https://doi.org/10.35662/unine-thesis-3079

FACULTÉ DES SCIENCES
Institut de Statistique

ON MEASURING INCOME INEQUALITY

by

Ziqing Dong

Thesis submitted in fulfillment of the requirements
for the degree of
Doctorat ès Sciences

Accepted by the examination committee:

Prof.	Alina	Matei	Université de Neuchâtel	Jury president
Prof.	Yves	Tillé	Université de Neuchâtel	Thesis director
Prof.	Camelia	Goga	Université de Franche-Comté	Jury member
Prof.	Anne	Ruiz-Gazen	Université Toulouse	Jury member
Dr.	Alessio	Guandalini	Istituto nazionale di statistica	Jury member

Thesis defended on 30 November 2023.

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THÈSE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Monsieur Ziqing DONG

Titre :

"On measuring Income Inequality"

sur le rapport des membres du jury composé comme suit :

- Prof. Yves Tillé, directeur de thèse, Université de Neuchâtel,
- Prof. Alina Matei, rapporteure, Université de Neuchâtel
- Prof. Camelia Goga, rapporteure, Université de Franche-Comté, France
- Prof. Anne Ruiz-Gazen, rapporteure, Université Toulouse, France
- Dr Alessio Guandalini, rapporteur, Instituto nazionale di statistica, Rome, Italie

Neuchâtel, le 1er décembre 2023

Le Doyen, Prof. R. Bshary

Table of Contents

Acknowledgements

This doctoral thesis could not be written without the help and guidance of my PhD supervisor Prof. Yves Tillé. It has been a great pleasure to work with Prof. Yves Tillé, who is always calm and approachable. His interest in income inequality measures influences and leads me to work on it for my PhD thesis. The Institute of Statistics at the University of Neuchâtel has been a pleasant and welcoming place to work in under his directorship. His supervision is decent, which smooths the everyday research work for everybody.

A part of my doctoral research was conducted at the Institute of Statistical Mathematics, Japan. I would like to express my gratitude to Prof. Satoshi Kuriki, who hosted and supervised my research stay in Japan. His deep grasp of statistical mathematics makes each discussion fruitful. I would also like to thank Prof. Kunio Shimizu, who has always been kind to me, both academically and personally.

My thanks also go to Prof. Alina Matei, Dr. Alessio Guandalini and Prof. Giovanni Giorgi. Prof. Matei has given me many constructive suggestions for my scientific research and career development. Dr. Alessio Guandalini and I have had several meetings during my Ph.D. years. It has been being always a great pleasure to collaborate with him. Prof. Giovanni Giorgi was an expert in the research area of income inequality measures, which for me was certainly an honour to have him in the team. Very sadly, I must mention that Prof. Giovanni Giorgi passed away two years ago.

Preface

Income inequality is a profound subject. My research by no means aims to cover every aspect of the subject. Comprehending income inequality, in my humble opinion, apart from statistical research, requires a deep investigation of human society, history and philosophy. Statistics does not help solve the questions of income inequality per se. Nevertheless, statistics provides an approach to measure it. This PhD research concentrates on the questions of measuring income inequality. It focuses on the objectivity of the income inequality measures, the accuracies of the estimation of them and the quantifications of the uncertainty of the estimation.

1 Introduction

Since the end of the nineteenth century, the degree of inequality in the distribution of income and wealth has been a matter of great interest. Its increase is considered as a brake for economic growth and a danger for the political and economical stability of a country. While, on the contrary, its decrease is a symptom of well-being and of best prospects for the future. The Nobel prize winner Joseph Stiglitz states that income inequality is an important measure for forecasting the wealth of a country. He has argued in his book, *The Price of Inequality*, that income inequality is detrimental to economic growth (Stigilitz, 2012).

The study of income inequality contributes to the understanding of economic and social disparities of a society. Its vital importance has drawn attention from not only governments and economic institutions, but also researchers in different disciplines such as philosophy, social sciences and as well as statistics (Arestis, 2018). It has been the subject of numerous publications (Silber, 1999). It is considered to be linked with many health and social problems (Dabla-Norris et al., 2015; Pickett and Wilkinson, 2010). Therefore, it is crucial to have an index for keeping track of the level and trend of income inequality and for monitoring economic policies or forecasting their effects on it.

The Gini (1914) index is the most famous and widespread inequality measure. Since Corrado Gini suggested the index, it has been the subject of numerous publications. It is available for almost all the countries in the world from various international organisations' datasets, such as the World Bank 'Inequality around the world' dataset, the UNU-WIDER World Income Inequality Database, the EurLIFE database and the UNDP annual report (Decancq and Lugo, 2012). Furthermore, in the most developed countries, national statistical institutes carry out yearly surveys and measure the income inequality and its evolution through time (see Osier, 2009). Its use is not restricted only to the economic field. It is surprising to note that even after a century, different applications of the Gini index pop up in new fields (see, e.g., Giorgi, 2019). To name a few, there have been applications of the Gini index to analyse earthquake damages to buildings (Jihui Tu and Han, 2017), vehicular terror attacks (Hasisi et al., 2020), demands in public transport (Hörcher and Graham, 2021), error distributions in computational chemistry (Pernot and

Savin, 2021) and coronavirus infection rates in cities (Arbel et al., 2022).

Recent studies suggest that the use of just one inequality index is not sufficient to have a complete picture on income inequality. Although the Gini index may have several excellent properties, it is inadequate to rely exclusively on it for inequality measurement. Figure 1.1 illustrates income inequality measurement for two distinctive economies using the Gini index. The red area represents an economy with the lowest 75% of the population owning 25% of the total income while the rest 25% of the population sharing the rest of the income, and the blue area represents an economy with half of the population having zero income and the rest equally sharing the whole income. The Gini index, of which the value equals twice the shaded area, is measured to be 0.5 for both economies. Obviously, in such cases, the Gini index is not able to capture the differences of the inequality between the red and the blue economies.

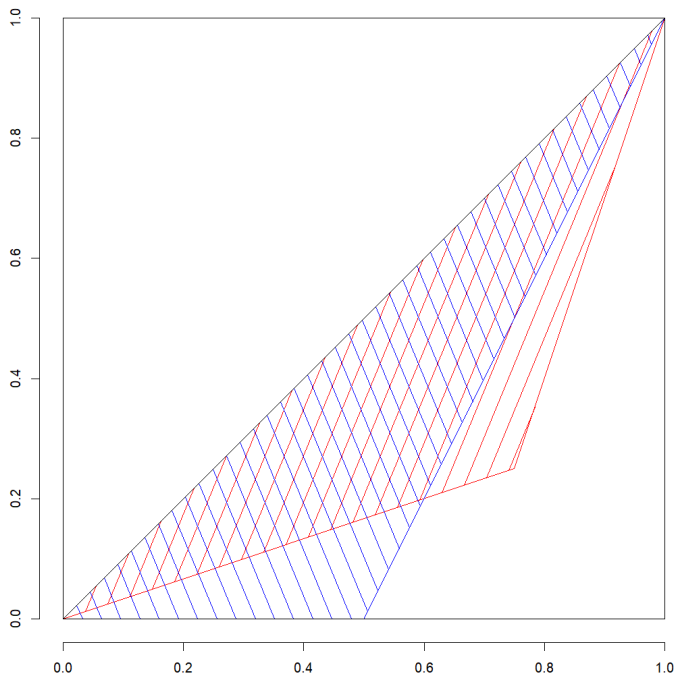


Figure 1.1: Gini index measuring the income inequality levels for the red and the blue economies.

As a consequence, researchers are more inclined to encourage the use of more than

one inequality index simultaneously to better catch the inequality in different parts of the income distribution and thereby to better understand the socio-economic reality and political significance of inequality (Piketty, 2015; Osberg, 2017). Indeed, each inequality measure proposed in the literature has its own sensitivity to different parts of the income distribution. That is, it pays more attention to a certain part of the income distribution while diminishing the other parts. Therefore, it is necessary to be backed up by more than one inequality measure to prevent this drawback and to have a global view of the inequality. The most suitable candidate to place side by side with the Gini index could be the Bonferroni inequality index (Bonferroni, 1933), of which the introduction is given in Chapter 2. As an illustration, suppose that both the red and the blue economies in the example of Figure 1.1 have a population of size $N = 20$. The Bonferroni index for the red economy is equal to the red shaded area in Figure 1.2 multiplying by a normalising constant $\frac{N}{N-1} = \frac{20}{19}$, and is measured to be 0.588. Similarly, the Bonferroni index measured for the blue economy equals the blue shaded area multiplying by $\frac{20}{19}$, and is 0.704. Thus, the Bonferroni index in this case has the ability to capture the differences of the two economies, which the Gini index fails to do. Of course, one might find a counter-example, which the Bonferroni index fails to measure the differences, but the Gini index succeeds to do.

Due to the fact that income data are usually collected through sample surveys, sampling properties of the income inequality indices estimators should not be overlooked. Although a vast amount of research on the inferential problems of the Gini index has been done, little is studied for the Bonferroni index. While revealing the properties of two indices, Chapter 2 gives a detailed analysis on the inferential aspects of the Bonferroni index. The Bonferroni index is estimated by its two estimators, and the Graf linearisation method (Graf, 2011; Graf and Tillé, 2014) is applied for the approximation of the variances of them. The Graf linearised variables are essentially the partial derivatives of the statistic with respect to the indicator variables of the presence of the units in the sample, adjusting for sample weights. The variances of the Bonferroni inequality index estimators are estimated under simple random sampling and stratified simple random sampling. The choice of the Graf method is demonstrated by its accuracy and

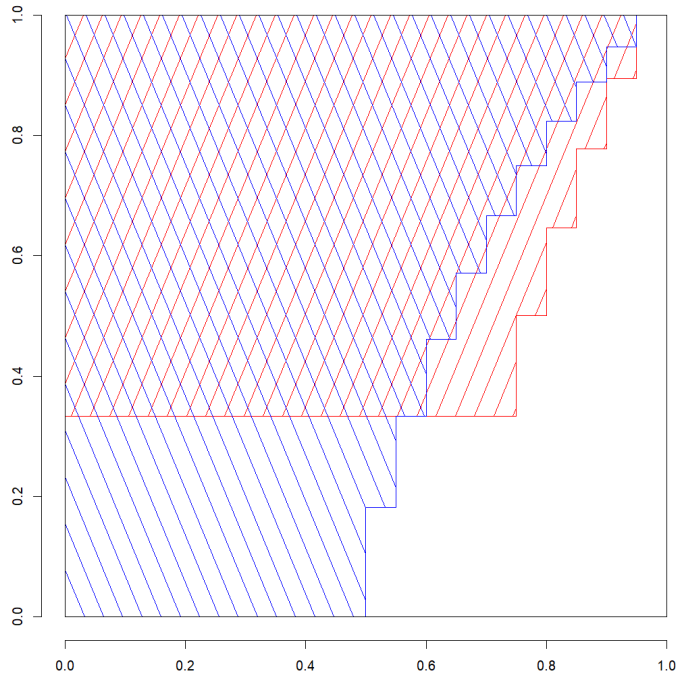


Figure 1.2: Bonferroni index measuring the income inequality levels for the red and the blue economies.

simplicity. In addition, the Graf method could be used as long as the expression of the variance estimator of the total estimator under the given sampling design is known. The Graf linearised variables could also be viewed as the discrete analogues of the influence functions in the cases of the Gini and the Bonferroni indices for illustrating the sensitivity to different parts of the income distribution.

Except from the Gini and the Bonferroni indices, many other income inequality indices have been proposed in the research literature of inequality measurement. Reflections on the various existing inequality indices in the same framework lead to the study of a generalised income inequality index (Dong et al., 2023). To be specific, a generalised index with two parameters controlling the sensitivity to different levels of the income distribution is proposed in Chapter 3. Tuning the two parameters, several well-known inequality indices (such as the Bonferroni index, the Gini index, the Mehran index, the Piesch index, the De Vergottini index, the Pietra index also known as the Robin Hood

index) descend from or are related to the expression of the generalised income inequality index. Some new inequality indices are created and derived according to the required sensitivity for filling the gap in the literature. Investigations into the generalised index make it possible to analyse and compare the influence of the different levels of incomes on the inequality indices in the same framework. Once and for all, their sensitivity to different parts of the income distribution could be established.

With the proposal of the generalised income inequality index, the estimation procedures (see, i.e., the estimators of the Bonferroni index in Chapter 2) could also be generalised and applied for the estimation of the various inequality indices in the same framework. Three methods, the rectangular rule, the trapezoidal rule and the reformulation method, are developed for defining the inequality indices and their estimators in finite populations. The estimators defined using the three methods for the same index converge very fast to the same value when sample size increases. Their properties, however, differ when the sample size is small. Furthermore, as a result of the generalised index, the influence functions of the various income inequality indices in the same framework are also unified.

The Lorenz curve and the Bonferroni curve are the two crucial curves for illustrating the concentration of income. The Gini index and the Bonferroni index are computed based on the two curves, respectively. The last part of the thesis studies the construction of simultaneous confidence bands for the two curves. The Lorenz curve and the Bonferroni curve include errors in explanatory variables. The construction of simultaneous confidence bands for such curves is treated completely differently from ordinary regression curves, and is a subject which seems to have not been studied so far. The procedure developed could be applied for building simultaneous confidence bands for similar curves when point estimators on the curves are established.

This doctoral thesis is organized as follows. Chapter 2 focuses on the linearisation and variance estimation of the Bonferroni inequality index, while Chapter 3 introduces and studies the generalised income inequality index. The construction of simultaneous confidence bands for the Lorenz and the Bonferroni curves is discussed in Chapter 4. Chapter 5 concludes the thesis and gives the final remark.

2 Linearisation and Variance Estimation of the Bonferroni Inequality Index

Abstract

The study of income inequality is important for predicting the wealth of a country. There is an increasing number of publications where the authors call for the use of several indices simultaneously to better account for the wealth distribution. Due to the fact that income data are usually collected through sample surveys, the sampling properties of income inequality measures should not be overlooked. The most widely used inequality measure is the Gini index, and its inferential aspects have been deeply investigated. An alternative inequality index could be the Bonferroni inequality index, although less attention on its inference has been paid in the literature. The aim of Chapter 2 is to address the inference of the Bonferroni index in a finite population framework. The Bonferroni index is linearized by differentiation with respect to the sample indicators which allows for conducting a valid inference. Furthermore, the linearized variables are used to evaluate the effects of the different observations on the Bonferroni and Gini indices. The result demonstrates once for all that the former is more sensitive to the lowest incomes in the distribution than the latter. [1]

Keywords: Bonferroni, Gini, inequality measures, inference, influence function.

2.1 Introduction

There is an increasing number of publications where the authors call for the use of several indices simultaneously to better account for the wealth distribution. The most widely used inequality measure is the Gini index (Gini, 1914), and its inferential aspects have been deeply investigated. The most suitable candidate to place side by side with the Gini index could be the Bonferroni inequality index (Bonferroni, 1933). In fact, Pundir et al. (2005) show that both can be derived from the Lorenz Curve (Lorenz, 1905). Indeed, the two indices share several properties while maintaining some very interesting peculiarities.

The opposition between the Bonferroni index and the Gini index is rooted when Carlo Emilio Bonferroni proposed his index in 1930. In the beginning, the Bonferroni index was fought by Corrado Gini and his followers who were very fond of the Gini index and who tried to avoid the use of any other measures that took the Gini index down the line (Giorgi, 1998). Only in the last forty years, the Bonferroni index has been rediscovered by

[1] This chapter is an adaptation of: Ziqing Dong, Yves Tillé, Giovanni M. Giorgi and Alessio Guandalini (2022). Linearisation and variance estimation of the Bonferroni inequality index. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 184(3)*, 1008-1029. The author received the Cochran-Hansen Prize 2023 awarded by the International Association of Survey Statisticians, a section of the International Statistical Institute, for the paper.

