Inference for the Zenga inequality index

Matti Langel\(^1\) and Yves Tillé\(^1\)

\(^1\) University of Neuchâtel, Switzerland, e-mail: matti.langel@unine.ch

Abstract

The Zenga Index is a recent inequality measure associated with a new inequality curve, the Zenga curve. The Zenga curve \(Z(\alpha)\) is the ratio of the mean income of the \(100\alpha\)% poorest to that of the \(100(1 – \alpha)\)% richest. The Zenga index can also be expressed by means of the Lorenz Curve and some of its properties make it an interesting alternative to the Gini index. Like most other inequality measures, inference on the Zenga index is not straightforward. Some research on its properties and on estimation has already been conducted but inference in the sampling framework is still needed. In this paper, we propose an estimator and variance estimator for the Zenga index when estimated from a complex sampling design. The proposed variance estimator is based on linearization techniques and more specifically on the direct approach presented by Demmati and Rao. The quality of the resulting estimators are evaluated in Monte Carlo simulation studies on real sets of income data and robustness issues are briefly discussed.

Keywords: inequality, sampling, linearization.

1 Introduction

The level of income inequality in a population is often accounted for by using the Gini index. Many discussions concerning the latter measure have arisen in the statistical and economic literature and a lot of competing inequality measures have been proposed. Some, like the Atkinson index, the Theil index or the Quintile Share Ratio, have been known and used for decades. The Zenga index (Zenga, 2007) on the other hand, is a very recent measure and a new alternative to the Gini index and other existing inequality measures and curves. Alike the Gini index, the Zenga index can be expressed by means of the Lorenz curve. However, it is also associated to a new inequality curve, the Zenga curve which provides interesting and direct interpretations on inequality. In this paper, finite population inference for the Zenga index is discussed. In the next section, the index and the curve are defined and an estimator allowing for complex sampling designs is derived. In Section 3, linearization techniques are used to present a variance estimator for the Zenga index. Some simulations on a real data set are then run in Section 4 to apply our theoretical results. The paper ends with some concluding remarks.
2 The Zenga Index and Zenga Curve

2.1 Definition and notation

Some inequality indices are synthetic values based on an underlying curve or functions. The most obvious example is the Gini index and the underlying Lorenz Curve. Although the Gini index is the main inequality measure, it does not have unanimous support from statisticians and practitioners and thus has motivated research on other curves and synthetic indices. Zenga (1984) had already proposed two curves as well as the inequality measures $\lambda$ and $\xi$. The $\xi$ index and its underlying curve have drawn particular attention (for a review see Zenga, 2007), but according to the author, it has not been widely used because it requires the estimation of the quantile function and of the inverse of the incomplete first moment.

Zenga (2007) has presented a new alternative to the Gini index and to other existing inequality measures and curves. Hereafter, the new inequality curve and index are denoted $Z(\alpha)$ and $Z$ respectively and referred to as the Zenga curve and Zenga index. They should therefore not be mistaken for anterior measures proposed in Zenga (1984) and often also referred to as Zenga indices. Although literature on the Zenga index is not as plentiful as on the Gini index, the former has had increasing attention in the scientific community. The literature includes some publications on the properties of the index (Maffenini and Polisicchio, 2010), inference and applications (Polisicchio, 2008; Greselin et al., 2009, 2010) as well as subgroup decomposition (Radaelli, 2008, 2010). The literature on the Zenga index and curve focuses also greatly on its advantages comparatively to the Gini index. Some of these features are described below.

Consider a continuous strictly increasing cumulative distribution function $F(y)$, also, let us denote $Q_{\alpha}$, the quantile of order $\alpha$, such that $F(Q_{\alpha}) = \alpha$. The quantile function can be written as the inverse of the cumulative distribution function $Q_{\alpha} = F^{-1}(\alpha)$. The Zenga curve $Z(\alpha)$ is the ratio of the mean income of the poorest $100\alpha\%$ in the distribution to that of the rest of the distribution, namely the $100(1-\alpha)\%$ richest. It is defined by

$$Z(\alpha) = 1 - \frac{L(\alpha)}{\alpha} \cdot \frac{1 - \alpha}{1 - L(\alpha)}, \quad (2.1)$$

where $0 \leq \alpha \leq 1$ and $L(\alpha)$ is the quantile share or Lorenz curve (Lorenz, 1905; Gastwirth, 1972; Cowell, 1977; Kovacevic and Binder, 1997; Langel and Tillé, 2009b), which is a central tool of inequality theory and is defined by

$$L(\alpha) = \frac{\int_0^{Q_{\alpha}} u dF(u)}{\int_0^{\infty} u dF(u)}. \quad (2.2)$$

