Mind the Income Gap: the Estimation of Inequality Indexes in Small Samples

Silvia De Nicolò, Maria Rosaria Ferrante, and Silvia Pacei

Abstract Inequality measures are biased in small samples leading to a dramatic underestimation. The nature of the bias is investigated and a bias-correction framework is proposed for a large class of inequality measures comprising Gini Index, Generalized Entropy and Atkinson families, accounting for complex survey designs. The methodology proposed is based on Taylor’s expansions and generalized linearization method, and does not require any parametric assumption on income distribution, being very flexible. Design-based performance evaluation of the suggested correction has been carried out, showing a noticeable bias reduction for all measures.

Keywords Bias Correction · Complex Surveys · Income Inequality · Small Sample Inference

JEL Classification C15 · D31

1 Introduction

The estimation of Income Inequality Measures (IM) in small samples can face several issues, a main one is the negative bias in small samples which leads to a dramatic underestimation of inequality, reaching in some cases the 10-15%. This bias depends on the shape and skewness of the underlying distribution of the variable of interest, such as income, known to be affected by positive skewness, and its intensity varies depending on the measure. Generally mea-

Silvia De Nicolò
Department of Statistical Sciences, University of Padova
e-mail: silvia.denicolo@phd.unipd.it

Maria Rosaria Ferrante and Silvia Pacei
Department of Statistical Sciences "P. Fortunati", University of Bologna
e-mail: maria.ferrante@unibo.it, silvia.pacei@unibo.it
sures structurally more sensible to values in the tails appears to be more biased, particularly those sensible to upper tail values. This aspect should be extremely important given that inequality measures values are used for comparisons across time and location, thus if not considered, discrepancies in estimated quantities may arise from different sample sizes or different amounts of distribution skewness rather than be true inequality differences.

As regards Gini index bias, a rich literature faces it such as [21], [9], [8], [31] and see [13] for the complex survey case. However, concerning alternative measures such as Atkinson Indexes and the Generalized Entropy (GE) measures, literature on bias and high variability is very scarce. Some contributions, strictly focused on bias evaluation of the latter measures, are provided by [15] and [27] for GE measures, [4] for GE measures and Atkinson Indexes. This references adopt different approaches to correct or reduce bias but limited to the i.i.d. case. In fact, generally income data are collected via specific survey with a complex sampling design. At the same time, income data collection involves stratification of the survey population and selection of sampling units in more than one stages. Thus, the survey sample selection process, together with ex-post treatment procedures such as recalibration and imputation, invariably introduces a complex correlation structure in the data, which makes the development of a theoretically valid bias correction challenging. In this paper we will focus on bias issue in small samples, providing a bias correction when dealing with complex survey data, for a set of measures including Gini, Atkinson Indexes and Generalized Entropy measures, together with the Coefficient of Variation measure, deterministically related to the latter family. The bias correction strategy, constituting as a generalization of [4] for the complex survey case, is set out in Section 2. A simulation study involving Italian EU-SILC income data is provided in Section 3. Conclusions are drawn in Section 4.

2 Bias Correction of Complex Survey Estimators

The bias of the IM complex survey estimators in small samples can be due to the structure of IM as non linear function of unbiased (or sometimes even biased) estimators, leading them to be consistent but biased. As stated by [4], the relationship between the expectation of the sample measure of inequality $\hat{\theta}$ and its true population value $\theta$ is:

$$E(\hat{\theta}) = \theta + O\left(\frac{1}{n}\right).$$

It is important to point out that the technique proposed hereafter does not require any parametric assumptions on the income distribution, in order to
provide a flexible framework for the bias approximation that does not account for different distributional assumptions.

Suppose we have a finite population \( U \) of \( N(< \infty) \) elements labeled as \( \{1, \ldots, N\} \). Let \( y_k \) be a characteristic of interest, in our case income, for the \( k \)-th unit of the finite population, \( k = 1, \ldots, N \), we are interested in a variety of non linear functions of income values. Let denote \( S \) a sample of size \( n \), drawn using a complex probability sampling design, with \( p(s) = Pr(S = s) \) the probability of selecting the particular sample \( s \subset U \), thus \( p(s) \geq 0 \) and \( \sum_{s \in S} p(s) = 1 \). The inclusion probability of unit \( k \) is denoted with \( \pi_k \) and defined such that \( \pi_k = Pr(k \in S) \). We consider the generic inequality measure, pertaining to the Generalized Entropy and Atkinson index family, written as a function of the mean \( \mu \) and the expected value of a generic function \( g(\cdot) \) which we will generically refer to as \( \gamma = E(g(y)) \). The population value for the generic inequality measure is

\[
\theta = f(\mu, \gamma).
\]  

