth parameter is crucial. One of the most widely used method to evaluate the global accuracy is the Mean Integrated Square Error (MISE), and minimizing an appropriate estimator of MISE is a way to estimate the bandwidth parameter. In a model-based approach MISE is defined as

$$\text{MISE}(\widehat{f}_K(y)) = E \int [\widehat{f}_K(y) - f(y)]^2 dy, \tag{47}$$

while Jones and Bradbury (1993) develop a generalization of MISE for samples from finite populations.

However, there are many ways to choice the parameter h, which may also depend on the characteristics of the distribution (*e.g.* its skewness). For instance, when $f(y)$ is suspected to come from a Log-normal or heavily skewed distribution, an optimal value for the bandwidth parameter h suggested by Silverman (1986) is $h_o = 0.79Rn^{-1/5}$, where R is the inter-quantile range. Graf and Tillé (2014) point out that the most common choices of the bandwidth parameter, based on rules given by Silverman (1986) (Osier, 2009, Verma and Betti, 2011), are not robust estimators because of outliers and irregularities of the empirical density function, and the consequence is that a strong bias in the variance estimation may occur. Consequently, they propose a method to obtain more robust estimation for density function.

4 Laeken indicators

Poverty and inequality are studied in the European Union (EU) by the realization of the European Union - Statistics on Income and Living Conditions (EU-SILC) survey, which is conducted by the European statistical office (Eurostat). A set of indicators (known as Laeken indicators - European Commission, 2003) is estimated and published each year, through the data obtained with the survey. The regulation is defined uniquely by Eurostat, expression of EU, and allows the standardization of the procedure and the comparability between countries, reflecting a balanced representation of EU social concerns. However, there are great differences between countries, especially for the choice of the sample design (presence/absence of stratification, multistage and systematic sampling)[1]. The survey involves not only the members of EU, but also some candidate countries and potential candidates of EU - the complete list can be found on the Eurostat website[2].

We decide to focus on some well-known monetary measures of the Laeken indicators. In this framework the variable of interest - y_i - represents the equivalized income for unit i (European Commission, 2019).

4.1 Poverty Rate

The Poverty Rate (PR) index, also known as Head Count Ratio, is defined as the proportion of units whose income is below the Poverty Threshold (PT), which is not defined uniquely. This choice is crucial and requires a certain level of subjectivity. Eurostat adopts the 60% of the median of the income distribution. It is estimated by

$$\widehat{PR} = \frac{\sum_{i\in\mathcal{S}} w_i \mathcal{I}_{]-\infty,PT]}(y_i)}{\sum_{i\in S} w_i}. \tag{48}$$

4.2 Quintile Share Ratio

The index Quintile Share Ratio (Qsr) - also known as $S80/S20$ - is defined as the proportion between the total income received by the richest 20% of the population, and the total income received by the poorest 20% of the population. It is estimated by

$$S\widehat{80/S}20 = \frac{\sum_{i\in S} w_i y_i I_{]\widehat{Y}_{0.8},+\infty]}(y_i)}{\sum_{i\in S} w_i y_i I_{[-\infty,\widehat{Y}_{0.2}]}(y_i)}, \tag{49}$$

where $\widehat{Y}_{0.8}$ and $\widehat{Y}_{0.2}$ are the quantiles estimated from the sample.

[1] https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology_-_sampling

[2] https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

4.3 Gini Index

The Gini Index is a popular inequality index proposed by the statistician Corrado Gini (1912, 1921), which has been widely studied in the literature. Langel and Tillé (2013), and Barabesi *et al.* (2016) offer a global overview of the studies about making inference on this index.

The simplest way to describe the Gini Index ($Gini$) is by the use of the Lorenz curve (1905). This curve is obtained by sorting the units on the basis of their income on the x axis, from the poorest to the richest, and the corresponding cumulative income distribution function is represented on the y axis. It is important to know that the index may also be defined in a different way, independently from the Lorenz curve (Langel and Tillé, 2013).