Piesch (1975) and Nygård and Sandström (1981). Several extensions and interpretations proposed for the Gini index (see, e.g., Giorgi, 2005, for a comprehensive review) have then been extended to the Bonferroni index, disclosing even more similarities and differences between the two indices. Among them, welfare implication (see, e.g., Benedetti, 1986; Aaberge, 2000; Chakravarty and Muliere, 2004; Chakravarty, 2007; Bárcena-Martín and Silber, 2013, 2017), socio-economic aspects (Bárcena and Imedio, 2008; Silber and Son, 2010; Bárcena-Martín and Silber, 2011; Imedio-Olmedo et al., 2012; Bárcena-Martín and Silber, 2013), application in fuzzy and reliability frameworks (Giordani and Giorgi, 2010; Giorgi and Crescenzi, 2001b) and decomposition by sources or by groups (Tarsitano, 1990; Bárcena-Martín and Silber, 2013; Giorgi and Guandalini, 2018) have been studied.

Because income data are usually collected through sample surveys, the sampling properties of the Bonferroni and Gini indices should not be overlooked. Inference on the Gini index is a tricky problem which has generated a large number of publications (see, e.g., Langel and Tillé, 2013; Graf and Tillé, 2014; Giorgi and Gigliarano, 2017, and reference therein). However, mainly for the reasons stated previously, less attention has been paid on the inference of the Bonferroni index. Giorgi and Mondani (1994, 1995a) derive the sampling distribution of the Bonferroni index from exponential population, while Giorgi and Crescenzi (2001a) propose Bayes estimators of it from a Pareto-type I population. Furthermore, Giorgi and Nadarajah (2010) explicate the Bonferroni index for several distributions. Aaberge (2007) studies the Bonferroni index as an inequality measure belonging to the Gini's nuclear family, and Pundir et al. (2005) provide asymptotically distribution-free statistical inference for it. Finally, Giorgi and Guandalini (2013) study a sampling estimator less biased for small samples than the plug-in estimator.

In Chapter 2, the linearisation technique developed by Graf (2011) has been applied for estimating the variances of the Bonferroni index estimators. Moreover, the obtained linearised variables are interpreted as influence function in the sense of Hampel (1974) and Hampel et al. (1985). The comparison with the linearised variables of the Gini index helps to evaluate the effects of the different observations on the Bonferroni and Gini indices. The relation between the Bonferroni index and the Gini index stated several times in the literature (De Vergottini, 1950; Pizzetti, 1951) is confirmed, and it is demonstrated once

for all that the former is more sensitive to the lowest incomes in the distribution than the latter.

Chapter 2 is organised as follows: In Section 2.3, the Bonferroni index is defined in an infinite population. Section 2.4 provides the estimation of the Bonferroni index in a finite population framework. In Section 2.5, the linearisation method developed by Graf is applied to the Bonferroni index estimators. In Section 2.6, the procedure for deriving the sampling variances of the Bonferroni index estimators is described. The accuracy of the Bonferroni index estimators and their associated variance estimators are tested in Section 2.7. In Section 2.8, the influence function of the Bonferroni index is derived and analysed along with the influence function of the Gini index. The relation between the influence function and the proposed linearised variables is discussed, and through an empirical demonstration, it is shown that the Bonferroni index is more sensitive to the lowest incomes in the distribution than the Gini index. Finally, Section 2.9 contains some concluding remarks.

2.2 Notation

Consider a non-negative continuous random variable Y with probability density function $f(y)$ and finite mean $\mu = \int_0^\infty yf(y)\ dy \neq 0$.

Let Y denote income (or, in some cases, wage, turnover, profit or consumption expenditure). Its cumulative distribution function is given by

$$F(y) = \int_0^y f(t)dt.$$

Assume also that $F(y)$ is absolutely continuous and at least twice differentiable. Such F will be interpreted as income distribution and $p = F(y)$ as the probability that Y will have a value less than or equal to y, $p \in [0, 1]$.

Let

$$\mu(y) = \frac{\int_0^y tdF(t)}{F(y)}$$

be the partial mean. Its quantile function is defined by $Q(p) = \inf\{y \,|\, F(y) \geq p\}$, $p \in [0, 1]$. The Lorenz curve (1905) can be expressed as

$$L\left(p\right) = \frac{1}{\mu}\int_0^p Q\left(\alpha\right) d\alpha$$

(see, e.g., Pietra, 1915; Gastwirth, 1971). The Lorenz curve can also be defined using the partial mean:

$$L(p) = \frac{p\,\mu(Q(p))}{\mu}.$$

The Bonferroni curve is

$$Bon(p) = \frac{L(p)}{p},$$

and the complementary Bonferroni curve is

$$\overline{Bon}(p) = 1 - Bon(p)$$

(Bonferroni, 1930; Giorgi and Mondani, 1995b; Pundir et al., 2005). When p approaches 0, according to the L'Hôpital's rule:

$$\lim_{p\to 0^+} \overline{Bon}(p) = \lim_{p\to 0^+} \left(1 - \frac{L(p)}{p}\right) = \lim_{p\to 0^+} \left(-\frac{Q(p)}{\mu}\right) = -\frac{Q(0)}{\mu}.$$

2.3 The Bonferroni index

The Gini index has been extensively studied in the literature. Among the several ways of defining the Gini index (Yitzhaki, 1998; Xu, 2003), one can write it as function of $L(p)$:

$$Gini = 1 - 2\int_0^1 L(p)dp. \tag{1}$$

Indeed, *Gini* is equal to the area between the 45-degree diagonal segment (line of perfect equality) and the Lorenz curve divided by the whole area under the diagonal.

The Bonferroni index can also be written as function of the Lorenz curve (Pundir et al., 2005):

$$Bonferroni = 1 - \int_0^1 Bon(p)dp, \tag{2}$$

where $Bon(p) = L(p)/p$ is the ordinate of the Bonferroni curve (Figure 2.1). The Bonferroni curve, $[p, L(p)/p]$, is defined on the orthogonal plane within a unit square. It does not always start from the origin of the orthogonal plane because when p goes to 0, $Bon(p)$ takes the form 0/0. Furthermore, it is strictly increasing and it can be convex in some parts and concave in others (Giorgi and Crescenzi, 2001b).

For *Bonferroni*, the line of perfect equality is the line which joins the coordinate points (0,1) and (1,1). The Bonferroni index is equal to the area enclosed by the axis of

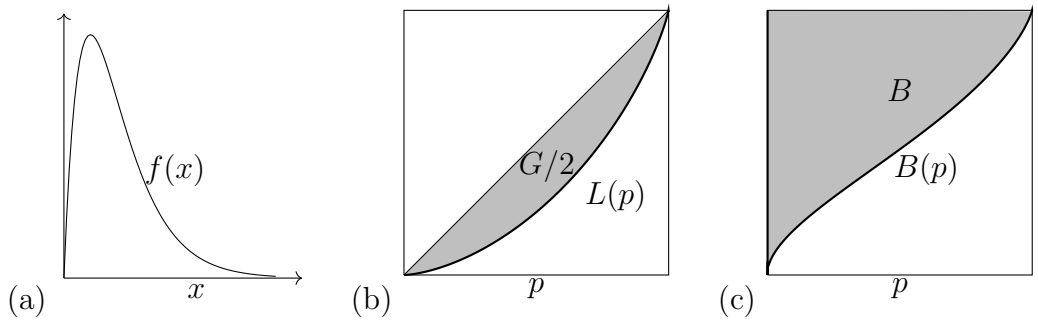


Figure 2.1: (a) Probability density function of a gamma random variable with shape $\alpha = 2$ and rate $\beta = 10$; (b) Lorenz curve, the hatched portion is equal to half of the Gini index; (c) Bonferroni curve, the hatched portion represents the Bonferroni index.

ordinates, the line of perfect equality and the Bonferroni curve. That is, the concentration area coincides with *Bonferroni*.

Except for extreme cases, maximum or null concentration, *Bonferroni* > *Gini* (De Vergottini, 1950). The difference between the two inequality indices becomes more clear by writing (1) and (2) in terms of the relative difference between the total mean and the partial mean, $r(y) = (\mu - \mu(y))/\mu$. That is,

$$Gini = \int_0^\infty r(y) \left[\frac{F(y)}{\int_0^\infty F(y)\ dF(y)} \right] dF(y)$$

and

$$Bonferroni = \int_0^\infty r(y)\ dF(y).$$

Hence, *Gini* is a weighted mean of the $r(y)$'s, while *Bonferroni* is their simple mean (Tarsitano, 1990, p. 230). The difference, *Bonferroni* − *Gini*, increases in the first stretch up to a threshold where the weights in *Gini* are large enough. It then decreases until *Gini* and *Bonferroni* achieve their values (Pizzetti, 1951, p. 302).

2.4 Definition and estimation in a finite population

The inequality measures are usually estimated by means of a sample survey. Consider a population U of size N, which could be individuals, households or enterprises. Let $y_1, \ldots, y_i, \ldots, y_N$ denote the the incomes of the N population units sorted in ascending order. The total, the mean and the partial mean in the finite population are respectively defined by

$$Y = \sum_{i \in U} y_i, \qquad \overline{Y} = \frac{Y}{N}, \qquad \overline{Y}_k = \frac{1}{N_k} \sum_{i \in U} y_i \mathbb{1}[y_i \leq y_k],$$

where

$$N_k = \sum_{i \in U} \mathbb{1}[y_i \leq y_k].$$

The Bonferroni inequality index in the finite population according to its original definition (Bonferroni, 1933, p. 58) is

$$Bonferroni = \frac{1}{(N-1)\,\overline{Y}} \sum_{k \in U} \left(\overline{Y} - \overline{Y}_k\right). \tag{3}$$

As explained in Giorgi (1998), Expression (3) is the discrete analogue of Expression (2) defined in Section 2.3.

The finite-population definition of the Gini index, written in terms of mean difference (Gini, 1914), is

$$Gini = \frac{\sum\limits_{i \in U} \sum\limits_{j \in U} |y_i - y_j|}{2NY}. \tag{4}$$

Consider a random sample of size n selected from U according to a sampling design $p(s) = \Pr(S = s)$, for all $s \in S$. Let $a_1, \ldots, a_i, \ldots, a_N$ be the indicator Bernoulli random variables for the presence of the units in the random sample. Let $\pi_i = \mathrm{E}[a_i]$ denote the first-order inclusion probability of unit i in U and its inverse $d_i = 1/\pi_i$, the Horvitz and Thompson (1952) weight. Let $\pi_{ij} = \mathrm{E}[a_i a_j]$ be the second-order inclusion probability.

A sampling weight w_i is associated to each sampling unit. The w_i could be equal to the inverse of the first-order inclusion probability $d_i = 1/\pi_i$ used in the Horvitz-Thompson estimator. The weights could also be the result of a more complex estimation procedure. For instance, the weights could be obtained through a calibration on known marginal totals (Deville and Särndal, 1992; Särndal, 2007). The weights could also contain a reweighting factor to compensate questionnaire nonresponse (Särndal and Lundström, 2005).

The estimators of the population size, the total, the mean and the partial mean are respectively defined by

$$\widehat{N} = \sum_{i \in U} w_i a_i, \qquad \widehat{Y} = \sum_{i \in U} w_i y_i a_i, \qquad \widehat{\overline{Y}} = \frac{\widehat{Y}}{\widehat{N}}, \qquad \widehat{\overline{Y}}_k = \frac{1}{\widehat{N}_k} \sum_{i \in U} w_i y_i a_i \mathbb{1}[y_i \leq y_k],$$

where

$$\widehat{N}_k = \sum_{i \in U} w_i a_i \mathbb{1}[y_i \leq y_k].$$

For the Bonferroni index, the plug-in estimator of (3) is

$$\widehat{B}_r = \frac{1}{\left(\widehat{N} - 1\right) \widehat{\overline{Y}}} \sum_{k \in U} a_k \left[w_k \left(\widehat{\overline{Y}} - \widehat{\overline{Y}}_k \right) \right], \tag{5}$$

while an alternative estimator (Giorgi and Guandalini, 2013, p. 154) is

$$\widehat{B}_t = \frac{1}{\left(\widehat{N} - 1\right) \widehat{\overline{Y}}} \sum_{k \in U} a_k \left\{ w_k \left(\widehat{\overline{Y}} - \frac{\widehat{\overline{Y}}_k + \widehat{\overline{Y}}_{k-1}}{2} \right) \right\}. \tag{6}$$

Since both estimators are non-linear functions of Horvitz-Thompson estimators, they are slightly biased. The former decomposes the concentration area in rectangles (for this reason, the notation $\widehat{B}_r$ is used, where r refers to *rectangles*), while the latter uses trapezoids (for this reason, the notation $\widehat{B}_t$ is used, where t refers to *trapezoids*) in order to reduce the bias of $\widehat{B}_r$ for samples of small sizes in particular (see simulation results in Section 2.7).

For the Gini index, the plug-in estimator of (4) is

$$\widehat{Gini} = \frac{\sum\limits_{i \in U} \sum\limits_{j \in U} a_i a_j w_i w_j |y_i - y_j|}{2 \widehat{N} \widehat{Y}} \tag{7}$$

(Langel and Tillé, 2013, p. 524).

2.5 Linearisation

In order to address the problem of inference for the Bonferroni index, a linearisation method is proposed. Linearisation includes a range of techniques used to approximate the variance of a non-linear statistic. With the linearisation techniques, a non-linear or a complex statistic, such as *Bonferroni*, is approximated by a sum of terms. The interest lies basically in finding linearised variables $z_{\bullet i}$'s such that

$$\widehat{B}_\bullet - Bonferroni \approx \sum_{i \in S} w_i z_{\bullet i} - \sum_{i \in U} z_{\bullet i},$$

where $\widehat{B}_\bullet$ stands for $\widehat{B}_r$ or $\widehat{B}_t$. The variance of $\widehat{B}_\bullet$ can thereby be simply approximated by the variance of the related total estimator

$$\widehat{Z} = \sum_{i \in S} w_i z_{\bullet i}. \tag{8}$$

In general, linearised variables $z_{\bullet i}$'s must be estimated because they depend on population parameters. They can be estimated from the sample and can be used to construct an estimator of the variance by plugging $\hat{z}_{\bullet i}$'s into the variance estimator of the total estimator under the given sampling design.

There are several ways of deriving linearised variables. For smooth functions of the totals, it is possible to linearise by performing a Taylor series expansion with respect to these totals (Woodruff, 1971). However, for parameters that are not functions of the totals, alternative methods must be used. Most of these parameters can be seen as the solution of an estimating equation. Following the estimating equations methodology developed in Binder (1983; 1991) and Binder and Patak (1994), linearised variables can be derived for estimating the sampling variance.

Deville (1999) proposes the use of influence function, already known in the field of robust statistics (see, e.g., Hampel, 1974; Hampel et al., 1985), as artificial variables for estimating the variance of complex estimators (i.e., calibration type estimators) and non-linear statistics. Furthermore, Demnati and Rao (2004) propose to use the Deville influence function on the estimated measure of mass equal to w_i.

In order to estimate the variances of the estimators presented in Section 2.4, the linearisation method developed by Graf (2011) is used. This method consists of computing the derivatives of the estimator with respect to the a_i indicator variables of the presence of the units in the sample (see also Vallée and Tillé, 2019). It can be applied to almost all sampling designs as long as the expression of the variance estimator of the total estimator under the sampling design is known.