The Zenga index, which can be written

$$Z = \int_0^1 Z(\alpha) d\alpha, \quad (2.3)$$

can thus be defined, alike the Gini index in terms of the Lorenz Curve. However, the Gini and Zenga indices differ in many ways. One argument in favor of index $Z$ is described by Greselin et al. (2010, p.3):

[...] the Gini index underestimates comparisons between the very poor and the whole population and emphasizes comparisons which involve almost identical population subgroups [...] the Zenga index detects, with the same sensibility, all deviations from equality in any part of the distribution.
A comparative simulation study regrouping 17 different inequality indices (Langel and Tillé, 2009a) seems to confirm this idea by showing that the Zenga index is one of the most appropriate measures to detect changes at any level of the income distribution and in many different situations. Another argument in favor of the Zenga index concerns interpretation. A lot of intuitive information can indeed be obtained from analyzing the curve itself. For instance, any point measure \( Z(\alpha) \) on the curve indicates that the mean income of the \( \alpha \% \) poorest is equal to \( 1 - Z(\alpha) \) times the mean income of the richest \( (1 - \alpha) \% \). Moreover, the Zenga index \( Z \) can be graphically represented together with the curve, letting the latter show directly the intervals of \( \alpha \) where inequality is lower or higher than the mean level of inequality, represented by the index itself. Finally, Maffenini and Polisicchio (2010) have shown that when adding an identical positive income to all observations, the effect on the Zenga curve is more intuitive than on the Lorenz curve. Indeed, the Zenga curve shows that, after translation, the level of inequality decreases more heavily for small incomes than for larger ones, whereas the latter intuition is not captured by the Lorenz curve.

2.2 Zenga index in finite population

Let now \( U \) denote a finite population of \( N \) identifiable units \( u_1, u_2, \ldots, u_N \). For the sake of simplicity, we will hereafter denote unit \( u_k \) by its identifier \( k \). Associated to each unit \( k \) is the value \( y_k \) of some characteristic of interest, for example the income. To lighten the notation, we will assume with no loss of generality that all \( y_k \)'s are distinct and sorted. We have

\[
Y = \sum_{\ell \in U} y_\ell, \tag{2.4}
\]

\[
Y_k = \sum_{\ell \in U} y_\ell [\ell \leq k], \tag{2.5}
\]

where \( [A] = 1 \) if \( A \) is true and 0 otherwise. Let us also denote partial sum \( Y(\alpha) \), the sum of incomes up to quantile \( \alpha \) by

\[
Y(\alpha) = \sum_{\ell \in U} y_\ell [\ell \leq k-1] + y_k [\alpha N - (k-1)] = Y_{k-1} + y_k [\alpha N - (k-1)], \tag{2.6}
\]

where the value of \( k \) is such that \( \alpha N < k \leq \alpha N + 1 \). With Expression (2.6), the finite population quantile share can be defined by

\[
L(\alpha) = \frac{Y(\alpha)}{Y}. \tag{2.7}
\]

The Zenga index for a population of size \( N \) is then

\[
Z = \sum_{k \in U} Z_k,
\]

where

\[
Z_k = \frac{1}{N} - \int_{(k-1)/N}^{k/N} \frac{Y(\alpha)}{\alpha} \cdot \frac{1 - \alpha}{Y - Y(\alpha)} d\alpha.
\]
Denoting $A_k = (k-1)y_k - Y_{k-1}$ for $k = 2, \ldots, N$ and assuming $A_0 = 0$, we obtain

$$Z_k = \begin{cases} \left( \frac{Y}{Ny_1} - 1 \right) \log \left( \frac{Y}{Y-Y_1} \right), & \text{if } k = 1, \\ \frac{A_k}{Y + A_k} \log \left( \frac{k}{k-1} \right) + \left[ \frac{Y}{Ny_k} - \frac{Y}{Y+A_k} \right] \log \left( \frac{Y-Y_{k-1}}{Y-Y_k} \right), & \text{if } k = 2, \ldots, N-1, \\ \left( 1 - \frac{Y}{Ny_N} \right) \log \left( \frac{N}{N-1} \right), & \text{if } k = N. \end{cases}$$ \hspace{1cm} (2.8)