The related estimator in our complex survey framework is \( \hat{\theta} = f(\hat{\mu}, \hat{\gamma}) \) in which Horvitz-Thompson complex survey estimators of the mean and the generic component \( \gamma \) are plugged in, i.e.

\[
\begin{align*}
\hat{\mu} &= \frac{\sum_{i=1}^{n} w_i y_i}{N} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \\
\hat{\gamma} &= \frac{\sum_{i=1}^{n} w_i g(y_i)}{N} = \frac{\sum_{i=1}^{n} w_i g(y_i, w_i)}{\sum_{i=1}^{n} w_i}.
\end{align*}
\]  

where \( w_k = 1/\pi_k \) or a treated and recalibrated version of it, see [19] who uses a similar approach to express inequality indices in order to derive their asymptotic standard error and [4]. Considering a full design-based approach, the framework of [4] developed for i.i.d. observations, still holds in case of finite population and dependency of observations.

By simply applying a second order Taylor’s series expansion of the sample estimators around the population values, the bias is:

\[
E(\hat{\theta} - \theta) = f_{\gamma}(\gamma, \mu)E(\hat{\gamma} - \gamma) + \frac{1}{2} f_{\gamma,\gamma}(\gamma, \mu)V(\hat{\gamma}) + f_{\gamma,\mu}(\gamma, \mu)Cov(\hat{\gamma}, \hat{\mu}) + \frac{1}{2} f_{\mu,\mu}(\gamma, \mu)V(\hat{\mu}) + O_p(n^{-2}),
\]  

where \( f_{\gamma} = \frac{\partial f(\gamma, \mu)}{\partial \gamma} \) and \( f_{\gamma,\gamma} = \frac{\partial^2 f(\gamma, \mu)}{\partial \gamma^2} \).

An alternative method that should fit better in the case of small sample, is the small-\( \sigma \) approximation described in [29]. However this framework when facing non i.i.d. assumptions requires high order moments and cross-moments estimation, which results quite challenging in cases of multi-stage surveys and distribution-free setting. For some uni-stage designs, the unbiased moments estimation issue is faced by [12].

Mind the Income Gap: the Estimation of Inequality Indexes in Small Samples
2.1 Estimation of the Direct Estimators Variances and Covariances

As regards the estimation of the design variances of linear and linearized estimators, we will consider a complex survey design involving stratification and multi-stage selection, with both Self-Representing (SR) - included at the first stage with probability one - and Non Self-Representing (NSR) strata, consistently with the most common income survey designs and, in general, with household surveys whereas firm surveys are based on a single stage stratified sample design, also a special case of the two stage selection. The Horvitz Thompson variance estimator in the case of linear (or linearized) estimators, such as \( \hat{\mu} \), and when \( w_i = 1/\pi_i \), is:

\[
V(\hat{\mu}) = \frac{1}{N^2} \sum_{i \in S} \frac{y_i^2}{\pi_i^2} \left( 1 - \pi_i \right) + \frac{1}{N^2} \sum_{i \in S} \sum_{j \in S, i \neq j} \frac{y_i y_j}{\pi_i \pi_j} \frac{\pi_{i,j} - \pi_i \pi_j}{\pi_{ij}}
\]

with \( \pi_{ij}, \forall i, j \in U, i \neq j \) second order inclusion probabilities i.e. the probability that the sample includes both \( i \)-th and \( j \)-th units, \( \pi_{ij} = \sum_{s \in S_i \cap S_j} p(s) \). However generally (a) \( w_i \neq 1/\pi_i \) and (b) \( \pi_{ij}, \forall i, j \in U, i \neq j \) are difficult to calculate under complex sampling designs. Thus, the variance estimator considered constitutes as a approximation, relying on simplified assumptions, such as assuming Primary Sampling Units (PSU) are sampled with replacement, and multi-stage sampling is reduced to a single stage process using the Ultimate Clusters technique [20]. Moreover we take into account the hybrid nature of the probability scheme, blending a stratified design variance estimator for the SR strata with a finite population correction factor and a typical Ultimate Cluster approximation-based variance estimator for the multi-stage scheme in NSR strata, widely used in official statistics, see [24] for Eurostat procedures. Therefore, considering without loss of generality a two-stage scheme, let \( \hat{\psi} = \sum_h \sum_d \sum_i w_{hd} y_{hd} \) with \( h \) stratum indicator, \( d \) Primary Sampling Unit (PSU) indicator and \( i \) Secondary Sampling Unit indicator (SSU), be a linear estimator for \( \psi \), its standard error estimate is as follows:

\[
\hat{V}(\hat{\psi}) \cong \sum_{h=1}^{H_{SR}} V(\hat{\psi}_h) + \sum_{h=1}^{H_{NSR}} V(\hat{\psi}_h)
\]

\[
= \sum_{h=1}^{H_{SR}} M_h^2 (1 - f_h) \frac{s_h^2}{m_h} + \sum_{h=1}^{H_{NSR}} n_h s_{\hat{\mu}}^2
\]

\[
= \sum_{h=1}^{H_{SR}} M_h \frac{M_h - m_h}{m_h(m_h - 1)} \sum_{i=1}^{m_h} (y_{hi} - \bar{y}_h)^2 + \sum_{h=1}^{H_{NSR}} n_h \frac{n_h - 1}{n_h} \sum_{d=1}^{n_h} (\hat{\psi}_{hd} - \bar{\psi}_h)^2,
\]
with $H_{SR}$ self-representative and $H_{NSR}$ non self-representative strata, $M_h$ number of resident households in strata $h$, $m_h$ number of sample households in strata $h$, $f_h = m_h / M_h$ finite population correction factor, $n_h$ number of PSUs in strata $h$. Consider $\bar{y}_h = \sum_{i=1}^{m_h} y_{hi} / m_h$, $\hat{\psi}_{hd} = \sum_{i=1}^{m_h} w_{hdi} y_{hdi}$ with $i$ household label of stratum $h$ and PSU $d$, with a total of $m_i$ per PSU and $\hat{\psi}_h = \sum_{d=1}^{n_h} \hat{\psi}_{hd} / n_h$, lastly $n_h$ is the number of PSU in stratum $h$. If however $n_h = 1$ for some strata, the estimator (8) cannot be used. A solution is to collapse strata to create “pseudo-strata” so that each pseudo-stratum has at least two PSUs. Common practice is to collapse a stratum with another one that is similar w.r.t. the target variables of the survey [25].

2.2 Linearization of a non-linear design estimator $\gamma$

When referring instead to the variance of $\hat{\gamma}$, this quantity has to be clearly linearized in order to apply the variance estimator hereabove. In doing so, we will apply the generalized linearization method [11], [10], and [23]). This method allows to encompass more non-linear statistics than the Taylor method, and at the same time it does not involve more calculations, therefore being in general more flexible [23]. Moreover, this method works better when dealing with small samples (e.g., for small domain estimation) [25].

In particular, this procedure as stated by [2], reconciles the two approaches introduced by [11] and [10], both relying on the concept of influence function, a framework borrowed from robust statistics [18]. The same linearized variables can be found by applying the method proposed by [17] and [2] that constructs a linearized variable on the basis of a Taylor expansion with respect to sample inclusion indicators. The same method has been used directly to estimate the variance of some inequality measures estimators by linearization by [16].

Following the theoretical framework of [2], we could say that the influence of a unit $k$ on a population parameter of interest $\theta$ in a population $\mathcal{U}$, is determined by an infinitesimal variation in the importance assigned to the unit. By expressing the parameter as a functional $\theta = T(M)$ based on a measure $M(\cdot)$ such that

$$
\begin{cases}
M(x) = 1 & x = x_k \quad \forall k \in \mathcal{U}, \\
M(x) = 0 & \text{otherwise.}
\end{cases}
$$