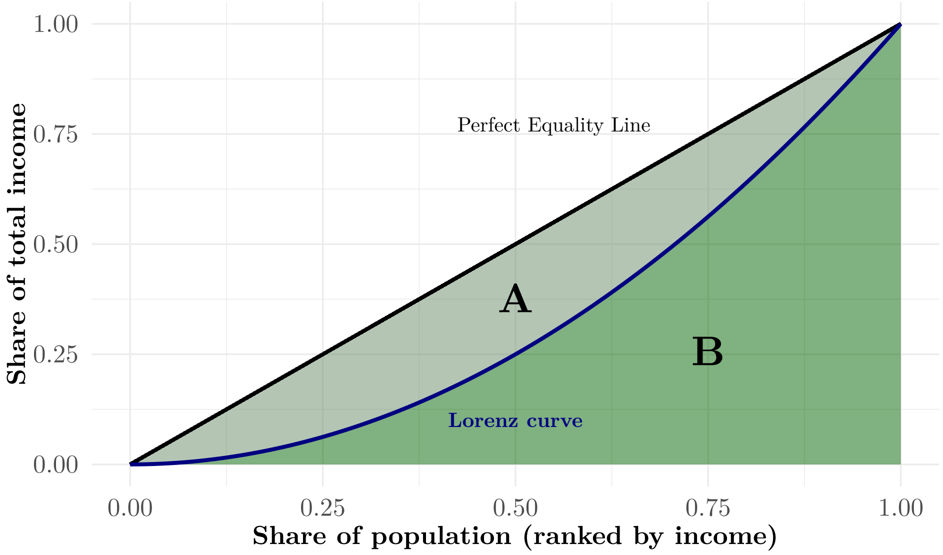

Figure 1: Gini Index and Lorenz curve

Hence, the Gini Coefficient is defined as

$$Gini = \frac{A}{A+B}.$$

It is clear that $0 \leq Gini < 1$, where a lower value implies a more equal income distribution. According to European Commission (2003) the Gini Index may be estimated by

$$\widehat{Gini} = \frac{2\sum_{i\in S}\left(w_i y_i \sum_{j\in S: y_j \leq y_i} w_j\right) - \sum_{i\in S} w_i^2 y_i}{\left(\sum_{i\in S} w_i\right)\sum_{i\in S} w_i y_i} - 1.$$

5 Empirical analysis - Monte Carlo simulation

A simulation study is required when we would like to comprehend the empirical properties of an estimation method, in the situation where they could not be formally proved. The idea is to draw a large number of sample from a complete population available, and thus to estimate the parameter of interest each time, to obtain as many estimates as the number of sample drawn, which is an arbitrary number as it is our choice. Obviously, a greater number of replications gives more reliable results. The following step involves the comparison between the expected value of the estimates, and their distribution, with the true value of the parameter of interest which is known.

Therefore, we decide to conduce a simulation study to understand how the JRR defined in Subsection 2.2 works. Furthermore, we compare the results with the linearization method - using the equation (22) - and the naive Bootstrap described in Alfons and Templ (2013), by taking account of the presence of stratification and clustering in the sample design.

5.1 Population and sampling design

The data used in this simulation study are obtained from the 2011 census of Albania, which contains a limited amount of information for the whole population, in combination with the Albanian Living Standard Measurement Survey (LSMS) of 2012, a multi-purpose survey which collects information to measure poverty and living conditions. The two sources of data have been used to simulate the consumption of each household by the use of a methodology named Poverty Mapping (Elbers *et al.*, 2003, Betti *et al.*, 2018). In few words, it combines census and sample data, the former having a huge size and the latter being more detailed, and also some extra data - *e.g.* geographical data - can be used if available, in order to describe the spatial poverty distribution on a country, through the construction of a database with a high level of disaggregation. It is important to get a large number of variables which can be matched between the two sources. The method follows a model-based approach and it is implemented by the use of econometric techniques.

Through the use of Poverty Mapping 100 simulations of the population consumption distribution have been obtained, thus the expected value for each household has been taken to obtain the per-capita consumption for each population unit - our variable of interest. We use the consumption as a proxy variable for the income, to compute the inequality measures and their errors. The monetary variable is expressed in terms of Albanian leks in the 2002 value of the currency. See Betti *et al.* (2018) for a description of the study made in Albania about poverty and inequality.