### 2.3 An estimator of the Zenga index

A random sample $S$ of size $n$ is drawn from a finite population $U$ of size $N$ from a random sampling design, such that $p(s) = \Pr(S = s)$ is the probability of selecting sample $s \subset U$. The probability for unit $k \in U$ to be included in the sample is written $\pi_k = \Pr(k \in S)$. Also, $d_k$ denotes the design weight of $k$ such that $d_k = 1/\pi_k$. Let us also denote

$$D = \sum_{\ell \in S} d_{\ell},$$

$$D_k = \sum_{\ell \in S} d_{\ell} \mathbb{1}[\ell \leq k]$$

and

$$\alpha_k = \frac{D_k}{D}.$$  

Expressions (2.4), (2.5) and (2.6) can be respectively estimated from a sample by

$$\hat{Y} = \sum_{\ell \in S} d_{\ell} y_{\ell},$$

$$\hat{Y}_k = \sum_{\ell \in S} d_{\ell} y_{\ell} \mathbb{1}[\ell \leq k],$$

$$\hat{Y}(\alpha) = \sum_{\ell \in S} d_{\ell} y_{\ell} \mathbb{1}[\ell \leq k-1] + y_k (\alpha D - D_{k-1}) = \hat{Y}_{k-1} + y_k (\alpha D - D_{k-1}),$$

where $k$ is an integer such that $D_{k-1} < \alpha D \leq D_k$. Thus, an estimator for $L(\alpha)$ is

$$\hat{L}(\alpha) = \frac{\hat{Y}(\alpha)}{\hat{Y}}.$$ \hspace{1cm} (2.9)

With $\hat{A}_k = D_{k-1} y_k - \hat{Y}_{k-1}$ for $k = 2, \ldots, n$ and $\hat{A}_0 = 0$, a natural estimator for the Zenga index is then:

$$\hat{Z} = \sum_{k \in S} \hat{Z}_k,$$ \hspace{1cm} (2.10)
where

\[ \hat{Z}_k = \begin{cases} \left( \frac{\hat{Y}}{Dy_1} - 1 \right) \log \left( \frac{\hat{Y}}{\hat{Y} - \hat{Y}_1} \right), & \text{if } k = 1, \\ \frac{\hat{A}_k}{\hat{Y} + A_k} \log \left( \frac{D_k}{D_{k-1}} \right) + \left[ \frac{\hat{Y}}{Dy_k} - \frac{\hat{Y}}{\hat{Y} + A_k} \right] \log \left( \frac{\hat{Y} - \hat{Y}_{k-1}}{\hat{Y} - \hat{Y}_k} \right), & \text{if } k = 2, \ldots, n - 1, \\ \left( 1 - \frac{\hat{Y}}{Dy_n} \right) \log \left( \frac{D_n}{D_{n-1}} \right), & \text{if } k = n. \end{cases} \]  

(2.11)

3 Approximation of the Variance by Linearization

3.1 Linearization by the Demnati-Rao approach

Linearization regroups a variety of techniques for computing an approximation of the variance of a statistic \( \hat{\theta} \), an estimator of a function of interest \( \theta \). The idea behind these techniques is to find a linearized variable \( v_\ell \) such that

\[ \hat{\theta} - \theta \approx \sum_{\ell \in S} d_\ell v_\ell - \sum_{\ell \in U} v_\ell. \]

The variance of \( \sum_{\ell \in S} d_\ell v_\ell \), the weighted sum of the linearized variable \( v_\ell \), is then used as an approximation of the variance of \( \hat{\theta} \):

\[ \text{var} \left( \sum_{\ell \in S} d_\ell v_\ell \right) \approx \text{var} \left( \hat{\theta} \right). \]  

(3.1)

Because the variance of statistic \( \theta \) is approximated by the variance of a total, linearization methods can easily provide a variance estimator for all complex sampling designs for which an expression for the variance of a total is known. To compute the values of \( v_\ell \), however, information at the population level is often needed. Thus, \( v_\ell \) is generally replaced by its sample counterpart \( \hat{v}_\ell \).

The linearization method has been introduced by Woodruff (1971) using Taylor series. Deville (1999) has presented a more general method based on influence functions (Hampel, 1974; Hampel et al., 1985). In both methods, the linearized variable is computed on the function of interest and is then estimated on the sample. Binder (1996) has proposed a direct approach in which the linearized variable is directly computed on the estimator. However, like in Woodruff (1971), it is only adapted for smoothed functions of totals. Demnati and Rao (2004) have proposed yet another direct approach. The latter uses influence functions, and is thus of broad application. In this paper, we use the Demnati and Rao (2004) approach to derive a linearized variable of the Zenga index, and consequently a variance estimator.