Result 1. *The linearised variable of* $\widehat{B}_r$ *is*

$$\begin{aligned}\hat{z}_{ri} := \left(\frac{\partial \widehat{B}_r}{\partial a_i}\right) / w_i = \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \Bigg\{ & \frac{1}{\widehat{N}}\left(y_i - \widehat{\overline{Y}}\right)\widehat{B}_r - y_i \widehat{B}_r + \left(y_i - \widehat{\overline{Y}}_i\right) \\ & - y_i \sum_{k \in U} \frac{w_k \mathbb{1}[y_i \leq y_k] a_k}{\widehat{N}_k} + \sum_{k \in U} \frac{w_k \mathbb{1}[y_i \leq y_k] \widehat{\overline{Y}}_k a_k}{\widehat{N}_k} \Bigg\}. \end{aligned} \tag{9}$$

Result 2. *The linearised variable of* $\widehat{B}_t$ *is*

$$\hat{z}_{ti} := \left(\frac{\partial \widehat{B}_t}{\partial a_i}\right)/w_i = \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}}\left\{\frac{1}{\widehat{N}}\left(y_i-\widehat{\overline{Y}}\right)\widehat{B}_t - y_i\widehat{B}_t + \left(y_i - \frac{\widehat{\overline{Y}}_i+\widehat{\overline{Y}}_{i-1}}{2}\right)\right.$$
$$- y_i\sum_{k\in U}\frac{w_k a_k}{2}\left(\frac{\mathbb{1}[y_i\le y_k]}{\widehat{N}_k}+\frac{\mathbb{1}[y_i\le y_{k-1}]}{\widehat{N}_{k-1}}\right)$$
$$\left.+\sum_{k\in U}\frac{w_k a_k}{2}\left(\frac{\mathbb{1}[y_i\le y_k]\widehat{\overline{Y}}_k}{\widehat{N}_k}+\frac{\mathbb{1}[y_i\le y_{k-1}]\widehat{\overline{Y}}_{k-1}}{\widehat{N}_{k-1}}\right)\right\}. \tag{10}$$

The proofs of Result 1 and Result 2 are given in the appendix of this chapter.

The linearised variable of $\widehat{Gini}$ can be found in Langel and Tillé (2013, see also references therein):

$$\hat{z}_{Gi} := \left(\frac{\partial \widehat{Gini}}{\partial a_i}\right)/w_i = \frac{1}{\widehat{N}\widehat{Y}}\left\{2\widehat{N}_i\left(y_i-\widehat{\overline{Y}}_i\right)+\widehat{Y}-\widehat{N}y_i-\widehat{Gini}\left(\widehat{Y}+y_i\widehat{N}\right)\right\}. \tag{11}$$

2.6 Variance estimation

2.6.1 Variance estimation formulas

Results 1 and 2 presented in Section 2.5 can be used to approximate the variances of $\hat{B}_r$ and $\hat{B}_t$. In fact, the proposed linearisation method can be easily implemented as long as the expression of the variance estimator of the total estimator under the given sampling design is known.

Consider a general sampling design whose first-order (π_i) and second-order (π_{ij}) inclusion probabilities are all positive, a generalised Horvitz-Thompson formulation for the variance estimator of the total estimator is

$$\widehat{var}(\widehat{Y}) = \sum_{i\in S}\sum_{j\in S}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}. \tag{12}$$

In the set-up when there is no weight adjustment for non-response or calibration, the variance of $\widehat{B}_r$ (resp. $\widehat{B}_t$) can be estimated by simply replacing y_i with $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) in Expression (12) as a general formulation or in the following Expressions (13), (14), (15) or more other expressions of the variance estimator of the total estimator for each specific sampling design.

Under simple random sampling without replacement with fixed sample size ($SRSWOR$), Expression (12) can be written as:

$$^{SRSWOR}\widehat{var}(\widehat{Y}) = N^2 \frac{N-n}{Nn(n-1)} \sum_{i \in S} \left(y_i - \widehat{\overline{Y}} \right)^2 . \tag{13}$$

Notice that for estimating the variance of $\hat{B}_r$ (resp. $\widehat{B}_t$) under $SRSWOR$, $\widehat{\overline{Y}}$ is calculated as the sample mean of the linearised variables $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$).

Under stratified simple random sampling ($StrSRS$), the population is stratified and $SRSWOR$ is used within each stratum. Because the samples are independent from stratum to stratum, the variance estimator of the total estimator is obtained by summing up the strata variances estimated using Expression (13) within each stratum:

$$^{StrSRS}\widehat{var}(\widehat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h n_h (n_h - 1)} \sum_{i \in S_h} \left(y_i - \widehat{\overline{Y}}_h \right)^2 , \tag{14}$$

where h is the label for each stratum ($h = 1, \ldots, H$), and $\widehat{\overline{Y}}_h$ is the sample mean calculated within stratum h:

$$\widehat{\overline{Y}}_h = \frac{1}{n_h} \sum_{i \in S_h} y_i .$$

For estimating the variance of $\hat{B}_r$ (resp. $\widehat{B}_t$) under $StrSRS$, $\widehat{\overline{Y}}_h$ is calculated as the sample mean of the linearised variables $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) within stratum h.

Under stratified two-stage sampling, a double sampling procedure is implemented: one for the primary sampling units (PSUs) and one for the secondary sampling units (SSUs). As PSUs are usually selected within each stratum with probabilities proportional to their sizes, second-order (π_{ij}) inclusion probabilities must be known to obtain an exact variance estimate. This is however generally unfeasible, and therefore, the ultimate cluster approximation (for details, see Kalton, 1979; Wolter, 2007) is used. In practice, it is assumed that PSUs are sampled with replacement even when in reality it is not the case. In such way, the leading contribution to the variance of the total estimator would come from the estimated PSUs totals, and the contribution from the second stage to the variance is disregarded. Within each stratum, the variance formula for cluster sampling selected with probability proportional to size with replacement is used, and by summing

up the estimated strata variances due to independency between strata:

$$\widehat{var}^{Str-Two}\left(\widehat{Y}\right) = \sum_{v \in V} \frac{m_v}{(m_v - 1)} \sum_{\ell=1}^{m_v} \left(\widehat{Y}_{v\ell} - \frac{\widehat{Y}_v}{m_v} \right)^2, \tag{15}$$

where V is the set of strata composed of PSUs, m_v is the number of PSUs selected within stratum v, while $\widehat{\overline{Y}}_{v\ell}$ is the sample total of Y in the ℓ^{th} PSU in stratum v:

$$\widehat{Y}_{v\ell} = \sum_{i \in S_{v\ell}} w_i y_i,$$

and $\widehat{Y}_v$ is the sample total of Y in stratum v:

$$\widehat{Y}_v = \sum_{l=1}^{m_v} \widehat{Y}_{v\ell}.$$

This approximation provides conservative variance estimates with an upward bias that becomes negligible as long as the sampling fractions of PSUs are very small. In order to estimate the variance of $\widehat{B}_r$ (resp. $\widehat{B}_t$) under stratified two-stage sampling, it is also sufficient to substitute $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) for y_i in the variance estimator of the total estimator $\widehat{Y}$, where $\widehat{Y}_v$ is calculated as the sample total of $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) in stratum v.

When the calibration estimator is used, a two-step procedure can be implemented for the variance estimation based on the proposed linearised variables. In the first step, the linearised variable $\hat{z}_{\bullet i}$ would be calculated as if we were dealing with the Horvitz-Thompson estimator. In the second step, residual of this linearised variable is computed by performing a regression of this linearised variable on the calibration variables:

$$\check{\hat{z}}_{\bullet i} = \hat{z}_{\bullet i} - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_z,$$

where $\mathbf{x}_i$ is the vector of the auxiliary variables on which the sample is calibrated and $\widehat{\boldsymbol{\beta}}_z$ is the regression coefficients estimated by

$$\widehat{\boldsymbol{\beta}}_z := \left(\sum_{i \in S} \frac{q_i \mathbf{x}_i \mathbf{x}_i^\top}{\pi_i} \right)^{-1} \sum_{i \in S} \frac{q_i \mathbf{x}_i \hat{z}_{\bullet i}}{\pi_i}.$$

The parameters q_i's make it possible to take into account the potential heteroscedasticity problem.

The linearised variance estimator of $\widehat{B}_r$ (resp. $\widehat{B}_t$) in the case of calibration can be obtained by substituting $\check{\hat{z}}_{ri}$ (resp. $\check{\hat{z}}_{ti}$) for y_i in Expression (12) as a general formulation

or more other expressions of the variance estimator of the total estimator for each specific sampling design. This procedure and some further knowledge on how to estimate sampling variance in the presence of non-response is, for example, described in Vallée and Tillé (2019).

2.6.2 Variance estimation through R packages

The proposed methodology can be easily implemented for almost all sampling designs using the R packages on survey estimation available online, such as `survey` (Lumley, 2011) and `ReGenesees` (Zardetto, 2015). Both packages are devoted to sampling estimation.

For this research, `ReGenesees`[2] is used. The variance estimation of $\hat{B}_r$ (resp. $\hat{B}_t$) can be carried out in three steps. The first step consists in computing $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) from Expression (9) (resp. Expression (10)). In the second step, through the function `e.svydesign`, the sampling design adopted must be declared by identifying the variables (i.e., elementary units, primary sampling units, secondary sampling units, stratification variables, Horvitz-Thompson sampling weights and self-representative strata) in the sample data. If calibration is concerned and if the calibration weights have not yet been calculated, the functions `pop.template` and `e.calibrate` should be used for defining the calibration totals and the calibration model. However, if the Horvitz-Thompson sampling weights have already been calibrated through other statistical softwares, instead of using the function `e.svydesign` at the beginning, the `ext.calibrated` function should be used for identifying the sampling design, the calibrated weights and the calibration model. The third step consists in estimating the variance of the total estimator of the linearised variables $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) from Expression (8) using the function `svystatTM`. If calibration is concerned, the `svystatTM` function automatically computes the residuals of the linearised variables, and therefore, besides the computation of $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$), no further computation is needed.

[2]The package and the user manual are available online at https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/elaborazione/strumenti-di-elaborazione/regenesees. Last access, 31 January 2021.

2.7 Simulation studies

In Section 2.7, the empirical demonstrations on the consistency of the Bonferroni index estimators and the accuracy of their variance estimation using the proposed linearisation technique are carried out on a series of simulations based on the real 2015 IT-SILC data (Istat, 2015).

2.7.1 IT-SILC data

The IT-SILC belongs to the framework of surveys yearly carried out by European countries according to the European Regulation n. 1177/2003 for providing data on income, poverty, social exclusion and living conditions. The 2015 IT-SILC implements a stratified two-stage sampling design. The municipalities are the PSUs and the households are the SSUs. Municipalities are stratified according to their sizes and are split into self-representative (SR) municipalities and non self-representative (NSR) municipalities. The former are included directly in the sample, and therefore, there is only one single stage of selection for the households. The latter are selected with probabilities proportional to their sizes, and thus, there are two stages of selection, one for the municipalities and one for the households. In both cases, households are selected under a systematic sampling design in each selected municipality. All individuals older than 16 years old inside each selected household are surveyed. The sample size is determined for constructing reliable estimates of the main parameters both at the regional and at the national level. The Italian sample in 2015 consists of 615 PSUs, in which there are 116 SR municipalities and 499 NSR municipalities. Finally, a total number of 17985 households for 49987 individuals are selected.

The income considered is the household gross income including imputed rents but excluding social contribution. In the simulation, the income distribution has been reconstructed by duplicating household incomes with respect to sampling weights. The reconstructed income distribution consists of N=25,775,872 households in Italy in 2015. The income distribution is heavily right-skewed as the mean is equal to 42223 €, which is much larger than the median (=34196 €).

The Bonferroni index is estimated by its two estimators introduced in Section 2.4,

namely $\widehat{B}_r$ from Expression (5) and $\widehat{B}_t$ from Expression (6). The variances of the two estimators are approximated using the aforementioned `ReGenesees` package in `R` and the confidence intervals are constructed using Expression (16) in Section 2.7.2. Both estimators, $\widehat{B}_r$ and $\widehat{B}_t$, yield the same result with the same length of confidence interval. The estimate of the Bonferroni index is equal to 0.490 with a 95% confidence interval $[0.484, 0.495]$. Similarly, the estimate of the Gini index is computed using the estimator $\widehat{Gini}$ from Expression (7). Its variance and the associated confidence interval are also estimated. The estimate of the Gini index is equal to 0.365 with a 95% confidence interval $[0.358, 0.371]$. The Bonferroni index, as demonstrated in Section 2.3, assumes by definition a higher value than the Gini index.

2.7.2 Simulation scheme and results

Simulation studies are performed based on the 2015 IT-SILC survey data. The obtained estimate of the Bonferroni index presented in the previous section is treated as the true value, *Bonferroni*.

SRSWOR and StrSRS are designed for the simulation studies. For SRSWOR, the population of households has been reconstructed by duplicating the original sample based on the calibration survey weights. For StrSRS, a pseudo population is constructed using the multivariate hypergeometric distribution, i.e., within one stratum, each calibration survey weight of the household income is viewed as a sub-population from which a sub-sample is drawn and the allocated sample size in that stratum (proportional allocation with a minimum of two sampling units guaranteed in each stratum) is the total number of draws. The multivariate hypergeometric distribution is applied to every stratum and the whole sample can be retrieved by aggregating the samples inside all strata. The strata are determined according to the regions of Italy. There are 21 strata in total and they are the Italian region Piedmont, Aosta Valley, Lombardy, Veneto, Friuli Venezia Giulia, Liguria, Emilia-Romagna, Tuscany, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria, Sicily and Sardinia, plus Province of Bolzano and Province of Trento from the Trentino-Alto Adige/Südtirol region. Samples with a large range of sample sizes are considered, i.e., n=100, 500, 1000, 2000, 5000 and 10000. The sample selection is repeated R=10000 times for each sample size.

Firstly, behaviors of the two Bonferroni index estimators are investigated. The empirical RB and NRMSE of $\widehat{B}_r$ and $\widehat{B}_t$ are computed for different sample sizes under SRSWOR and StrSRS. The empirical RB of $\widehat{B}_r$ is defined as

$$RB_{sim}(\widehat{B}_r) = \frac{\frac{1}{R}\sum_{i=1}^{R} \widehat{B}_r^{(i)} - Bonferroni}{Bonferroni},$$

while the empirical NRMSE of $\widehat{B}_r$ is

$$NRMSE_{sim}(\widehat{B}_r) = \frac{\sqrt{\frac{1}{R}\sum_{i=1}^{R}\left(\widehat{B}_r^{(i)} - Bonferroni\right)^2}}{\frac{1}{R}\sum_{i=1}^{R}\widehat{B}_r^{(i)}}.$$

The empirical RB and the empirical NRMSE of $\widehat{B}_t$ are similar by simply replacing $\widehat{B}_r^{(i)}$ in the previous two expressions with $\widehat{B}_t^{(i)}$.