The population consists of 722,262 households for a total of 2,784,539 individuals, which are subdivided in 24 strata. The strata are obtained by joining the 12 prefectures of Albania with the dummy variable Urban, which indicates whether the household lives in an Urban or a Rural context. Moreover, the households are grouped in 11,579 PSUs, which have been defined by the Albanian

Institute of Statistics on the basis of the geographical and political subdivisions.

We have drawn a total of 1,000 samples from the Albanian population. Each sample drawn follows the instructions of the survey LSMS 2012, implemented by the Albanian Institute of Statistics. A two-stage design has been adopted: at the first stage, 834 PSUs have been drawn with a systematic stratified sampling, where the sample size for each stratum has been decided by the Albanian Institute to represent the whole country, while within each stratum the inclusion probability of each PSU is proportional to its number of households contained. Afterwards, at the second stage 8 households are selected within each PSU previously selected, with Simple Random Sampling without replacement. Finally, the sample of 6,672 households is obtained. All the individuals of the households selected are included.

5.2 Numeric results

Four inequality indices are considered. ($PR\text{-}fix$) represents the Poverty Rate with a fixed poverty line, (PR-60) represents the Poverty Rate adopted by Eurostat, while (Qsr) stands for the Quintile share ratio. Finally, ($Gini$) represents the Gini Coefficient. All measures are considered at individual level. Note that the Poverty Rates and the Gini Index are represented with the percentage values.

At first, we present the population parameter θ and the expected value ($E[\widehat{\theta}]$) of the 1,000 estimates obtained. Besides, concerning a measure of accuracy, we adopt the square root of the variance - known as Standard Error ($Se[\theta]$) - which has the advantage of having the same unit of measure of the point estimator, in such a way that it may give clearer results. In order to comprehend the performance of the standard error estimator ($\widehat{Se}[\widehat{\theta}]$), we take account of the expected value ($E[\widehat{Se}[\widehat{\theta}]]$) and the relative bias ($RB[\widehat{Se}[\widehat{\theta}]]$), say

$$RB[\widehat{Se}[\widehat{\theta}]] = \frac{E[\widehat{Se}[\widehat{\theta}]] - Se[\widehat{\theta}]}{Se[\widehat{\theta}]}. \tag{50}$$

Furthermore, we also report the distribution of the standard error estimator over the 1,000 samples, which evidences how the estimator diverges around its expected value. We use an estimated kernel distribution for better visualization.

Table 1: Simulation results for the first population

				Jackknife		Bootstrap		Linearization	
	θ	$E[\widehat{\theta}]$	$Se[\widehat{\theta}]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$
$PR\text{-}fix$	14.300	14.320	0.704	0.740	0.051	0.729	0.036	0.736	0.045
PR-60	7.113	7.093	0.496	0.687	0.385	0.536	0.081	0.639	0.288
Qsr	3.076	3.016	0.068	0.088	0.294	0.069	0.015	0.071	0.044
$Gini$	22.500	22.503	0.523	0.524	0.002	0.506	-0.033	0.522	-0.002

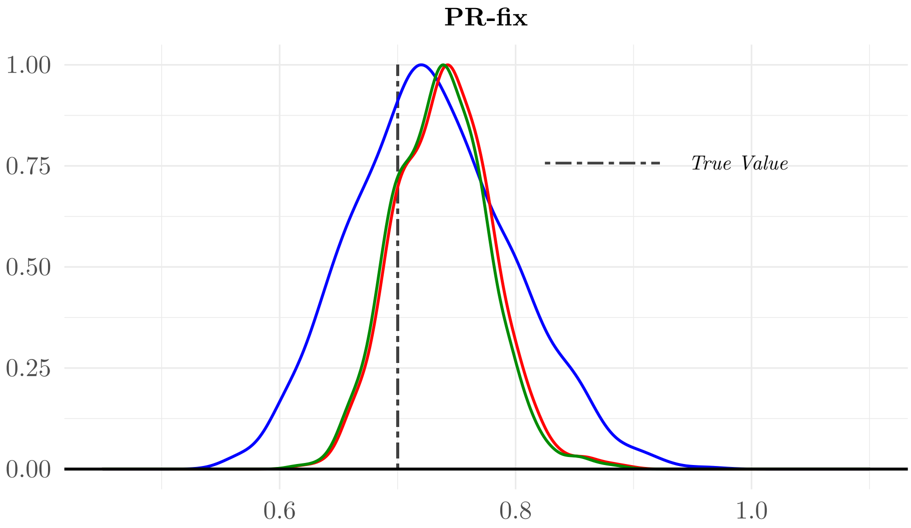

Figure 2: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for $PR\text{-}fix$, both populations