The specialization of the general measure $M$ into a discrete measure turns the functional $T$, predefined on a continuum, into a discrete functional, in the same way as the total $Y$ is defined as the sum of all $y_k$ over the given finite population. The influence function of $T$ that in this case constitutes as the linearized variable $z_k$, is defined as the functional derivative
where \( \delta_k \) is a Dirac measure for unit \( k \). In practice, we have known data only from a sample \( s \) and \([11]\) defines a linearized variable \( \hat{z}_k \) or empirical influence function, by determining the limit above using differential calculus and replacing the unknowns in the evaluation with the corresponding estimated quantities using the sample. \([2]\) present the Demnati-Rao approach as resulting from Deville’s framework when the measure \( M \) is not the discrete measure defined for \( U \) as described above, but rather the following measure defined for \( s \) such that

\[
M(x) = w_k \quad x = x_k \quad \forall k \in s, \\
M(x) = 0 \quad \text{otherwise}.
\]

This alternative approach completely overcomes the problem that the starting point of Deville’s approach is the population parameter and not the estimator that is proposed to be used for the evaluation using the sample. Note that to the extent that the functional is expressed as an explicit function of the variables that are the weights assigned by the measure \( M \) to the observations, this linearized variable is in fact a function of the partial derivatives with respect to the weights, i.e.

\[
I[T(\hat{M})]_k = \hat{z}_k = \frac{\partial T(\hat{M})}{\partial w_k} = \frac{\partial}{\partial w_k} \sum_{i=1}^{n} w_i g(y_i) = g(y_k),
\]

once obtained the linearized variable \( V(\hat{\gamma}) \approx V(\sum_{i=1}^{n} w_i \hat{z}_i / \sum_{i=1}^{n} w_i) \). As regards the estimation of the direct estimators covariance \( Cov(\hat{\gamma}, \hat{\mu}) \) let us consider that

\[
Cov(\hat{\gamma}, \hat{\mu}) \approx \frac{1}{2} \left[ V \left( \frac{\sum_{i=1}^{n} w_i \hat{z}_i}{N} + \hat{\mu} \right) - V \left( \frac{\sum_{i=1}^{n} w_k \hat{z}_i}{N} \right) - V(\hat{\mu}) \right].
\]

Thus, a possible estimator for the design covariance would be simply obtained by plugging in the variance estimates previously mentioned, while the first term of the left-hand side of the formula could be estimated by simply applying (8).

### 3 Simulation Study

A first design-based simulation study has been carried out in order to evaluate the bias correction framework exposed in the previous section. In this simulation, the cross-section EU-SILC sample (2017 wave) has been assumed as the target population and the administrative Regions as small areas. The study
is based on real data, rather than use data generated under some distribution models, in order to check whether this specific framework can work with close-to-reality income data, mostly affected by contamination and therefore quite "dirty".

However, those data have been treated in order to circumvent non-robustness problems. The causes related to the non-robustness of the inequality measures are widely explained by [7]. The intrinsic properties of those measures are even exacerbated when applied to income data, which are traditionally affected by the extreme values problem, see [30]. It has been widely tested that generally those measures appear to be unrobust even to an infinitesimal amount of data contamination, especially when dealing with extreme values on the tails, this unrobustness depends clearly on the type of measure we are dealing with. The issue of robust estimation of economic indicators based on a semi-parametric Pareto upper tail model is well-established in literature see [6] for a review and [1] for a specification suitable for survey data. On the contrary, the issue of robust treatment of outlier in the lower tail of income distribution appears less established, see [30], [22]. As regards the upper tail we operated a semi-parametric pareto-tail modeling procedure using a proper robust estimator, PITSE proposed by [14], which blends very good performances in small samples and a fast computational implementation, as suggested by [6]. As regards the lower tail extreme value treatment, we used an inverse pareto modification of PITSE estimator, suggested by [22]. In our simulations the treatment has been done to the original EU-SILC sample and the detection of outliers has been carried out following [26] by using a Generalized Boxplot outlier detection procedure proposed by [5]. In order to deal with skewed and heavy-tailed distribution, [5] suggested modifying the whiskers of the boxplot by a simple rank-preserving transformation that allows for a so-called fitting for Tukey g– and h– distribution. Based on the simulations performed by [26], the generalized boxplot gives a higher power value compared to the adjusted and standard boxplot.