Table 1: RB and NRMSE of $\widehat{B}_r$ and $\widehat{B}_t$ for different sample sizes under SRSWOR

	sample size					
	100	500	1000	2000	5000	10000
RB(%)						
$\widehat{B}_r$	-2.059	-0.478	-0.203	-0.105	-0.029	-0.024
$\widehat{B}_t$	-1.265	-0.229	-0.085	-0.054	-0.016	0.001
NRMSE(%)						
$\widehat{B}_r$	6.872	2.941	2.085	1.461	0.922	0.656
$\widehat{B}_t$	6.558	2.928	2.051	1.469	0.935	0.660

The results are presented in Table 1 and 2. They show that both estimators slightly underestimate the Bonferroni index when $n = 100$. They are both asymptotically approximately unbiased as the relative biases of the two estimators are merely negligible when n reaches 10000. Among the two estimators, $\widehat{B}_t$ is much less biased than $\widehat{B}_r$. Under SRSWOR, the relative bias of $\widehat{B}_t$ is around one half of the relative bias of $\widehat{B}_r$ when $n \leq 5000$, and the bias decreases drastically when $n = 10000$. Under StrSRS, the relative bias of $\widehat{B}_t$ is around one half of the relative bias of $\widehat{B}_r$ when $n \leq 2000$, and the bias drops massively when n reaches 5000. In terms of NRMSE, $\widehat{B}_t$ also performs better than

Table 2: RB and NRMSE of $\widehat{B}_r$ and $\widehat{B}_t$ for different sample sizes under StrSRS

	sample size					
	100	500	1000	2000	5000	10000
RB(%)						
$\widehat{B}_r$	-2.377	-0.440	-0.216	-0.111	-0.058	-0.006
$\widehat{B}_t$	-1.215	-0.245	-0.129	-0.050	-0.008	0.001
NRMSE(%)						
$\widehat{B}_r$	7.055	2.952	2.074	1.453	0.922	0.654
$\widehat{B}_t$	6.752	2.942	2.075	1.465	0.926	0.655

$\widehat{B}_r$ for small sample sizes, i.e., $n \leq 1000$ under SRSWOR and $n \leq 500$ under StrSRS. For large sample sizes, it seems that $\widehat{B}_r$ outperforms $\widehat{B}_t$ regarding NRMSE; however, the differences between the two estimators are negligible in those cases. The relative bias (in absolute value) and NRMSE of both estimators decrease when the sample size increases.

Furthermore, the performance of the variance estimator $\widehat{var}_{lin}(\widehat{B}_r)$ (resp. $\widehat{var}_{lin}(\widehat{B}_t)$) constructed using the linearised variable $\hat{z}_{ri}$ (resp. $\hat{z}_{ti}$) computed from Expression (9) (resp. (10)) is assessed. Table 3 and 4 present the variance estimates of $\widehat{B}_r$ and $\widehat{B}_t$, enlarged by 10000 times, obtained through linearisation and the Monte Carlo method for different sample sizes under SRSWOR and StrSRS. For a specific sample size n, given that a variance estimate through linearisation is obtained for each selected sample, $E_{sim}\widehat{var}_{lin}(\widehat{B}_r)$ is computed by averaging the 10000 replicates of $\widehat{var}_{lin}(\widehat{B}_r)$, while

$$var_{sim}(\widehat{B}_r) = \frac{1}{R-1}\sum_{i=1}^{R}\left(\widehat{B}_r^{(i)} - \overline{\widehat{B}_r}\right)^2,$$

where

$$\overline{\widehat{B}_r} = \frac{1}{R}\sum_{i=1}^{R}\widehat{B}_r^{(i)}.$$

Similarly, $E_{sim}\widehat{var}_{lin}(\widehat{B}_t)$ and $var_{sim}(\widehat{B}_t)$ are computed. Since the variance estimates obtained through Monte Carlo experiments, namely $var_{sim}(\widehat{B}_r)$ (resp. $var_{sim}(\widehat{B}_t)$), approximate the true variance of $\widehat{B}_r$ (resp. $\widehat{B}_t$) for each sample size, they can be considered as benchmarks.

As shown in Table 3 and 4, the variance estimators constructed using the linearisation method are approximately unbiased, at least asymptotically. Underestimation of the

Table 3: Variance of $\widehat{B}_r$ and $\widehat{B}_t$ based on linearisation and the Monte Carlo method for different sample sizes (results enlarged by 10000 times) under SRSWOR

	sample size					
	100	500	1000	2000	5000	10000
$E_{sim}\widehat{var}_{lin}(\widehat{B}_r)$	8.638	1.983	1.010	0.511	0.207	0.103
$var_{sim}(\widehat{B}_r)$	9.839	1.999	1.028	0.508	0.203	0.103
$E_{sim}\widehat{var}_{lin}(\widehat{B}_t)$	8.872	1.997	1.021	0.514	0.206	0.103
$var_{sim}(\widehat{B}_t)$	9.664	2.032	1.005	0.516	0.209	0.104

Table 4: Variance of $\widehat{B}_r$ and $\widehat{B}_t$ based on linearisation and the Monte Carlo method for different sample sizes (results enlarged by 10000 times) under StrSRS

	sample size					
	100	500	1000	2000	5000	10000
$E_{sim}\widehat{var}_{lin}(\widehat{B}_r)$	9.062	1.973	1.004	0.506	0.204	0.102
$var_{sim}(\widehat{B}_r)$	10.015	2.024	1.015	0.502	0.203	0.102
$E_{sim}\widehat{var}_{lin}(\widehat{B}_t)$	9.130	1.990	1.005	0.508	0.204	0.102
$var_{sim}(\widehat{B}_t)$	10.307	2.050	1.025	0.513	0.205	0.103

variances of both $\widehat{B}_r$ and $\widehat{B}_t$ is observed when the sample sizes are small and the bias could be large. For example, when n=100, the relative bias of the variance estimators could reach 8-12% under SRSWOR and 9%-11% under StrSRS. However, once the sample size becomes sufficiently large, that is, when n reaches 1000, the variance estimates obtained using the linearisation method are almost equal to the Monte Carlo variances for both sampling designs.

The linearised variance estimators can be utilised to construct empirical confidence intervals. By assuming standard normal deviates, the empirical 95% confidence interval of $\widehat{B}_r$ is

$$\left[\widehat{B}_r - 1.96 \cdot \sqrt{\widehat{var}_{lin}(\widehat{B}_r)},\ \widehat{B}_r + 1.96 \cdot \sqrt{\widehat{var}_{lin}(\widehat{B}_r)} \right]. \tag{16}$$

The empirical 95% confidence interval of $\widehat{B}_t$ is similar by substituting $\widehat{B}_t$ for $\widehat{B}_r$ in Expression (16).

The empirical coverage rate is the proportion of the successful coverage of the true

value of the Bonferroni index (*Bonferroni*) within the 10000 empirical confidence intervals for each sample size. In Table 5 and 6, the empirical coverage rates of the 95% confidence intervals for both $\widehat{B}_r$ and $\widehat{B}_t$ under SRSWOR and StrSRS are presented. The empirical coverage rate rises with the increase of the sample size. The true coverage rates of *Bonferroni* are mostly slightly underestimated, which could be a consequence of the underestimation of *Bonferroni* inherited from the two estimators $\widehat{B}_r$ and $\widehat{B}_t$. The underestimation of the variance estimators and the potential failure of the normality assumption could also be factors which cause the undercoverage for small sample sizes. However, when the sample size reaches 5000, there seems no significant differences with the true 95% coverage rates in terms of a t-test at a significance level of 5% for both sampling designs. An approach recently proposed by Berger and Gedik Balay (2020) is devoted to the inference for the Gini index using the empirical likelihood method for overcoming the normality assumption. The transposition of this method to the Bonferroni index could be a promising alternative method to compute the confidence interval.

Table 5: Empirical coverage rates of 95% confidence intervals for $\widehat{B}_r$ and $\widehat{B}_t$ under SRSWOR

	sample size					
	100	500	1000	2000	5000	10000
$\widehat{B}_r$	0.889	0.934	0.942	0.944	0.949	0.950
$\widehat{B}_t$	0.917	0.940	0.947	0.945	0.946	0.948

Table 6: Empirical coverage rates of 95% confidence intervals for $\widehat{B}_r$ and $\widehat{B}_t$ under StrSRS

	sample size					
	100	500	1000	2000	5000	10000
$\widehat{B}_r$	0.886	0.931	0.939	0.944	0.948	0.951
$\widehat{B}_t$	0.908	0.938	0.939	0.944	0.948	0.950

2.8 Sensitivity to different levels of distribution

2.8.1 Influence function and linearised variable

Influence function (Hampel, 1974; Hampel et al., 1985) is a statistical tool used for deriving asymptotic variances and investigating local robustness properties. It is essentially the first derivative of an estimator viewed as functional. Let $T(F)$ be a functional. The influence function is

$$IF_x(T) := \lim_{\varepsilon \to 0^+} \frac{T(\varepsilon\delta_x + (1-\varepsilon)F) - T(F)}{\varepsilon},$$

where δ_x denotes the pointmass 1 at x. In the research of income inequality measures, $x \in R_{\geq 0}$ as it is a variable representing the point of income which receives an infinitesimal perturbation. Thus, influence function of an inequality measure shows the effect of an infinitesimal perturbation at the point of income x on the inequality measure. The linearised variable obtained with the Graf method in Section 2.5 could be interpreted as the discrete analogue of the influence function.

The influence function of the Lorenz curve (see among others Essama-Nssah and Lambert, 2011) is

$$IF_x\{L(p)\} = \frac{1}{\mu}\left\{x\mathbb{1}(x \leq Q(p)) + Q(p)\left[p - \mathbb{1}(x \leq Q(p))\right]\right\} - L(p)\frac{x}{\mu}.$$

Monti (1991) has shown that the influence function of the Gini index is

$$IF_x(Gini) = \frac{1}{\mu}\left\{2F(x)\left[x - \mu(x)\right] + \mu - x - (\mu + x)\,Gini\right\} \tag{17}$$

(see also Langel and Tillé, 2013, for a synthesis). It is important to point out that, up to a multiplicative factor, the linearised variable of $\widehat{Gini}$ from Expression (11) in Section 2.5 is the discrete analogue of the influence function of the Gini index given in (17).

The first derivative of (17) is

$$\begin{aligned}
\frac{\partial IF_x(Gini)}{\partial x} &= \frac{1}{\mu}\left[2f(x)x + 2F(x) - 2f(x)\mu(x) - 2F(x)\frac{\partial \mu(x)}{\partial x} - 1 - Gini\right] \\
&= \frac{2}{\mu}f(x)x + \frac{2}{\mu}F(x) - \frac{2}{\mu}xf(x) - \frac{1}{\mu} - Gini\frac{1}{\mu} \\
&= \frac{1}{\mu}\left[2F(x) - 1 - Gini\right],
\end{aligned}$$

since $\frac{\partial \mu(x)}{\partial x} = \frac{xf(x) - \mu(x)f(x)}{F(x)}$. The influence function of the Gini index is a convex function as the second derivative of (17) is positive:

$$\frac{\partial^2 IF_x(Gini)}{\partial x^2} = \frac{2f(x)}{\mu} > 0$$

(see also Monti, 1991, p. 566).

Result 3. *The influence function of the Bonferroni index is*

$$\begin{aligned} IF_x(Bonferroni) &= \frac{x - \mu(x)}{\mu} - \frac{x}{\mu}\int_0^\infty \frac{\mathbb{1}(x \leq y)}{F(y)} dF(y) + \frac{1}{\mu}\int_0^\infty \frac{\mu(y)\mathbb{1}(x \leq y)}{F(y)} dF(y) \\ &\quad - \frac{x Bonferroni}{\mu} \\ &= \frac{x \log F(x)}{\mu} + \frac{x - \mu}{\mu} + \frac{1}{\mu}\int_0^\infty \frac{y\mathbb{1}(x \leq y)}{F(y)} dF(y) - Bonferroni\frac{x}{\mu}. \end{aligned} \tag{18}$$

The proof is given in the appendix of this chapter.

Notice that, up to a multiplicative factor, the linearised variable of the plug-in estimator of the Bonferroni index $\hat{B}_r$ from Expression (9) is the discrete analogue of (18).

The first derivative of (18) is

$$\begin{aligned} \frac{\partial IF_x(Bonferroni)}{\partial x} &= \frac{1}{\mu}\left[\log F(x) + x\frac{1}{F(x)}f(x)\right] + \frac{1}{\mu} - \frac{1}{\mu}\frac{x}{F(x)}f(x) - \frac{Bonferroni}{\mu} \\ &= \frac{1}{\mu}\log F(x) + \frac{1}{\mu} - \frac{Bonferroni}{\mu}. \end{aligned}$$

The influence function of the Bonferroni index is also a convex function as the second derivative of (18) is positive:

$$\frac{\partial^2 IF_x(Bonferroni)}{\partial x^2} = \frac{f(x)}{\mu F(x)} > 0.$$

The comparison between the second derivatives of the influence functions of the Bonferroni and Gini indices shows that

$$\begin{cases} \dfrac{\partial^2 IF_x(Gini)}{\partial x^2} < \dfrac{\partial^2 IF_x(Bonferroni)}{\partial x^2} & \text{if} \quad x < Q(1/2) \\ \dfrac{\partial^2 IF_x(Gini)}{\partial x^2} \geq \dfrac{\partial^2 IF_x(Bonferroni)}{\partial x^2} & \text{if} \quad x \geq Q(1/2). \end{cases}$$

On the one hand, the curvature of the influence function of the Bonferroni index is stronger for the small x values than that of the Gini index. On the other hand, the curvature of the influence function of the Gini index is stronger for the large x values than that of the Bonferonni index. Furthermore, the curvature of the influence function of the Bonferroni index at the smallest x values is of size $1/2F(x)$ greater than that of the Gini index. That is, asymptotically, for a well-behaved income distribution, the curvature of the influence function of the Bonferroni index at the smallest x values can be much stronger than that of the Gini index. This result thus clearly confirms that the Bonferroni index is more influenced by the smallest incomes than the Gini index.

2.8.2 Empirical demonstration

The significance of the relation between the linearised variables in Section 2.5 and the influence functions is two-fold. Deville (1999) proposes a generalised linearisation method based on the concept of influence function that under mild conditions provides an approximately unbiased variance estimation for non-linear statistics, for instance, $\widehat{B}_\bullet$ and $\widehat{Gini}$ in this chapter. Viewing the linearised variables derived with the Graf method as discrete analogues of the influence functions, the simulation results in Section 2.7 provides a further demonstration of the use of influence function for variance estimation. In such view, the linearised variables in Section 2.5 could also be utilised for understanding and comparing the sensitivity of the Bonferroni and Gini indices to different levels of the income distribution.

In Figure 2.2, the values of $\hat{z}_{ri}$'s (i.e., the linearised variables of $\widehat{B}_r$) on each observed equivalised income estimated from a sample of 100 households selected under SRSWOR from the reconstructed population of the Italian households based on the real IT-SILC data are compared with those of $\hat{z}_{Gi}$'s (i.e., the linearised variables of $\widehat{Gini}$). The values of $\hat{z}_{ti}$'s (i.e., the linearised variables of $\widehat{B}_t$) are almost identical with those of $\hat{z}_{ri}$'s, and therefore, only $\widehat{B}_r$ is represented. The contributions of the different observations to the inequality indices can thus be directly appreciated.