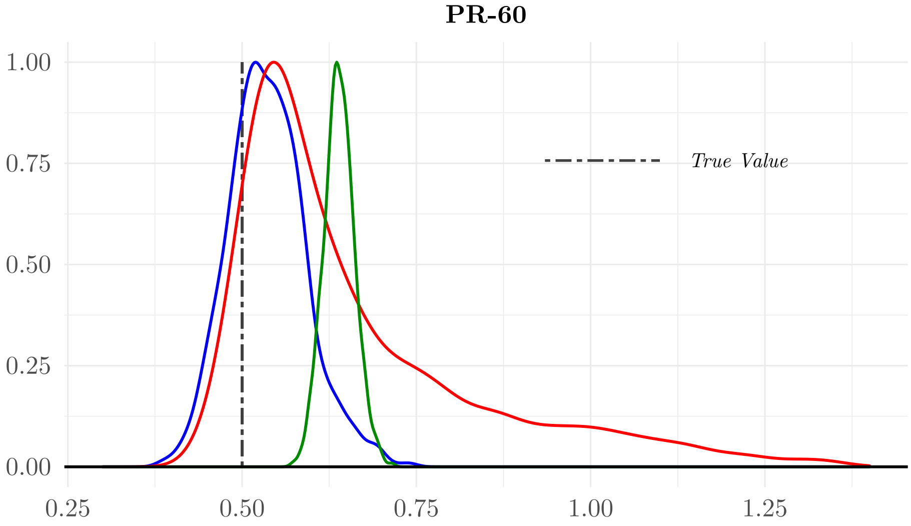

Figure 3: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for PR-60, first population

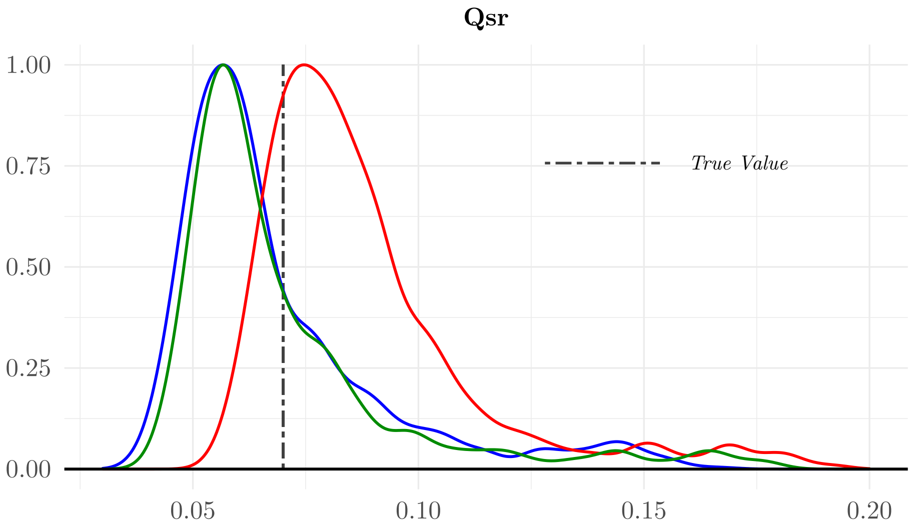

Figure 4: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for *Qsr*, first population