<table>
<thead>
<tr>
<th>CV</th>
<th>GE(0)</th>
<th>GE(1)</th>
<th>GE(2)</th>
<th>Atk(1)</th>
<th>Atk(2)</th>
<th>Atk(0.5)</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5%</td>
<td>ARB: -11.9 -13.1 -15.3 -17.0 -14.3 -18.3 -14.2 -14.2</td>
<td>AARE: 25.8 44.0 42.6 47.6 40.6 38.2 41.6 24.3</td>
<td>ARB: -5.6 -3.5 -6.3 -8.7 -4.7 -8.6 -4.9 -8.2</td>
<td>AARE: 25.9 46.2 44.6 49.9 42.5 39.3 43.6 26.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.0%</td>
<td>ARB: -7.4 -6.4 -8.3 -10.3 -7.1 -9.4 -7.3 -7.1</td>
<td>AARE: 19.8 31.6 31.7 37.8 29.1 27.2 30.2 16.5</td>
<td>ARB: -2.8 -0.6 -2.2 -3.5 -1.1 -2.8 -1.4 -1.5</td>
<td>AARE: 20.4 33.0 33.3 40.2 30.4 28.4 31.7 19.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Percentage ARB and AARE for the 21 small domains.
Fig. 1: Relative Bias of non-corrected measures (grey line), and of corrected measures (blue line) in 3% samples.
From the assumed population, we repeatedly select 1,000 two-stage stratified samples, mimicking the sampling strategy adopted in the survey itself: in the first stage, Self-Representing (SR) strata are always included in the sample, while a stratified sample of PSU in Non Self-Representing (NSR) strata is selected; in the second stage, a systematic sample of households is selected from each municipality included in the first stage. We repeated the drawing for two different overall sampling rates, 1.5% and 3% respectively. Results are set out in Table 1 with the Average Relative Bias (ARB) and the Average Absolute Relative Error (AARE) in percentage, calculated for the 1000 samples and the 21 regions defined as:

\[
ARB = \frac{1}{21} \sum_{r=1}^{21} \frac{1}{1000} \sum_{m=1}^{1000} \left( \frac{\hat{\theta}_{m,r}}{\theta_r} - 1 \right)
\]

\[
AARE = \frac{1}{21} \sum_{r=1}^{21} \frac{1}{1000} \left( \frac{\sum_{m=1}^{1000} \left| \frac{\hat{\theta}_{m,r}}{\theta_r} - 1 \right|}{\sum_{m=1}^{1000} \left| \frac{\hat{\theta}_{m,r}}{\theta_r} - 1 \right|} \right)
\]

In our simulation setting the small area sample size ranges from 6 to 32 households (from 6 to 96 individuals) for the 1.5% sampling rate, and from 10 to 74 households (11 to 196 individuals) for the 3% sampling rate. Figure 1 clearly illustrates the negative correlation between sample size and average relative bias in the 21 Italian regions for both the direct estimator \( \hat{\theta} \) and the bias corrected estimator \( \hat{\theta}_{corr} \). The reduction of the bias provided by the correction is noticeable for all measures, leading to slightly biased (1-2%) or even approximately unbiased estimates when \( n \geq 20 \) depending on the measure. Notice that the bias correction works well for those measures not particularly sensitive to extreme observations such as GE(0), Atk(0.5) and Atk(1). In case of Gini, CV and GE(2), the correction provides good results, but it seems however to not capture totally all the bias components. This confirms the results of [3], suggesting that the coefficient of variation squared bias depends on the coefficient of skewness of the income distribution (remember that GE(2) = CV^2 / 2). Nonetheless, a reliable estimation of that quantity appears cumbersome in case of weighted data. Furthermore, the bias-correction induces a slight but negligible error increase.

### 4 Conclusions

A bias correction framework has been set out for a set of inequality measures constituting as an extension of the method proposed by [4] to the finite population case, more specifically accounting for stratified and multi-stage survey designs. The methodology proposed is based on Taylor’s expansions, where the estimation of direct estimator variances and covariances has been performed via the generalized linearization method, proposed by [11] and [10],
relying on the concept of influence functions. This methodology does not require any parametric assumptions on the income distribution, providing a very flexible framework applicable to any income distributional assumption and allowing for inequality comparisons between different sized samples. Design-based performance evaluation of the suggested bias correction has been carried out, showing a noticeable bias reduction for all measures which works particularly well for \( n \geq 20 \). Moreover, the underlined heterogeneity of sensibilities and bias across measures can guide analysts in choosing the most suitable inequality measure depending on the context.

References