As shown in Figure 2.2, in the beginning, the curvature of the influence function of $\widehat{B}_r$ is stronger than that of $\widehat{Gini}$. This is particularly true for the smallest incomes. In the middle part of the distribution, the observations contribute approximately the same

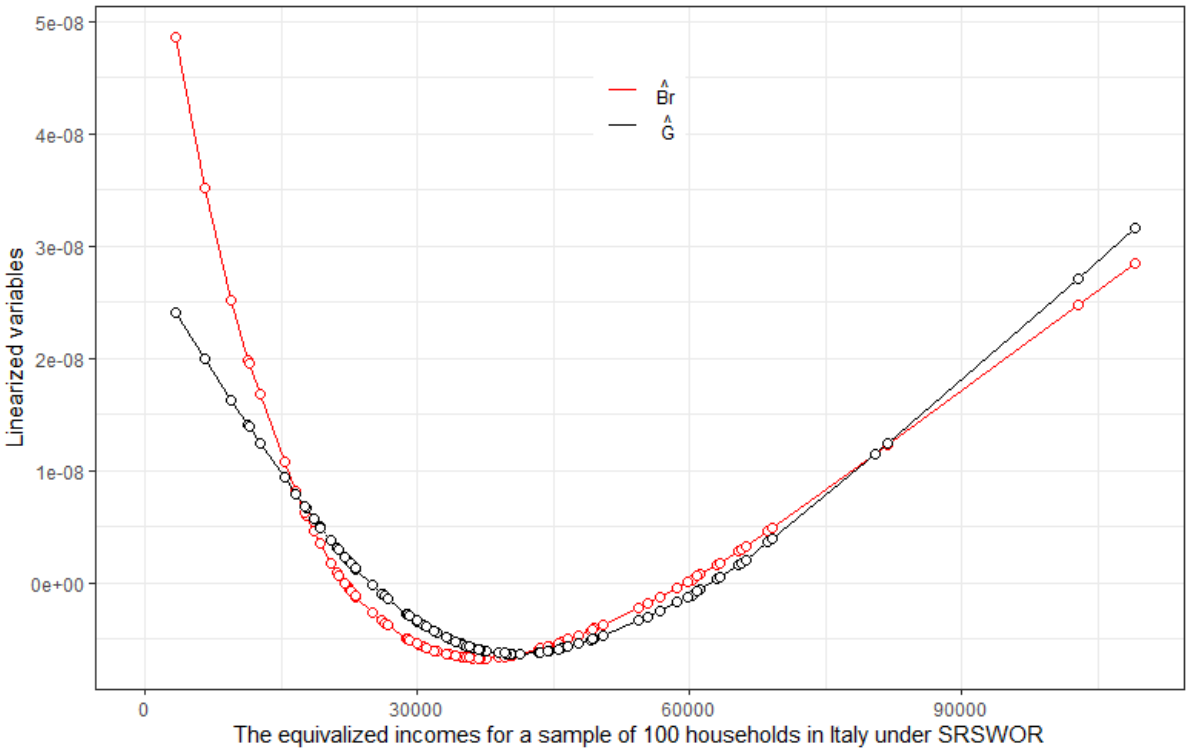


Figure 2.2: Influence of incomes on $\widehat{B}_r$ and $\widehat{Gini}$.

amount to the indices. Finally, in the right tail of the distribution, the curvature of the influence function of $\widehat{Gini}$ is stronger than that of $\widehat{B}_r$. Thus, the observations at the very beginning which are composed of households with the smallest incomes are more "influential" for $\widehat{B}_r$ than for $\widehat{Gini}$, while the observations in the tail part of the distribution play a relatively more important role in $\widehat{Gini}$.

Figure 2.2, as a typical example, confirms the theoretical results obtained in the previous section. They stress the different sensitivity of the Bonferroni and Gini indices to different levels of the income distribution. It demonstrates, from a different and new perspective, the evidence stated by De Vergottini (1950) and Pizzetti (1951). Furthermore, it serves to relaunch the Bonferroni index as a complementary and not only as an alternative inequality measure to the Gini index for studying income inequality.

2.9 Some concluding remarks

Researchers increasingly advocate the use of further inequality measures besides the Gini index. Their use, simultaneously with the evergreen Gini index, can help to better catch

the inequality that lurks in different parts of the wealth distribution. Chapter 2 focuses on the Bonferroni inequality index. Because income data are usually collected through sample surveys, attention has been paid to its inferential aspects.

As the main aim, the linearised variables for the plug-in estimator and an alternative sampling estimator of the Bonferroni index have been computed. The Graf method has been used, that is, the estimators have been linearised by differentiating them with respect to their sample indicators. A Monte Carlo simulation carried out on real income data demonstrates that both estimators are approximately unbiased, at least asymptotically, and the proposed linearisation method provides a valid inference on the variances of the estimators. Furthermore, interesting results have been obtained by interpreting the linearised variable as the influence function. In this sense, the influence function provides a sensitivity measure of the inequality index to different levels of the income distribution. In particular, the curvature of the influence function of the Bonferroni index are compared with that of the Gini index. An example on a sample of real income data has been illustrated. In this chapter, it has been demonstrated, from a different and new perspective, that the Bonferroni index is more sensitive to the lowest incomes in the distribution than the Gini index. Hence, the Bonferroni inequality index is relaunched not only as an alternative but also as a complement to the Gini index.

2.10 Appendix

Proof of Result 1

First, denote u_i and $u_{i,k}$ as the derivatives of $\widehat{\overline{Y}}$ and $\widehat{\overline{Y}}_k$ with respect to a_i:

$$u_i = \frac{\partial \widehat{\overline{Y}}}{\partial a_i} = \frac{w_i y_i \left(\sum_{j \in U} w_j a_j \right) - \left(\sum_{j \in U} w_j y_j a_j \right) w_i}{\left(\sum_{j \in U} w_j a_j \right)^2} = \frac{w_i}{\widehat{N}} \left(y_i - \widehat{\overline{Y}} \right) \tag{19}$$

and

$$u_{i,k} = \frac{\partial \widehat{\overline{Y}}_k}{\partial a_i} = \frac{w_i y_i \mathbb{1}[y_i \leq y_k] - \widehat{\overline{Y}}_k w_i \mathbb{1}[y_i \leq y_k]}{\sum_{i \in U} w_i \mathbb{1}[y_i \leq y_k] a_i} = \frac{w_i \mathbb{1}[y_i \leq y_k]}{\widehat{N}_k} \left(y_i - \widehat{\overline{Y}}_k \right). \tag{20}$$

One can write $\widehat{B}_r = A \times P$, where $A = \mathbb{1}/[(\widehat{N}-1)\widehat{\overline{Y}}]$ and $P = \sum_{k \in U} a_k \left[w_k \left(\widehat{\overline{Y}} - \widehat{\overline{Y}}_k \right) \right]$.

It is then possible to calculate

$$\frac{\partial \widehat{B}_r}{\partial a_i} = P \frac{\partial A}{\partial a_i} + A \frac{\partial P}{\partial a_i}. \tag{21}$$

As

$$\frac{\partial A}{\partial a_i} = -\frac{\widehat{\overline{Y}} \frac{\partial \widehat{N}}{\partial a_i} + (\widehat{N}-1) \frac{\partial \widehat{\overline{Y}}}{\partial a_i}}{[(\widehat{N}-1)\widehat{\overline{Y}}]^2} = -\frac{\frac{\partial}{\partial a_i}\left(\widehat{\overline{Y}}\widehat{N}\right) - u_i}{[(\widehat{N}-1)\widehat{\overline{Y}}]^2} = \frac{u_i - w_i y_i}{[(\widehat{N}-1)\widehat{\overline{Y}}]^2} \tag{22}$$

and

$$\frac{\partial P}{\partial a_i} = \sum_{k \in U} \left[w_k \left(u_i - u_{i,k} \right) a_k \right] + w_i \left(\widehat{\overline{Y}} - \widehat{\overline{Y}}_i \right), \tag{23}$$

by plugging (22) and (23) into (21):

$$\begin{aligned} \frac{\partial \widehat{B}_r}{\partial a_i} = & \frac{u_i - w_i y_i}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \widehat{B}_r + \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \sum_{k \in U} \left[w_k \left(u_i - u_{i,k} \right) a_k \right] \\ & + \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times w_i \left(\widehat{\overline{Y}} - \widehat{\overline{Y}}_i \right). \end{aligned}$$

Through further simplification:

$$\frac{\partial \widehat{B}_r}{\partial a_i} = \frac{u_i - w_i y_i}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \widehat{B}_r + \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times w_i \left(y_i - \widehat{\overline{Y}}_i \right) - \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \sum_{k \in U} w_k u_{i,k} a_k. \tag{24}$$

After replacing u_i and $u_{i,k}$ with Expression (19) and (20) and dividing Expression (24) by w_i, Result 1 is obtained.

□

Proof of Result 2

Similar to the proof of Result 1, one can write $\widehat{B}_t = A \times R$, where $A = \mathbb{1}/[(\widehat{N}-1)\widehat{\overline{Y}}]$ and $R = \sum_{k \in U} a_k \left\{ w_k \left(\widehat{\overline{Y}} - \frac{\widehat{\overline{Y}}_k + \widehat{\overline{Y}}_{k-1}}{2} \right) \right\}$.

It is then possible to calculate

$$\frac{\partial \widehat{B}_t}{\partial a_i} = R \frac{\partial A}{\partial a_i} + A \frac{\partial R}{\partial a_i}. \tag{25}$$

As

$$\frac{\partial R}{\partial a_i} = \sum_{k \in U} \left[w_k \left(u_i - \frac{u_{i,k} + u_{i,k-1}}{2} \right) a_k \right] + w_i \left(\widehat{\overline{Y}} - \frac{\widehat{\overline{Y}}_i + \widehat{\overline{Y}}_{i-1}}{2} \right), \tag{26}$$

by plugging (22) and (26) into (25):

$$\frac{\partial \widehat{B}_t}{\partial a_i} = \frac{u_i - w_i y_i}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \widehat{B}_t + \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \sum_{k \in U} \left\{ w_k \left(u_i - \frac{u_{i,k} + u_{i,k-1}}{2} \right) a_k \right\}$$
$$+ \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times w_i \left(\widehat{\overline{Y}} - \frac{\widehat{\overline{Y}}_i + \widehat{\overline{Y}}_{i-1}}{2} \right).$$

Through further simplification:

$$\frac{\partial \widehat{B}_t}{\partial a_i} = \frac{u_i - w_i y_i}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \widehat{B}_t + \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times w_i \left(y_i - \frac{\widehat{\overline{Y}}_i + \widehat{\overline{Y}}_{i-1}}{2} \right)$$
$$- \frac{1}{\left(\widehat{N}-1\right)\widehat{\overline{Y}}} \times \sum_{k \in U} w_k \left(\frac{u_{i,k} + u_{i,k-1}}{2} \right) a_k. \tag{27}$$

After replacing u_i and $u_{i,k}$ with Expression (19) and (20) and dividing Expression (27) by w_i, Result 2 is obtained.

□

Proof of Result 3

$$\begin{aligned}
IF_x(Bonferroni) &= IF_x\left\{1 - \frac{1}{\mu}\int_0^\infty \mu(y)dF(y)\right\} = -IF_x\left\{\frac{1}{\mu}\int_0^\infty \mu(y)dF(y)\right\} \\
&= -\frac{1}{\mu}\left\{\mu(x) - \int_0^\infty \mu(y)dF(y)\right\} - \frac{1}{\mu}\int_0^\infty IF_x\{\mu(y)\}dF(y) + \frac{1}{\mu^2}\int_0^\infty \mu(y)dF(y)(x-\mu) \\
&= -\frac{\mu(x)}{\mu} + (1 - Bonferroni) - \frac{1}{\mu}\int_0^\infty \frac{(x-\mu(y))\mathbb{1}(x \le y)}{F(y)}dF(y) \\
&\quad + \frac{1}{\mu}(1 - Bonferroni)(x-\mu) \\
&= -\frac{\mu(x)}{\mu} - (Bonferroni - 1) - \frac{x}{\mu}\int_0^\infty \frac{\mathbb{1}(x \le y)}{F(y)}dF(y) + \frac{1}{\mu}\int_0^\infty \frac{\mu(y)\mathbb{1}(x \le y)}{F(y)}dF(y) \\
&\quad + \frac{1}{\mu}(1 - Bonferroni)x \\
&= \frac{x - \mu(x)}{\mu} + \frac{x}{\mu}\log F(x) + \frac{1}{\mu}\int_0^\infty \frac{\mu(y)\mathbb{1}(x \le y)}{F(y)}dF(y) - \frac{x Bonferroni}{\mu}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\int_0^\infty \frac{\mu(y)\mathbb{1}(x \le y)}{F(y)}dF(y) = \int_x^\infty \frac{\int_0^y t dF(t)}{F^2(y)}dF(y) \\
&= \int_0^\infty \int_{\max(x,t)}^\infty \frac{1}{F^2(y)}dF(y)tdF(t) = \int_0^\infty -\left\{1 - \frac{1}{F(\max(x,t))}\right\}tdF(t)
\end{aligned}$$

$$
\begin{aligned}
&= -\mu + \int_0^\infty \frac{1}{F(\max(x,t))} t dF(t) = -\mu + \int_0^x \frac{t}{F(x)} dF(t) + \int_x^\infty \frac{t}{F(t)} dF(t) \\
&= -\mu + \mu(x) + \int_x^\infty \frac{t}{F(t)} dF(t).
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
IF_x(Bonferroni) &= \frac{x-\mu(x)}{\mu} + \frac{x}{\mu}\log F(x) + \frac{-\mu + \mu(x) + \int_x^\infty \frac{t}{F(t)} dF(t)}{\mu} - \frac{xBonferroni}{\mu} \\
&= \frac{x-\mu}{\mu} + \frac{x}{\mu}\log F(x) + \frac{1}{\mu}\int_0^\infty \frac{y\mathbb{1}(x \le y)}{F(y)} dF(y) - \frac{xBonferroni}{\mu}.
\end{aligned}
$$

3 Generalised Income Inequality Index

Abstract

In Chapter 3, a deep generalisation for income inequality indices is proposed. A generalised income inequality index that depends on two parameters and that involves a large set of income inequality indices in the same framework is studied. The two parameters control the sensitivity of the generalised index to different levels of the income distribution. A thorough investigation of the generalised index paves the way for understanding the influence of the low, middle and high incomes on various income inequality indices and thereby facilitates the choice of multiple indices simultaneously for a better analysis of inequality as advocated by several recent studies. Moreover, three methods, the rectangular rule, the trapezoidal rule and a reformulation method of defining and estimating the income inequality indices in finite populations are proposed. [3]

Keywords: Bonferroni, De Vergottini, Gini, Mehran, Piesch, Pietra, inequality indices.

3.1 Introduction

Chapter 3 defines a generalised income inequality index that can be seen as the result of the continuation of the work by Nygård and Sandström (1981, 1985, 1989) and Sandström et al. (1985, 1988). Indeed, Nygård and Sandström (1981, 1985) consider a generalisation for several inequality measures. The generalised income inequality index depends on two parameters which then control the sensitivity of the inequality index to different levels of the income distribution. Several well-known inequality indices descend from or are related to the expression of the generalised index, and therefore, once and for all it is possible to establish their sensitivity to different parts of the income distribution. In fact, the study of the generalised income inequality index makes it possible to analyse and compare the influence of the low, middle and high incomes on various inequality indices. Furthermore, several new inequality measures can be simply derived by tuning the two parameters according to the required sensitivity. Chapter 3 is organised as follows. In Section 2.2, the notation is introduced and some propaedeutic results are covered. In Section 3.2, the expression of the generalised index is presented. Section 3.3 demonstrates how the well-known inequality indices descend from the generalised index, while their sensitivity to different parts of the income distribution is discussed in Section 3.4. Furthermore, two

[3]This chapter is an adaptation of: Ziqing Dong, Yves Tillé, Giovanni M. Giorgi and Alessio Guandalini (2023). Generalised Income Inequality Index. *International Statistical Review,* https://doi.org/10.1111/insr.12551.

methods for approximating the generalised index in the case of finite populations and the related sampling estimators are shown in Sections 3.6 and 3.7.

3.2 Generalisation

For unification and comprehensive understanding of the existing inequality indices in the same framework, a generalised income inequality index is studied. The generalised index is composed of the complementary Bonferroni curve and the beta distribution.