The linearized variable \( \hat{v}_\ell \) is the influence function computed directly on the sample obtained by deriving the partial derivative of the estimator with respect to the weight \( d_\ell \). Once the linearized variable \( \hat{v}_\ell \) is computed, variance estimation is done in the standard framework and usual asymptotic conditions of linearization techniques (Woodruff, 1971; Isaki and Fuller,
1982; Deville and Särndal, 1992; Binder, 1996; Kovacevic and Binder, 1997; Deville, 1999). Note that the design weights $d_\ell$ are used, but the methods hold for calibration weights as well (Demnati and Rao, 2004).

### 3.2 Linearization of the Zenga index

The influence function, or linearized variable $\hat{v}_\ell$ of the Zenga Index at $y_\ell$ can be computed by

$$I(\hat{Z})_\ell = \hat{v}_\ell = \frac{\partial \hat{Z}}{\partial d_\ell} = \sum_{k \in S} \frac{\partial \hat{Z}_k}{\partial d_\ell}. \quad (3.2)$$

Thus, for each sample element $\ell$, the partial derivative with respect to $d_\ell$ of $\hat{Z}_k$ for all $k \in S$ is computed. Similarly as for point estimation, three cases are derived. We present hereafter the final expressions for $\frac{\partial \hat{Z}_k}{\partial d_\ell}$. Complete derivation is available from the authors.

$$\frac{\partial \hat{Z}_k}{\partial d_\ell} = \begin{cases} \frac{Dy_\ell - \hat{Y}}{D^2y_1} \log \left( \frac{\hat{Y}}{\hat{Y} - \hat{Y}_1} \right) + y_\ell \left( \frac{\hat{Y}}{Dy_1} - 1 \right) \left[ 1 - \frac{1}{\hat{Y}} \right] \sum_{\ell > 1} \mathbb{I}(\ell > 1), & \text{if } k = 1, \\
\frac{\hat{Y}_k (y_k - y_\ell) \mathbb{I}(\ell < k) - \hat{A}_k y_\ell}{(\hat{Y} + \hat{A}_k) Dy_k} \log \left( \frac{\hat{Y} - \hat{Y}_k}{\hat{Y} - \hat{Y}_k} \right) + \frac{\hat{A}_k}{\hat{Y} + \hat{A}_k} \left[ \frac{\mathbb{I}(\ell < k)}{D_{k-1}} - \frac{\mathbb{I}(\ell < k)}{D_k} \right] + \frac{Dy_\ell - \hat{Y}}{D^2y_k} \log \left( \frac{\hat{Y} - \hat{Y}_k}{\hat{Y} - \hat{Y}_k} \right), & \text{if } k = 2, \ldots, n - 1, \\
\frac{\hat{Y} - Dy_\ell}{D^2y_n} \log \left( \frac{D}{D_{n-1}} \right) + \frac{1 - \hat{Y}}{D} \left[ 1 - \frac{1}{D_{n-1}} \frac{1}{D} \mathbb{I}(\ell < n) \right], & \text{if } k = n. \end{cases} \quad (3.3)$$

Hence, for example, a variance estimator for the Zenga index under a simple random sampling design without replacement of size $n$ is

$$\text{var} (\hat{Z}) = \frac{N(N - n)}{n(n - 1)} \sum_{\ell \in S} (\hat{v}_\ell - \bar{v})^2, \quad (3.4)$$

with $\bar{v} = n^{-1} \sum_{\ell \in S} \hat{v}_\ell$.

In statistics, influence functions are mainly used as a tool to study robustness. Although the influence function of Deville (1999) and Demnati and Rao (2004) are modified versions of that proposed by Hampel (1974), the influence curve of the Zenga index reveals high sensitivity of the statistic to outliers. Figure 1 plots the influence curve for one of the data sets used in the simulations in Section 4 and illustrates this issue. As a result, inference is heavily affected by very large incomes. Similar results can be found in robust statistics regarding the influence function of the Gini index (Monti, 1991; Cowell and Victoria-Feser, 1996, 2003) and the Quintile Share Ratio (Hulliger and Schoch, 2009), which are both unbounded above.
4 Simulation Studies