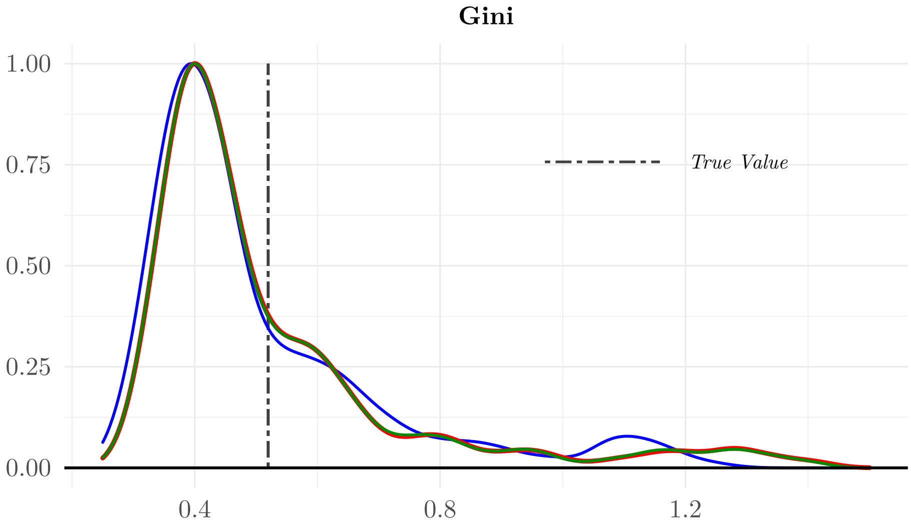

Figure 5: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for *Gini*, first population

We see that the estimators of the inequality indices are nearly unbiased. The small differences are due to the non-proportional sample size between strata.

In terms of accuracy, we do not have univocal results. First of all, we see some asymmetric distributions which suggest to take care not only of the expected value, but mainly of their dispersion. For $PR\text{-}fix$ we have a non-skewed distribution, the Bootstrap is nearly unbiased but more unstable. On the contrary, for PR-60 the Bootstrap has better results, while Jackknife has a problematic long right tail. For *Qsr* we have a long right tail for all the measures, Jackknife seems preferable because the other two methods usually underestimated the true error. Finally, the *Gini* has nearly identical distributions for all the methods, with a long right tail and a usual underestimation.

Again, we decide to apply the simulation study also to a different consumption distribution. This is because we have taken the expected value of the consumption as our proxy variable, which may have artificially reduced the tails of the distribution. Therefore, we adopt a Log-normal model (Aitchinson and Brown, 1969). Firstly, the parameters are estimated on each simulation of consumption. Secondly, the consumption for the 722,262 households is generated from the model, whose parameters are equal to the expected value of their estimates over the 100 simulated distribution.

Table 2: Simulation results for the second population

				Jackknife		Bootstrap		Linearization	
	θ	$E[\widehat{\theta}]$	$Se[\widehat{\theta}]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$	$E[\widehat{Se}[\widehat{\theta}]]$	$RB[\widehat{Se}[\widehat{\theta}]]$
$PR\text{-}fix$	14.300	14.320	0.704	0.740	0.051	0.729	0.036	0.736	0.045
PR-60	16.454	16.444	0.620	0.865	0.395	0.655	0.056	0.642	0.035
Qsr	4.552	4.487	0.104	0.145	0.394	0.117	0.125	0.119	0.144
$Gini$	29.598	29.605	0.479	0.537	0.121	0.524	0.094	0.536	0.119

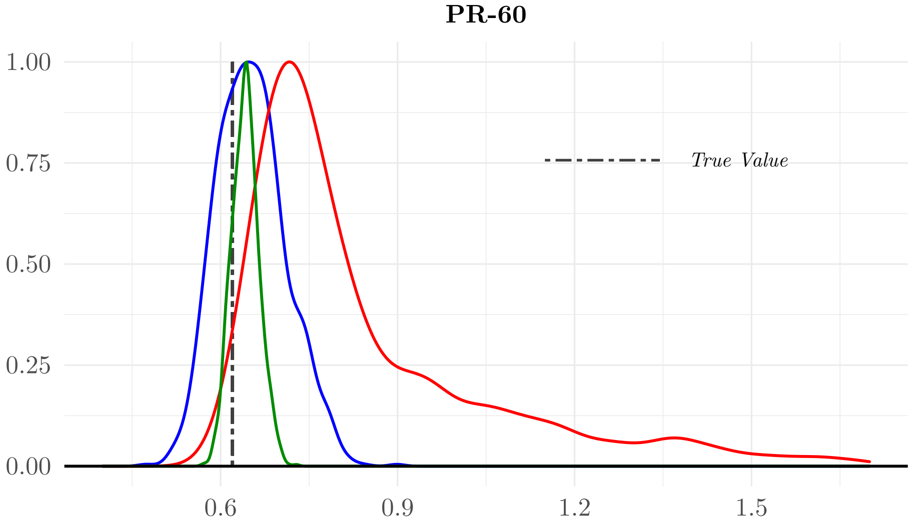

Figure 6: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for PR-60, second population