Consider the beta distribution whose PDF is

$$g(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)}, \text{ with } a > 0, b > 0, 0 \leq p \leq 1,$$

where $\mathrm{B}(a,b)$ is the beta function:

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt.$$

The generalised index can be defined as follows:

$$\mathrm{GI}(a,b) = \int_0^1 \overline{Bon}(p)\; g(p|a,b)\; dp. \tag{28}$$

It is bounded between 0 and 1 as $\overline{Bon}(p)$ is bounded between 0 and 1.

Some particular cases of the generalised index are given in Table 7, which involves the Bonferroni index, the Gini index, the Mehran index and the Piesch index. The Pietra index (also known as the Robin Hood index) and the De Vergottini index are also related to the generalised index. All these indices will be briefly presented in the next section.

By substituting p by $F(y)$ in Expression (28),

$$\mathrm{GI}(a,b) = \frac{1}{\mathrm{B}(a,b)} \int_0^\infty \left\{1 - \frac{\mu(y)}{\mu}\right\} \{F(y)\}^{a-1}\{1-F(y)\}^{b-1} dF(y). \tag{29}$$

This integral does not always converge because it depends on $F(.)$ and the parameters a and b.

Consider the incomplete beta function:

$$\mathrm{B}(y;a,b) = \int_0^y t^{a-1}(1-t)^{b-1} dt.$$

The regularised incomplete beta function is defined as

$$I_y(a,b) = \frac{\mathrm{B}(y;a,b)}{\mathrm{B}(a,b)}.$$

Result 4. *If $a > 1$, the generalised index GI(a,b) can also be written as*

$$GI(a,b) = \frac{a+b-1}{\mu(a-1)} \int_0^\infty y I_{F(y)}(a-1,b) dF(y) - \frac{b}{a-1},$$

where $I_{F(y)}(a-1,b)$ is the regularised incomplete beta function.

The proof of Result 4 is given in the appendix of this chapter.

Corollary 1. *If the expectation of the random variable Y exists and if $a > 1$, the integral given in* (29) *converges.*

Proof. According to Result 4, since $I_{F(y)}(a-1,b) \leq 1$ and the expectation of the random variable Y exists, the integral converges. □

3.3 Family of GI(a,b) and counterexamples

The income inequality indices associated with the generalised index are presented in this section. Their historical background and development have been introduced. It has been shown mathematically how these indices are linked with the generalised index. Moreover, indices that cannot be expressed as special cases of the generalised index are added in order to clarify the family of inequality indices that GI(a,b) encompasses.

3.3.1 Bonferroni index

The Bonferroni index has been proposed by Carlo Emilio Bonferroni in 1930. At the begining, it was ostrasized by Corrado Gini and his followers (see Giorgi, 1998, for more details). However, it has been re-discovered 40 years later by Piesch (1975) and Nygård and Sandström (1981). The Bonferroni index shares a lot of properties with the Gini index, but it is more sensitive to the left tail of the income distribution than the Gini index (Pizzetti, 1951; Dong et al., 2021). Furthermore, several extensions and interpretations proposed for the Gini index hold also for the Bonferroni index (see Tarsitano, 1990, for a comprehensive review).

The Bonferroni index is defined as

$$\begin{aligned} Bonferroni &= \mathrm{GI}(1,1) \\ &= \int_0^\infty \left\{ 1 - \frac{\mu(y)}{\mu} \right\} dF(y) \end{aligned} \tag{30}$$

$$= \quad 1 + \frac{1}{\mu} \int_0^{\infty} y \ln F(y) dF(y). \tag{31}$$

If one substitutes p for $F(y)$ in Expression (30), one obtains: $Bonferroni = 1 - \int_0^1 \frac{L(p)}{p} dp$.

3.3.2 Gini index

The Gini (1914) index is the most famous and widespread inequality measure, also known in the literature as the Gini coefficient or the Gini ratio. The Gini index has been proposed for the first time by the namesake author in 1914 with the name of Concentration Ratio.

The success and spread of the Gini index are mainly justified by its simplicity and ease of interpretation due to its intuitive graphical relation with the Lorenz curve (see Giorgi, 1992, 2020). It satisfies the anonymity, scale independence, population independence and the Pigou-Dalton transfer principle[4] (De and Chattopadhyay, 2017). Furthermore, it can be decomposed by sources and by groups in several different ways. It has good inferential properties (see Langel and Tillé, 2013; Graf and Tillé, 2014; Giorgi and Gigliarano, 2017) as well as original interpretations, extensions and application in different fields.

The Gini index is defined as

$$\begin{aligned} Gini &= \mathrm{GI}(2,1) \\ &= 2 \int_0^{\infty} \left\{ 1 - \frac{\mu(y)}{\mu} \right\} F(y) dF(y) \qquad (32) \\ &= \frac{2}{\mu} \int_0^{\infty} t F(t) dF(t) - 1 \qquad (33) \end{aligned}$$

If one substitutes p for $F(t)$ in Expression (32), one obtains: $Gini = 1 - 2 \int_0^1 L(p) dp$.

3.3.3 Mehran index

The Mehran (1976) index belongs to the class of linear measures proposed by Piesch (1975). Mehran derived the expression of his index looking at the condition under which linear measures satisfy the Pigou–Dalton transfer principle. The Mehran index satisfies this principle even when stronger transfer principles are stipulated and is defined as

$$Mehran = \mathrm{GI}(2,2)$$

[4]Proposed by Pigou (1912) and Dalton (1920), the Pigou-Dalton transfer principle imposed a condition on income inequality measures that an income transfer from the rich to the poor should reduce inequality level.

$$= 6\int_0^{\infty}\left\{1-\frac{\mu(y)}{\mu}\right\}F(y)\{1-F(y)\}dF(y) \tag{34}$$

$$= \frac{3}{\mu}\int_0^{\infty} tF(t)\{2-F(t)\}dF(t)-2. \tag{35}$$

If one substitutes p for $F(y)$ in Expression (34), one obtains: $Mehran = 1-6\int_0^1 L(p)(1-p)dp$.

3.3.4 Piesch index

The Piesch index has been proposed in 1975 in a pioneeristic volume on income inequality measures by the namesake author. Same as the Mehran index, it belongs to the class of linear meausures. Both indices can be seen as special cases of a general algorithm proposed by Giaccardi (1950) and by Benedetti (1980) (see Giorgi and Pallini, 1990, for further details). The Piesch index is defined as

$$Piesch = \mathrm{GI}(3,1)$$

$$= 3\int_0^{\infty}\left\{1-\frac{\mu(y)}{\mu}\right\}\{F(y)\}^2dF(y) \tag{36}$$

$$= \frac{3}{2\mu}\int_0^{\infty} t\{F(t)\}^2dF(t)-\frac{1}{2}. \tag{37}$$

If one substitutes p for $F(t)$ in Expression (36), one obtains: $Piesch = 1-3\int_0^1 L(p)pdp$.

3.3.5 1st new index

A new index is here proposed. The generalised index expression for a=1 and b=2 is developed to fill the gap of the indices defined in the previous subsections for adding a more combination of the parameters a and b between the integers 1 and 3, and introduce an index with this level of sensitivity in addition to those already known in the literature. The 1st new index of income inequality is defined as

$$1^{st}\ New = \mathrm{GI}(1,2)$$

$$= 2\int_0^{\infty}\left\{1-\frac{\mu(y)}{\mu}\right\}\{1-F(y)\}\,dF(y) \tag{38}$$

$$= \frac{2}{\mu}\int_0^{\infty} t\,\{\ln F(t)-F(t)\}\,dF(t)+3. \tag{39}$$

If one substitutes p for $F(y)$ in Expression (38), one obtains: $1^{st}\ New = 1-2\int_0^1 \frac{L(p)}{p}(1-p)dp$.

3.3.6 2nd new index

As done for the 1st new index, the two parameters in the expression of the generalised index are tuned for obtaining a second new index with a different level of sensitivity for completing the combinations of the parameters a and b among the integers 1, 2 and 3. The 2nd new index of income inequality is obtained setting $a = 1$ and $b = 3$ in the expression of the generalised index, and it is defined as

$$\begin{aligned} \mathcal{2}^{nd}\ New &= \mathrm{GI}(1,3) \\ &= 3\int_0^\infty \left\{1 - \frac{\mu(y)}{\mu}\right\} \{1 - F(y)\}^2\, dF(y) \qquad (40) \\ &= \frac{3}{2\mu}\int_0^\infty t\left\{2\ln F(t) + \{F(t)\}^2 - 4F(t)\right\} dF(t) + \frac{11}{2}. \qquad (41) \end{aligned}$$

If one substitutes p for $F(y)$ in Expression (40), one obtains: $\mathcal{2}^{nd}\ New = 1 - 3\int_0^1 \frac{L(p)}{p}(1-p)^2 dp$.

3.3.7 De Vergottini index

Mario De Vergottini (1950) proposed another measure of inequality, the De Vergottini index, which is more sensitive to the right tail of the income distribution than the Gini index. Furthermore, he defined a class of inequality indices that includes both the Gini and Bonferroni indices and, of course, the De Vergottini index.

The De Vergottini index is not a particular case of the generalised index but is related to it. The De Vergottini index is defined as

$$\begin{aligned} \textit{De Vergottini} &= \mathrm{B}(2,0)\,\mathrm{GI}(2,0) \\ &= \int_0^\infty \left\{1 - \frac{\mu(y)}{\mu}\right\} \frac{F(y)}{1 - F(y)} dF(y) \qquad (42) \\ &= -\frac{1}{\mu}\int_0^\infty t\ln\{1 - F(t)\} dF(t) - 1. \qquad (43) \end{aligned}$$

If one substitutes p for $F(y)$ in Expression (42), one obtains: $\textit{De Vergottini} = \int_0^1 \frac{p - L(p)}{1-p} dp$.

The De Vergottini index does not always exist since the integral

$$\int_0^\infty t\ln\{1 - F(t)\} dF(t)$$

does not converge for any cumulative distribution function.

3.3.8 Pietra index

The Pietra index is also related to the generalised index. Few months after the publication of the Gini Concentration Ratio, Gaetano Pietra proposed a simple geometrical interpretation of it. He provided the formulation of the Lorenz curve in the continuous case for the first time in the literature, and he derived the expression of the longest vertical distance between the Lorenz curve and the line of perfect equality, which would later receive the appellation the Pietra index. The same result has been proposed by Hoover (1936; 1984) and Schutz (1951). Like the Gini index, it satisfies the anonymity, scale independence, population independence and the Pigou–Dalton transfer principle. The Pietra index is defined as

$$Pietra = \frac{1}{2\mu}\int_0^\infty |y-\mu| f(y)dy.$$

This index is also known as the Robin Hood index due to its simple economic connotation as it can be directly interpreted as the proportion of incomes that should be taken from the wealthier half of the population to the poorer half of the population in order to reach the state of perfect equality.

Result 5. *The Pietra index can also be written as*

$$Pietra = F(\mu) - L(F(\mu)).$$

Proof.

$$\begin{aligned} Pietra &= \frac{1}{2\mu}\int_0^\infty |y-\mu| f(y)dy = \frac{1}{\mu}\int_0^\mu (\mu - y)dF(y) \\ &= \frac{1}{\mu}\int_0^{F(\mu)} \{\mu - Q(p)\}\, dp = F(\mu) - L(F(\mu)). \end{aligned}$$

□

Result 6. *A third way of writing the Pietra index is*

$$Pietra = \max_p (p - L(p)).$$

Proof. The first derivative of $\{p - L(p)\}$ with respect to p is

$$\frac{\partial\{p - L(p)\}}{\partial p} = 1 - \frac{Q(p)}{\mu}.$$

The second derivative of $\{p - L(p)\}$ with respect to p is negative:

$$\frac{\partial^2\{p - L(p)\}}{\partial p^2} < 0.$$

Thus, the maximum distance between p and $L(p)$ is obtained when

$$\arg\max_p\{p - L(p)\} = F(\mu),$$

which corresponds to Result 5. □

The Pietra index is related to the generalised index as

$$\lim_{\substack{a,b\to\infty \\ a/(a+b)\to F(\mu)}} \frac{a\mathrm{GI}(a,b)}{a+b} = F(\mu)\overline{Bon}(F(\mu)) = Pietra.$$

3.3.9 Counterexamples

Counterexamples are the inequality indices that cannot be expressed as special cases of GI(a,b), which include the Zenga concentration index (Zenga, 1984; Tarsitano, 1990; Zenga, 2007; Greselin et al., 2010; Langel and Tillé, 2012). Zenga index can be expressed using the (complementary) Bonferroni curve. See,

$$Zenga = \int_0^1 Z(p)dp,$$

where the Zenga function $Z(p)$ is defined by

$$1 - \frac{L(p)}{p}\frac{1-p}{1-L(p)} = \frac{1 - \mathrm{Bon}(p)}{1 - p\,\mathrm{Bon}(p)}.$$

One argument in favour of the Zenga index is described by Greselin et al. (2010, p. 3): '[...] the Zenga index detects, with the same sensibility, all deviations from equality in any part of the distribution'. However, the Zenga index is not a particular case of the generalised index.

Additionally, the Theil (1967) index is based on the divergence of Kullback and Leibler (1951), which follows a completely different principle from that of the Lorenz curve. The Theil index is a particular case of the generalised entropy index (Shorrocks, 1980). The whole family of Atkinson indices (Atkinson, 1970) can be presented as a monotone transformation of the generalised entropy index. All these indices are unrelated to the generalised index in this chapter.

3.4 Comparisons of inequality indices

From Expression (28), the generalised index can be seen as a functional in which the complementary Bonferroni curve is weighted by the beta density function. In view of this, the parameters a and b defined in the generalised index control the sensitivity of the income inequality indices to different levels of the income distribution. The beta density function exhibits a wide variety of shapes with different values of a and b assumed. To be specific, the beta density function is positively skewed for $a < b$, negatively skewed for $a > b$ and symmetric for $a = b$. When $a = b = 1$, GI(1,1) equals the Bonferroni index and the beta distribution is the same as the standard uniform distribution. Thus, the complementary Bonferroni curve is weighted equally for all income levels for GI(1,1), which consequently is used as a baseline for discussion. When $a < 1$ and $b < 1$, the beta density function is U-shaped, and therefore, compared with the Bonferroni index, both the highest and lowest incomes have relatively significant impacts on the generalised index. On the contrary, when $a > 1$ and $b > 1$, the beta density function is unimodal with its mode equal to $(a-1)/(a+b-2)$, and thus, the generalised index takes more into account low incomes for $a < b$, high incomes for $a > b$ and middle incomes for $a = b$ than the Bonferroni index. Furthermore, when $a < 1$, $b \geq 1$ or $a = 1$, $b > 1$, the beta density function is strictly decreasing, which implies that the generalised index is relatively sensitive to the lowest incomes compared with the Bonferroni index. On the other hand, when $a \geq 1$, $b < 1$ or $a > 1$, $b = 1$, the beta density function is strictly increasing, which results in the generalised index being more sensitive to the highest incomes as opposed to the Bonferroni index.