4.1 Austrian EU-SILC data set

At first, a simulation is run on a synthetic data set based on original EU-SILC data for Austria. The data consists of 14'827 individual observations from 6'000 households. The data set is available from the laeken R-package. Here, the data at the individual level is considered as the finite population from which random samples of size \( n = 3000 \) are selected with a simple random sampling design without replacement. One thousand replications are made. In each sample, the Zenga index (Expression 2.10) and its linearization variance (Expression 3.4) are estimated. Results are summarized in Table 1. The relative bias for point and variance estimation are defined respectively by

\[
\text{RB}(\hat{Z}) = \frac{E(\hat{Z}) - Z}{Z},
\]

and

\[
\text{RB} \left[ \text{var}_{\text{lin}}(\hat{Z}) \right] = \frac{E \left[ \text{var}_{\text{lin}}(\hat{Z}) \right] - \text{var}_{\text{sim}}(\hat{Z})}{\text{var}_{\text{sim}}(\hat{Z})},
\]

where \( \text{var}_{\text{lin}}(\hat{Z}) \) stands for the estimated variance obtained with the linearization technique and \( \text{var}_{\text{sim}}(\hat{Z}) \) denotes the Monte-Carlo variance computed on the 1000 replications. Results show that point estimation is very successful and that the linearization technique only very slightly underestimates the variance with a relative bias of -1.65%. The coverage rate for a 95% confidence interval is close to the desired level.
Table 1: Simulation results (Austrian EU-SILC data, 1000 replications).

<table>
<thead>
<tr>
<th>Point estimation</th>
<th>Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Z )</td>
<td>( E(\hat{Z}) )</td>
</tr>
<tr>
<td>0.5872</td>
<td>0.5870</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \hat{\text{var}}_{\text{lin}}(\hat{Z}) )</th>
<th>( E[\hat{\text{var}}_{\text{lin}}(\hat{Z})] )</th>
<th>( RB(\hat{\text{var}}_{\text{lin}}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0310 \cdot 10^{-5}</td>
<td>2.9811 \cdot 10^{-5}</td>
<td>-1.65%</td>
</tr>
</tbody>
</table>

Coverage Rate of 95% for \( Z \) 95.9%

4.2 Taxable incomes of Canton of Neuchâtel, Switzerland

In the previous example, outliers do not have a large effect on the accuracy of estimation. Our second simulation study is run on real taxable income data in the Canton of Neuchâtel, Switzerland for year 2006. It is composed of 82’489 non-null income earners and incorporates some extreme outliers. The same strategy is used as for the first simulation study in order to allow for comparisons. To account for the outliers issue, two sets of simulations are performed: one on the full data set and one on truncated data. In the truncated data, all observations lying above \( Q_{0.999} \) are deleted, involving the 83 richest income earners. As a comparison the ratio between the largest income and the median income in the full data set is 175.5, while it is only of 13.3 for the truncated data, and is even smaller for the Austrian EU-SILC data (8.4). The results are summarized in Table 2.

Table 2: Simulation results, Neuchâtel data (1000 replications).

<table>
<thead>
<tr>
<th>Point estimation</th>
<th>Variance estimation</th>
<th>Coverage Rate (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( RB(\hat{Z}) )</td>
<td>( RB(\hat{\text{var}}_{\text{lin}}) )</td>
<td>( CR ) (95%)</td>
</tr>
<tr>
<td>Full data</td>
<td>-0.08%</td>
<td>-6.79%</td>
</tr>
<tr>
<td>Truncated data</td>
<td>-0.06%</td>
<td>0.22%</td>
</tr>
</tbody>
</table>

Although point estimation is accounted for in a satisfactory manner for both situations, we can see how the relative bias of the variance estimator is affected by the presence of outliers in the sample. However, it can be advocated that even in such a severe level of skewness and presence of outliers, which we believe to be frequent when working on income data, quality of inference for the Zenga index remains reasonable with a relative bias for the variance of -6.79% and a coverage rate of 93.5 for a 95% confidence interval.

5 Conclusion

To effectively bring new insights in the study of income inequality a recent measure like the Zenga index needs a general and valid framework for inference in finite population. In this
paper, we have firstly proposed an estimator of the Zenga index which accounts for the presence of weights. Secondly, a variance estimator has been presented. The Demnati and Rao linearization technique has been used to derive an estimator that can be applied to samples selected from a complex sampling design. The theoretical results have then been tested successfully in simulation studies. Skewness and robustness issues have been briefly discussed. These issues should probably be studied in depth as they are probably the most challenging questions regarding income inequality, a topic of statistics and economics in which outliers and asymmetry have to be handled with particular care.

Acknowledgements

The authors are grateful to the Office Cantonal de la Statistique (Canton de Neuchâtel, Switzerland) and especially to Gérard Geiser for the dataset. This research is supported by grant no. 200021-121604 of the Swiss National Science Foundation.

References