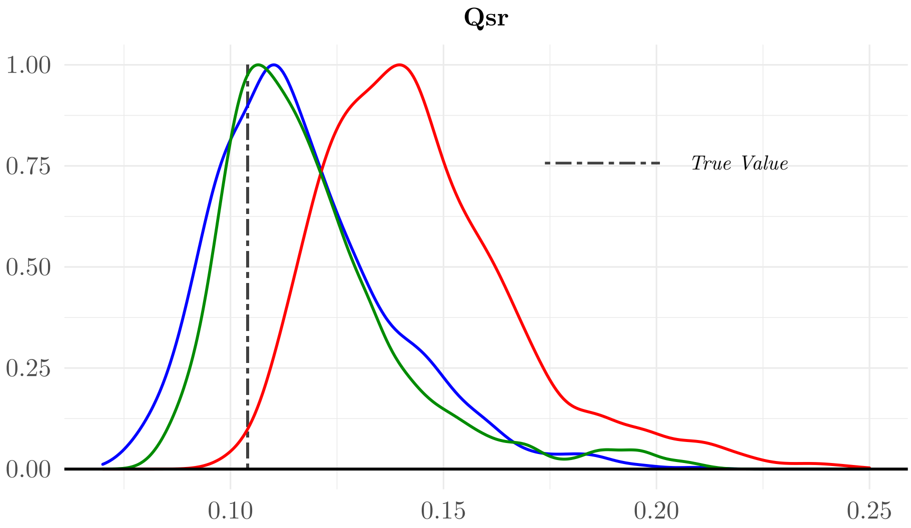

Figure 7: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for Qsr, second population

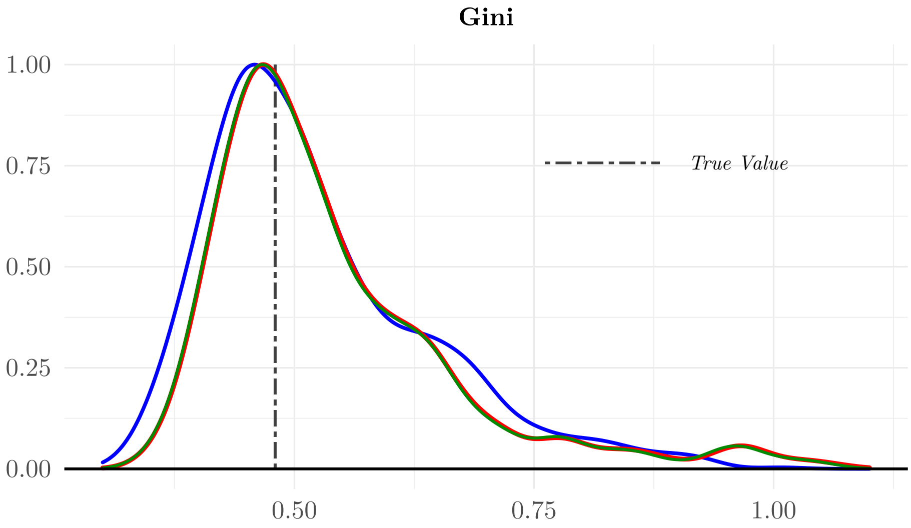

Figure 8: Jackknife (Red), Bootstrap (Blue), and Linearization (Green) Standard Error distribution for $Gini$, second population

Due to the decision of having uniform results with Betti *et al.* (2018), the fixed poverty line has been moved from 5,773.69 to 4,418.82 leks. However, it is useful to compare the values with the second index.