Consequently, the Gini index plays down lower incomes and emphasises higher incomes in contrast to the Bonferroni index because the Gini index has $a\,(=2)$ and $b\,(=1)$ with the underlying weighting distribution being a straight line with slope +2 towards the complentary Bonferroni curve. The Piesch index puts more weights on the highest incomes than the Gini index since for the Piesch index, $a\,(=3) > 2$ and $b\,(=1) = 1$, resulting in the beta density function being convex and strictly increasing. These results complement the recent findings of Gastwirth (2017). The 1$^{\text{st}}$ new index instead takes more into account lower incomes and downplays higher incomes compared with the Bon-

ferroni index because for the 1st new index, $a=1$ and $b=2$, leading the weight function to be a straight line with slope -2. Subsequently, the 2nd new index places more importance on the lowest incomes than the 1st new index since for the 2nd new index, $a\,(=1)=1$ and $b\,(=3)>2$, bringing about the beta density function being convex and strictly decreasing. The Mehran index focuses more on middle incomes in comparison with the Bonferroni index since for the Mehran index, $a=b\,(=2)>1$, for which the beta density function is symmetric and unimodal with its mode equaling 1/2.

The Pareto (1897) distribution marks the starting point of statistical investigations on personal incomes. Another commonly used distribution to model incomes is the log-normal distribution. There is scientific evidence that the distribution pattern of the log-normal with power law tail is the universal structure of personal income distribution (Souma, 2001). Out of courtesy for history, the Pareto distribution is chosen as a first-step example for calculating the indices.

Suppose Y follows the Pareto distribution. Its PDF is

$$f(y)=\begin{cases}\dfrac{\alpha y_m^{\alpha}}{y^{\alpha+1}} & \text{if } y\in[y_m,+\infty),\\[2ex] 0 & \text{if } y\in(-\infty,y_m),\end{cases}$$

where $y_m>0$ is the minimum possible value of Y and $\alpha>0$ is its shape parameter. The Pareto distribution has an infinite expected value when $0<\alpha\leq 1$. Therefore, it is sensible to restrict the shape parameter to $\alpha>1$ henceforth for a reasonable model of income distribution. The various indices including the new indices derived from the generalised income inequality index are illustrated in Table 7.

Result 7. *The generalised income inequality index under the Pareto distribution is*

$$\begin{cases}\dfrac{B(a-1,b+1-\frac{1}{\alpha})}{B(a,b)}-\dfrac{b}{a-1} & \textit{when } a\neq 1,\\[2ex] \dfrac{\psi(b+1)-\psi(b+1-\frac{1}{\alpha})}{B(1,b)} & \textit{when } a\to 1,\end{cases}$$

where $\psi(x)$ is the digamma function:

$$\psi(x)=\frac{d}{dx}\ln\left(\Gamma(x)\right)=\frac{\Gamma'(x)}{\Gamma(x)}.$$

The proof of Result 7 is given in the appendix.

Table 7: Particular cases of the generalised index (H_n is the n$^{\text{th}}$ harmonic number).

a	b	Index name	Index formula	Pareto
1	1	*Bonferroni*	$\int_0^1 \overline{Bon}(p)dp$	$1 - H_{(\alpha-1)/\alpha}$
2	1	*Gini*	$2\int_0^1 \overline{Bon}(p)pdp$	$\frac{1}{2\alpha-1}$,
1	2	*1st New*	$2\int_0^1 \overline{Bon}(p)(1-p)dp$	$3 - 2H_{2-1/\alpha}$
2	2	*Mehran*	$6\int_0^1 \overline{Bon}(p)p(1-p)dp$	$\frac{2}{3\alpha-1}$
3	1	*Piesch*	$3\int_0^1 \overline{Bon}(p)p^2dp$	$\frac{5\alpha-1}{12\alpha^2-10\alpha+2}$
1	3	*2nd New*	$3\int_0^1 \overline{Bon}(p)(1-p)^2dp$	$\frac{11}{2} - 3H_{3-1/\alpha}$
$a \to 0$	$b \to 0$		$\frac{\overline{Bon}(0)+\overline{Bon}(1)}{2}$	$\frac{1}{2\alpha}$
$a \to 0$	1		$\overline{Bon}(0)$	$\frac{1}{\alpha}$
$a \to 1$	$b \to 0$		$\overline{Bon}(1)$	0
$a, b \to \infty$	$\frac{a}{a+b} \to q$		$\overline{Bon}(q)$	$\frac{1-q}{q}\left[\frac{1}{(1-q)^{1/\alpha}} - 1\right]$

Result 8. *For the Pareto distribution, the De Vergottini index takes the form:*

$$De\ Vergottini = \frac{1}{\alpha - 1}.$$

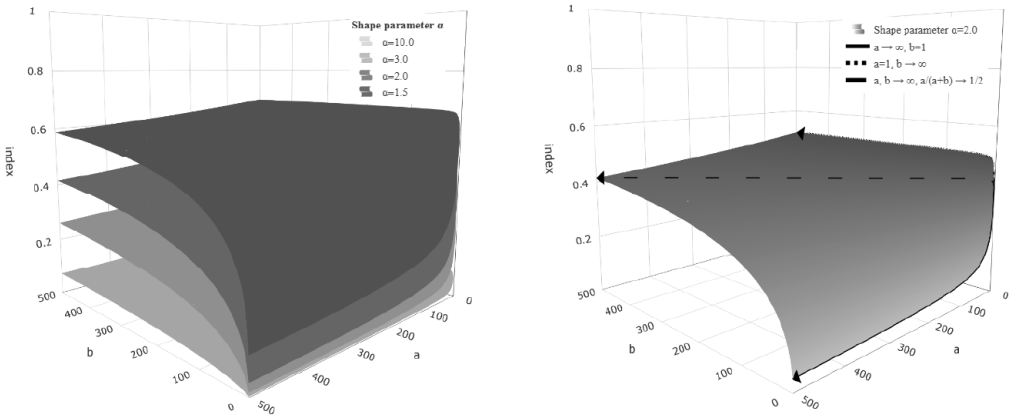
Result 9. *For the Pareto distribution,*

$$F(\mu) = 1 - \left(\frac{\alpha-1}{\alpha}\right)^{\alpha},$$

and thus, the Pietra index takes the form:

$$Pietra = \left(\frac{\alpha-1}{\alpha}\right)^{\alpha} \frac{1}{\alpha-1}.$$

In Figure 1a, the values of the indices with parameters $a, b \in \{1, 2, 3, ..., 500\}$ under the Pareto distribution for different shape parameter α are plotted. When the shape parameter α increases, the value of the index decreases for any given parameters a and



(a) Indices under the Pareto distribution for different values of the shape parameter α.

(b) Indices under the Pareto distribution with the shape parameter $\alpha = 2.0$.

Figure 3.1: The generalised income inequality index with parameters $a, b \in \{1, 2, 3, ..., 500\}$ under the Pareto distribution.

b. This is due to the fact that when the shape parameter α increases, the income data become less dispersed, and therefore, the inequality level shrinks.

In Figure 1b, only the values of the indices under the Pareto distribution with shape parameter $\alpha = 2.0$ are presented. Thus, the aforementioned results can be appreciated by a simple visual inspection. As explained, parameters a and b of the beta distribution define a system of weights exerting on the complementary Bonferroni curve. The complementary Bonferroni curve is a non-increasing curve. As observed, when $a, b \geq 1$, fixing a and enlarging b result in the generalised index being gradually more (resp. less) sensitive to the low (resp. high) incomes and the left side of the complementary Bonferroni curve being increasingly weighted, and thus, the value of the index rises. On the other hand, when $a, b \geq 1$, fixing b and increasing a cause the generalised index to be gradually more (resp. less) sensitive to the high (resp. low) incomes and the right side of the complementary Bonferroni curve being increasingly weighted, and therefore, the value of the index falls. When $a = b \to \infty$ and $a/(a+b) \to 1/2$, the index tends to $\overline{Bon}(1/2)$. Furthermore, the generalised index seems to be sensitive to changes in a, b when both parameters are of very small values. For the Pareto distribution in particular, the index changes also vastly in its value when fixing a to be large and increasing/decreasing b when b is relatively small. This is due to the fact that the Pareto distribution is heavy-tailed and indices of

such forms are sensitive to the highest incomes.

3.5 Influence function of GI(a,b)

As demonstrated in Chapter 2, influence function is an useful tool for illustrating the influence of the different levels of the income distribution on the Gini and the Bonferroni indices. With the proposal of the generalised index GI(a,b), the influence functions of the various income inequality indices in the same framework could also be unified, thereby facilitating the sensitivity analyses of the indices to different levels of income distribution.

Result 10. *If $a \neq 1$, the influence function of the general index of income inequality is*

$$\begin{aligned} IF_x\{GI(a,b)\} &= \frac{x}{\mu}\frac{a+b-1}{a-1}I_{F(x)}(a-1,b) + \frac{1}{\mu B(a,b)}\int_0^\infty y\{F(y)\}^{a-2}\{1-F(y)\}^b dF(y) \\ &\quad -\frac{1}{\mu B(a,b)}\int_0^x y\{F(y)\}^{a-2}\{1-F(y)\}^{b-1}dF(y) - \left\{\frac{b}{a-1} + GI(a,b)\right\}\frac{x}{\mu} \end{aligned} \tag{44}$$

If $a = 1$, the influence function of the general index of income inequality is

$$\begin{aligned} IF_x\{GI(a,b)\} &= IF_x\{GI(1,b)\} \\ &= \frac{x}{\mu}b\left\{\gamma + [F(x) - bF(x)]_3F_2(1,1,2-b;2,2;F(x)) + \ln F(x) + \psi(b)\right\} \\ &\quad +\frac{1}{\mu B(1,b)}\int_0^\infty \frac{y(1-F(y))^b}{F(y)}dF(y) - \frac{1}{\mu B(1,b)}\int_0^x \frac{y(1-F(y))^{b-1}}{F(y)}dF(y) \\ &\quad +\{1 - GI(1,b)\}\frac{x}{\mu}, \end{aligned} \tag{45}$$

where γ is the Euler-Mascheroni constant and ${}_pF_q(a_1, a_2, ..., a_p; b_1, b_2, ..., b_q; y)$ is the generalized hypergeometric function:

$${}_pF_q(a_1, a_2, ..., a_p; b_1, b_2, ..., b_q; y) = \sum_{n=0}^{\infty}\frac{(a_1)_n \cdots (a_p)_n}{(b_1)_n \cdots (b_q)_n}\frac{y^n}{n!},$$

where the Pochhammer symbol $(a)_n$ represents the rising factorial $(a)_0 = 1$ and $(a)_n = a(a+1)(a+2)\cdots(a+n-1),\ n \geq 1$. The proof of Result 10 is in the appendix of this chapter.

From Expression (44), when $a = 2$ and $b = 1$, the influence function of the Gini index proposed for the first time by Monti (1991) is obtained, while from Expression (45), when $a = 1$ and $b = 1$, the result provided by Dong et al. (2021) for the influence function of the Bonferroni index is recovered.

3.6 Finite-population representations

Since, in the real-world applications, finite populations are of interest, the expression of the generalised index is further developed for finite populations. Three rules of defining the income inequality indices in the finite-population case are presented.

3.6.1 Rectangular rule

Consider a finite population $U := \{1, \ldots, i, \ldots, N\}$. The variable of interest takes the value y_i on unit $i \in U$. Without loss of generality, assume that y_i's are sorted in ascending order.

The rectangular version of defining the generalised income inequality index is

$$\mathrm{GI}(a,b)_r = \frac{1}{\mathrm{B}(a,b)} \frac{1}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \left(\frac{k}{N}\right)^{a-1} \left(\frac{N-k}{N}\right)^{b-1},$$

where

$$\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i \quad \text{and} \quad \overline{Y}_k = \frac{1}{k} \sum_{i=1}^{k} y_i \,.$$

Particular cases applying the rectangular rule are

$$\begin{aligned}
\textit{Bonferroni}_r &= \mathrm{GI}(1,1)_r = \frac{1}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \\
\textit{Gini}_r &= \mathrm{GI}(2,1)_r = \frac{2}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \frac{k}{N} \\
\textit{1}^{st}\ \textit{New}_r &= \mathrm{GI}(1,2)_r = \frac{2}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \frac{N-k}{N} \\
\textit{Mehran}_r &= \mathrm{GI}(2,2)_r = \frac{6}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \frac{k}{N} \frac{N-k}{N} \\
\textit{Piesch}_r &= \mathrm{GI}(3,1)_r = \frac{3}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \left(\frac{k}{N}\right)^2 \\
\textit{2}^{nd}\ \textit{New}_r &= \mathrm{GI}(1,3)_r = \frac{3}{N} \sum_{k \in U} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \left(\frac{N-k}{N}\right)^2 \\
\textit{De Vergottini}_r &= B(2,0)\,\mathrm{GI}(2,0)_r = \frac{1}{N} \sum_{\substack{k \in U \\ k < N}} \left(1 - \frac{\overline{Y}_k}{\overline{Y}}\right) \left(\frac{k}{N-k}\right).
\end{aligned}$$

3.6.2 Trapezoidal rule

The trapezoidal rule of defining the income inequality indices is here introduced. Following the insight of Pietra (1915) who used this rule for computing the concentration area between the Lorenz curve and the equidistribution line for the Gini index, Giorgi and Guandalini (2013) adopt it for estimating the Bonferroni index. Needless to say, this technique can be extended to the generalised index and to all the income inequality indices associated.

The trapezoidal formula of defining the generalised income inequality is

$$\mathrm{GI}(a,b)_t = \frac{1}{\mathrm{B}(a,b)}\frac{1}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\left(\frac{k}{N}\right)^{a-1}\left(\frac{N-k}{N}\right)^{b-1}.$$

Particular cases applying the trapezoidal formula are

$$\begin{aligned}
\textit{Bonferroni}_t &= \mathrm{GI}(1,1)_t = \frac{1}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\\
\textit{Gini}_t &= \mathrm{GI}(2,1)_t = \frac{2}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\frac{k}{N}\\
\textit{1}^{st}\ \textit{New}_t &= \mathrm{GI}(1,2)_t = \frac{2}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\frac{N-k}{N}\\
\textit{Mehran}_t &= \mathrm{GI}(2,2)_t = \frac{6}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\frac{k}{N}\frac{N-k}{N}\\
\textit{Piesch}_t &= \mathrm{GI}(3,1)_t = \frac{3}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\left(\frac{k}{N}\right)^2\\
\textit{2}^{nd}\ \textit{New}_t &= \mathrm{GI}(1,3)_t = \frac{3}{N}\sum_{k\in U}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\left(\frac{N-k}{N}\right)^2\\
\textit{De Vergottini}_t &= B(2,0)\,\mathrm{GI}(2,0)_t = \frac{1}{N}\sum_{\substack{k\in U\\ k<N}}\left(1-\frac{\overline{Y}_k+\overline{Y}_{k-1}}{2\overline{Y}}\right)\left(\frac{k}{N-k}\right).
\end{aligned}$$

3.6.3 Reformulation version

By discretising the integrals in Expressions (31), (33), (39), (35), (37), (41) and (43) in Section 3.3 and by replacing $F(t)$ by $(k-1/2)/N$, we propose a new method of defining the income inequality indices. It is named as reformulation version as the method comes

from the reformulation of the formulas of the income inequality indices.

From Expression (31) : $Bonferroni_f = 1 + \frac{1}{\overline{Y}N}\sum_{k\in U} y_k \ln\frac{k-\frac{1}{2}}{N}$.

From Expression (33) : $Gini_f = \frac{2}{\overline{Y}N}\sum_{k\in U} y_k \frac{k-\frac{1}{2}}{N} - 1$.

From Expression (39) : $1^{st}\ New_f = \frac{2}{\overline{Y}N}\sum_{k\in U} y_k \left\{\ln\frac{k-\frac{1}{2}}{N} - \frac{k-\frac{1}{2}}{N}\right\} + 3$.

From Expression (35) : $Mehran_f = \frac{3}{\overline{Y}N}\sum_{k\in U} y_k \frac{k-\frac{1}{2}}{N}\left\{2 - \frac{k-\frac{1}{2}}{N}\right\} - 2$.