The results for the estimators of the inequality indices are similar to the first population. In terms of accuracy, for PR-60 Linearization is now nearly unbiased, with an high concentration, while again Jackknife has a problematic right tail. For Qsr we face distributions similar to the first population, but now Linearization and Bootstrap are concentrated around the true value. Finally, $Gini$ has the same curves of the first population, with a heavy right tail, but now most of the observations are concentrated on the true value.

Some conclusions seem to be interesting. First of all, it may be noted that in some situations there are quite identical distributions, not always for the same methods. Secondly, we see that the distributions for each parameter are equal between the two populations, moving the curves horizontally. As we have already said, the second population is less concentrated. We see that in this case we have more accurate estimated errors rather than the first population.

Finally, we can conclude that there is not an univocal result. Jackknife seems to be more unstable, and usually gives conservative estimates. Bootstrap and Linearization give nearly always similar results, but in the first population they sometimes tend to underestimate the standard error.

6 Conclusion

In the ambit of sampling from finite population, we see that the theme of variance estimation can be faced safely with different approaches, for many measures, even if we meet some problematic. In presence of complex surveys, and complex measures, some approximations are required. The purpose is to get un unbiased variance estimator, if it exists. Otherwise, we look for getting an estimator which is not downward biased, whose bias decreases as the sample size increases. Therefore, after having shown the main methods, we decide to focus on a common resampling method - Jackknife Repeated Replication - to try to understand how it works, and we have implemented a comparison with Linearization and Bootstrap.

The results say that we do not have a method which has always a major reliability. We see different distributions for each statistic, and also a different bias between the two populations. The Jackknife seems to be more conservative and sometimes more unstable, while - in the first population - the Bootstrap and the Linearization gives sometimes a systematic underestimation. Finally, we can conclude that there is not a clear superiority of any approach over the others, and the preference for one method may be influenced also by practical considerations.

References

[1] Aitchinson, J. and Brown, J.A.C. *The Lognormal distribution with special reference to its uses in economics.* Cambridge University Press, 1969.

[2] Alfons, A. and Templ, M. *Estimation of social exclusion indicators from complex surveys: the R package Laeken.* Journal of Statistical Software, Vol. 54, No. 15, pp. 1-25, 2013.

[3] Arnab, R. *Survey sampling theory and applications.* Academic Press, 2017.

[4] Barabesi, L., Diana G. and Perri P. F. *Linearization of inequality indices in the design-based framework.* Statistics, Vol. 50, No. 5, pp. 1161-1172, 2016.

[5] Betti, G., Bici, R., Neri, L., Sohnesen, T.P. and Thomo, L. *Local poverty and inequality in Albania.* Eastern European Economics, Vol. 56, pp. 223-245, 2018.

[6] Binder, D. A. and Kovacevic, M. S. *Estimating some measures of income inequality from survey data: an application of the estimating equation approach.* Survey Methodology, Vol. 21, pp. 137-145, 1995.

[7] Binder, D. A. and Kovacevic, M. S. *Variance estimation for measures of income inequality and polarization - The estimating equations approach.* Journal of Official Statistics, Vol. 13, No. 1, pp. 41-58, 1997.

[8] Binder, D. A. and Patak, Z. *Use of estimating functions for interval estimation from complex surveys.* Journal of the American Statistical Association, Vol. 89, pp. 1035-1043, 1994.

[9] Davison, A.C. and Hinkley, D.V. *Bootstrap methods and their application.* Cambridge University Press, 1997.

[10] Demnati, A. and Rao, J. N. K. *Linearization variance estimators for survey data.* Survey Methodology, Vol. 30, pp. 17-26, 2004.

[11] Deville, J. C. *Variance estimation for complex statistics and estimators: Linearization and residual techniques.* Survey Methodology, Vol. 25, No. 2, pp. 193-203, 1999.

[12] Efron, B. *The Jackknife, the Bootstrap and other resampling plans.* Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[13] Elbers, C., Lanjouw J. O. and Lanjouw, P. *Micro-level estimation of poverty and inequality.* Econometrica, Vol. 71, No. 1, pp. 355-364. 2003.

[14] European Commission, Eurostat *Laeken indicators. Detailed calculation methodology.* DOC. E2/IPSE/2003, 2003.

[15] European Commission, Eurostat *Methodological guidelines and description of EU-SILC target variables.* Doc/SILC/065 (2018 operation), 2019.

[16] Francisco, C. A. and Fuller, W. A. *Quantile estimation with a complex survey design.* The Annals of Statistics, Vol. 19, No. 1, pp. 454-469, 1991.

[17] Gini, C. *Variabilità e mutabilità.* Bologna, Tipografia di Paolo Cuppini, 1912.

[18] Gini, C., *Measurement of inequality and incomes.* The Economic Journal, Vol. 31, pp. 124-126, 1921.

[19] Graf, E. and Tillé, Y. *Variance estimation using linearization for poverty and social exclusion indicators.* Survey Methodology, Vol. 40, No. 1, pp. 61-79, 2014.

[20] Hampel, F. R. *The influence curve and its role in robust estimation.* Journal of the American Statistical Association, Vol. 69, pp. 383-393, 1974.

[21] Hansen, M., Hurwitz, W. and Madow, W. *Sample survey methods and theory.* Vol. I. New York, Wiley, 1953.

[22] Horvitz, D. G. and Thompson, D. J. *A generalization of sampling without replacement from a finite universe.* Journal of the American Statistical Association, 47, pp. 663-685, 1952.

[23] Jones, M. C. and Bradbury, I. S. *Kernel smoothing for finite populations.* Statistics and Computing, Vol. 3, No. 1, pp. 45-50, 1993.

[24] Kott, P. S. *The Delete-a-Group Jackknife.* Journal of Official Statistics, Vol. 17, No. 4, pp. 521-526, 2001.

[25] Kovacevic, M. S. and Yung, W. *Variance estimation for measures of income inequality and polarization - An empirical study.* Survey Methodology, Vol. 23, No. 1, pp. 41-52, 1997.

[26] Langel, M. and Tillé, Y. *Variance estimation of the Gini index: revisiting results several times published.* Journal of the Royal Statistical Society, Series A 176, pp. 521-540, 2013.

[27] Lorenz, M. O. *Methods of measuring the concentration of wealth.* Publications of the American Statistical Association, Vo. 9, pp. 209-219, 1905.

[28] Munnich, R. and Zins, S. *Variance estimation for indicators of poverty and social exclusion.* Research Project Report WP3 - D3.2, FP7-SSH-2007-217322, AMELI., University of Trier, 2011.

[29] Osier, G. *Variance estimation for complex indicators of poverty and inequality using linearization techniques.* Survey Research Methods, Vol. 3, No. 3, pp. 167-195, 2009.

[30] Osier, G., Berger, Y. G. and Goedemé, T. *Standard error estimation for the EU-SILC indicators of poverty and social exclusion.* Eurostat Methodologies and Working Papers, Eurostat, Luxembourg, 2013.

[31] Rust, K. and Kalton, G. *Strategies for collapsing strata for variance estimation.* Journal of Official Statistics, Vol. 3, No. 1, pp. 69-81, 1987.

[32] Qualité, L. and Tillé, Y. *Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added.* Survey Methodology, Vol. 34, No. 2, pp. 173-181, 2008.

[33] Särndal, C. E., Swensson, B. and Wretman, J. *Model assisted survey sampling.* Springer Series in Statistics, New York, 1992.

[34] Silverman B. W. *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1986.

[35] Thompson, M. E. *Theory of sample surveys.* Chapman & Hall, London, 1997.

[36] Thompson, S. K. *Sampling.* John Wiley & Sons, Hoboken, New Jersey, 2012.

[37] Verma, V. and Betti, G. *Taylor linearization sampling errors and design effects for poverty measures and other complex statistics.* Journal of Applied Statistics, Vol. 38, pp. 1549-1576, 2011.

[38] Verma, V., Betti, G. and Gagliardi, F. *An assessment of survey errors in EU-SILC.* Eurostat Methodologies and Working Papers, Eurostat, Luxembourg, 2010.

[39] Verma, V., Betti, G. and Ghellini, G. *Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC.* Statistics in Transition, Vol. 8, No. 1, pp. 5-50, 2007.