From Expression (37) : $Piesch_f = \frac{3}{2\overline{Y}N}\sum_{k\in U} y_k \left(\frac{k-\frac{1}{2}}{N}\right)^2 - \frac{1}{2}$.

From Expression (41) : $2^{nd}\ New_f = \frac{3}{2\overline{Y}N}\sum_{k\in U} y_k \left\{2\ln\frac{k-\frac{1}{2}}{N} + \left(\frac{k-\frac{1}{2}}{N}\right)^2 - 4\left(\frac{k-\frac{1}{2}}{N}\right)\right\} + \frac{11}{2}$.

From Expression (43) : $De\ Vergottini_f = -\frac{1}{\overline{Y}N}\sum_{k\in U} y_k \ln\left(1 - \frac{k-\frac{1}{2}}{N}\right) - 1$.

3.7 Estimation from a sample

Income data are often collected by means of sample surveys, and therefore, estimation and its properties for the inequality indices should not be overlooked. Consider a random sample S selected from population U using a probability sampling design with the inclusion probabilities denoted by π_k, $k \in U$. The values taken by the variable of interest are still assumed to be y_i's but are only known for the units selected in the sample. Consider also that the unit y_i takes weight w_i. The weight w_i could be the inverse of the inclusion probability: $w_i = 1/\pi_i$ (Horvitz and Thompson, 1952). The weights may also be subjected to a calibration procedure (Deville and Särndal, 1992; Särndal, 2007) and could be adjusted in order to compensate questionnaire non-response (Särndal and Lundström, 2005).

One plausible and intuitive estimation for the indices is to use the plug-in estimator. Define

$$\widehat{N} = \sum_{i\in S} w_i,\ \ \widehat{N}_k = \sum_{i\in S} w_i \mathbb{1}(y_i \le y_k),\ \ \widehat{Y} = \sum_{i\in S} w_i y_i,\ \ \widehat{Y}_k = \sum_{i\in S} w_i y_i \mathbb{1}(y_i \le y_k),$$

$$\widehat{\overline{Y}} = \frac{\widehat{Y}}{\widehat{N}} \quad \text{and} \quad \widehat{\overline{Y}} = \frac{\widehat{Y}_k}{\widehat{N}_k}.$$

The plug-in estimator of the generalised index under the rectangular rule is

$$\widehat{\mathrm{GI}}(a,b)_r = \frac{1}{\mathrm{B}(a,b)} \frac{1}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \left(\frac{\widehat{N}_k}{\widehat{N}}\right)^{a-1} \left(\frac{\widehat{N} - \widehat{N}_k}{\widehat{N}}\right)^{b-1},$$

while the plug-in estimator of it following the trapezoidal rule is

$$\widehat{\mathrm{GI}}(a,b)_t = \frac{1}{\mathrm{B}(a,b)} \frac{1}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k + \widehat{\overline{Y}}_{k-1}}{2\widehat{\overline{Y}}}\right) \left(\frac{\widehat{N}_k}{\widehat{N}}\right)^{a-1} \left(\frac{\widehat{N} - \widehat{N}_k}{\widehat{N}}\right)^{b-1}.$$

Particular cases of the plug-in estimators of the income inequality indices built with the rectangular rule are

$$\widehat{Bonferroni}_r = \widehat{\mathrm{GI}}(1,1)_r = \frac{1}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right)$$

$$\widehat{Gini}_r = \widehat{\mathrm{GI}}(2,1)_r = \frac{2}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \frac{\widehat{N}_k}{\widehat{N}}$$

$$\widehat{1^{st}\ New}_r = \widehat{\mathrm{GI}}(1,2)_r = \frac{2}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \frac{\widehat{N} - \widehat{N}_k}{\widehat{N}}$$

$$\widehat{Mehran}_r = \widehat{\mathrm{GI}}(2,2)_r = \frac{6}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \frac{\widehat{N}_k}{\widehat{N}} \frac{\widehat{N} - \widehat{N}_k}{\widehat{N}}$$

$$\widehat{Piesch}_r = \widehat{\mathrm{GI}}(3,1)_r = \frac{3}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \left(\frac{\widehat{N}_k}{\widehat{N}}\right)^2$$

$$\widehat{2^{nd}\ New}_r = \widehat{\mathrm{GI}}(1,3)_r = \frac{3}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \left(\frac{\widehat{N} - \widehat{N}_k}{\widehat{N}}\right)^2$$

$$De\ \widehat{Vergottini}_r = B(2,0)\, \widehat{\mathrm{GI}}(2,0)_r = \frac{1}{\widehat{N}} \sum_{k \in S} w_k \left(1 - \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \left(\frac{\widehat{N}_k}{\widehat{N} - \widehat{N}_k}\right).$$

The plug-in estimators of the associated income inequality indices following the trapezoidal rule and the reformulation version are derived similarly to the rectangular rule. In fact, the inequality indices estimated using the aforementioned three approaches converge very quickly to the same value when the number of observation increases. As the estimators are non-linear, estimation of the standard errors associated to sample estimates can be difficult. Variance estimation is not the focus of this chapter. Nevertheless, a practical method of estimating the sampling variances is promoted as follows. For the estimation of the sampling variance of a non-linear statistic, one approach is to use the

linearisation method, for example, the Graf (2011) method. By computing the derivatives of the sample estimator with respect to the indicator variables of the presence of the units in the sample, the linearised variables can be derived. The linearised variables are then used in the expression of the variance estimator of the total estimator for estimating the sampling variance (see, e.g. Dong et al., 2021). The Graf method can be applied to almost all sampling designs as long as the expression of the variance estimator of the total estimator under the sampling design is known. The details of applying the Graf method for variance estimation can be found in Chapter 2.

3.8 Simulation

The aforementioned income inequality indices (viz. Bonferroni, Gini, 1st new index, Mehran, Piesch, 2nd new index and Pietra) estimators are investigated based on the 2015 IT-SILC data. The indices are computed firstly on the basis of the whole population using the expressions in Section 3.6. They are then estimated based on R=500 replicated samples selected under stratified simple random sampling with proportional allocation (Str-SRS)[5] with sample size ($n = 10000$) for studying the inferential properties of the estimators proposed in Section 3.7.

To begin with, the relative bias (RB) and the normalised root-mean-square error (NRMSE) of each estimator are computed. The empirical RB of $\widehat{GI}(a,b)_r$, $\widehat{GI}(a,b)_t$ or $\widehat{GI}(a,b)_f$ is defined as

$$RBsim\left(\widehat{GI}(a,b)_\bullet\right) = \frac{\frac{1}{R}\sum_{i=1}^{R}\left(\widehat{GI}(a,b)_\bullet^{(i)} - GI(a,b)\right)}{GI(a,b)},$$

while the empirical NRMSE of $\widehat{GI}(a,b)_r$, $\widehat{GI}(a,b)_t$ or $\widehat{GI}(a,b)_f$ is

$$NRMSE_{sim}\left(\widehat{GI}(a,b)_\bullet\right) = \frac{\sqrt{\frac{1}{R}\sum_{i=1}^{R}\left(\widehat{GI}(a,b)_\bullet^{(i)} - GI(a,b)\right)^2}}{\frac{1}{R}\sum_{i=1}^{R}\widehat{GI}(a,b)_\bullet^{(i)}}.$$

[5]For Str-SRS, the strata are determined according to the 21 Italian regions (Piedmont, Aosta Valley, Lombardy, Veneto, Friuli Venezia Giulia, Liguria, Emilia-Romagna, Tuscany, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria, Sicily and Sardinia, plus Province of Bolzano and Province of Trento from the Trentino-Alto Adige/Südtirol region).

Table 8: Empirical RB and NRMSE of the plug-in estimators of the Bonferroni, Gini, 1st new, Mehran, Piesch and 2nd new indices defined using the rectangular rule, the trapezoidal rule and the reformulation method for samples selected under Str-SRS with sample size n = 10000.

Rectangular rule	Sample	Bonferroni GI(1,1)	Gini GI(2,1)	1st new GI(1,2)	Mehran GI(2,2)	Piesch GI(3,1)	2nd new GI(1,3)
Population		0.4942	0.3668	0.6216	0.4947	0.3029	0.6831
	RB	-0.0003	-0.0004	-0.0003	-0.0003	-0.0005	-0.0003
	NRMSE	0.6771	0.9232	0.5705	0.7195	1.1164	0.5362
Trapezoidal rule	Sample	Bonferroni GI(1,1)	Gini GI(2,1)	1st new GI(1,2)	Mehran GI(2,2)	Piesch GI(3,1)	2nd new GI(1,3)
Population		0.4942	0.3668	0.6216	0.4947	0.3029	0.6831
	RB	-0.0002	-0.0003	-0.0002	-0.0002	-0.0003	-0.0002
	NRMSE	0.6766	0.9226	0.5700	0.7192	1.1157	0.5356
Reformulation method	Sample	Bonferroni GI(1,1)	Gini GI(2,1)	1st new GI(1,2)	Mehran GI(2,2)	Piesch GI(3,1)	2nd new GI(1,3)
Population		0.4942	0.3668	0.6216	0.4947	0.3029	0.6831
	RB	-0.0003	-0.0004	-0.0002	-0.0003	-0.0005	-0.0001
	NRMSE	0.6766	0.9231	0.5698	0.7195	1.1164	0.5352

In each selected sample, the inequality indices are estimated using the expressions defined under the rectangular rule, the trapezoidal rule and the reformulation version. The empirical RB and the empirical NRMSE for each index estimator are computed. The results are presented in Table 8.

The indices show different magnitudes of inequality for the same population. The values obtained under the rectangular rule, the trapezoidal rule and the reformulation version for each index are identical when measurement is performed on the whole population, which validates the use of the three definitions in the finite-population case.

Besides the well-investigated relation between the Bonferroni index and the Gini index, further information on the order of magnitude of the indices can be drawn by examining the values that the parameters of the generalised index assume. It is well-known that the Bonferroni index GI(1,1) puts more weights on the lower incomes with respect to the Gini index GI(1,2), resulting in the Bonferroni index assuming larger values than the Gini index except for the extreme cases of minimum and maximum concentration (De Vergottini, 1940, 1950; Pizzetti, 1951). Furthermore, the numerical computation at the population level confirms that when $a, b \geq 1$, the indices with $a < b$ assume larger values and their values increase as b increases, while the indices with $a > b$ assume smaller values, which decrease when a increases, and the values of the indices with $a = b$ tend to

the complementary Bonferroni curve for increasing values of a and b.

Plug-in estimators of the indices built following the trapezoidal rule and the reformulation method have similar results. Their performances are better than those built with the rectangular rule. The advantage of the trapezoidal rule over the rectangular rule in estimation is that it is less biased and shows smaller NRMSE according the simulation results for all indices. The reformulation method has some advantages over the trapezoidal rule for some indices but is outperformed by it for the others.

3.9 Conclusion

In Chapter 3, a deep generalisation for income inequality indices in the same framework is presented. The two parameters of the generalised index control its sensitivity to different parts of the income distribution. It is demonstrated that the family of the generalised index encompasses some most famous inequality indices, such as Gini, Bonferroni, Mehran, Piesch, De Vergottini and Pietra (known also as the Robin Hood index). Two new indices are developed to fill a gap in the literature. In fact, many more indices could be easily defined by tuning the two parameters of the generalised index.

By considering the generalised index as a functional in which the complementary Bonferroni curve is weighted by the beta density function, it is possible to analyse its sensitivity to the income distribution for all different values of a and b. Two definitions of the generalised index in finite populations have been presented. The finite-population representations of the generalised index is practical because it can be directly applied in the real world. The family of the generalised index has been defined following the rectangular rule, the trapezoidal rule and the reformulation version. The three definitions for the same index tend to converge fast to the same value when the number of the observations increases.

Numerical computation and simulation study have been performed based on the 2015 IT-SILC data. The theoretical analyses on the generalised index have been confirmed by the simulation study. Furthermore, it has been shown that trapezoidal rule and reformulation method have certain advantages over the rectangular rule.

All in all, the generalised index provides a unification of the inequality indices in the same framework. With the aim of catching inequality comprehensively, an exhaustive

study on income inequality requires the use of a broad class of inequality measures that is sensitive to different parts of the distribution. For this reason, the generalised index offers a non-trivial approach to the understanding of the sensitivity of inequality indices of the same family to different levels of the distribution.

3.10 Appendix

Proof of Result 4

Proof.

$$\begin{aligned}
\mathrm{GI}(a,b) &= \frac{1}{\mathrm{B}(a,b)}\int_0^\infty \left\{1-\frac{\mu(y)}{\mu}\right\}\{F(y)\}^{a-1}\{1-F(y)\}^{b-1}dF(y)\\
&= 1-\frac{1}{\mathrm{B}(a,b)}\int_0^\infty \frac{1}{\mu F(y)}\int_0^y tdF(t)\{F(y)\}^{a-1}\{1-F(y)\}^{b-1}dF(y)\\
&= 1-\frac{1}{\mu\mathrm{B}(a,b)}\int_0^\infty \int_0^y tdF(t)\{F(y)\}^{a-2}\{1-F(y)\}^{b-1}dF(y)\\
&= 1-\frac{1}{\mu\mathrm{B}(a,b)}\int_0^\infty t\int_t^\infty \{F(y)\}^{a-2}\{1-F(y)\}^{b-1}dF(y)dF(t)\\
&= 1-\frac{1}{\mu\mathrm{B}(a,b)}\int_0^\infty t\{B(a-1,b)-\mathrm{B}(F(t);(a-1),b)\}dF(t)\\
&= 1-\frac{a-1+b}{\mu(a-1)\mathrm{B}(a-1,b)}\int_0^\infty t\{B(a-1,b)-\mathrm{B}(F(t);(a-1),b)\}dF(t)\\
&= 1-\frac{a+b-1}{\mu(a-1)}\int_0^\infty t\{1-I_{F(t)}(a-1,b)\}dF(t)\\
&= \frac{a+b-1}{\mu(a-1)}\int_0^\infty tI_{F(t)}(a-1,b)dF(t)-\frac{b}{a-1}\\
&= \frac{a+b-1}{\mu(a-1)}\int_0^\infty yI_{F(y)}(a-1,b)dF(y)-\frac{b}{a-1}.
\end{aligned}$$

□

Proof of Result 7

Proof.

$$\begin{aligned}
\textit{If } a\neq 1,\ \mathrm{GI}_{Pareto}(a,b) &= \frac{1}{\mathrm{B}(a,b)}\int_0^1\left[1-\frac{1-(1-p)^{\frac{\alpha-1}{\alpha}}}{p}\right]p^{a-1}(1-p)^{b-1}dp\\
&= \frac{1}{\mathrm{B}(a,b)}\int_0^1\left[\frac{(1-p)^{\frac{\alpha-1}{\alpha}}-(1-p)}{p}\right]p^{a-1}(1-p)^{b-1}dp\\
&= \frac{1}{\mathrm{B}(a,b)}\int_0^1\left[\frac{(1-p)^{\frac{\alpha-1}{\alpha}}-(1-p)}{p}\right]p^{a-1}(1-p)^{b-1}dp
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mathrm{B}(a,b)}\left[\int_0^1 p^{a-2}(1-p)^{b-\frac{1}{\alpha}}dp - \int_0^1 p^{a-2}(1-p)^b dp\right] \\
&= \frac{1}{\mathrm{B}(a,b)}\left[\mathrm{B}(a-1,b+1-\frac{1}{\alpha}) - \mathrm{B}(a-1,b+1)\right] \\
&= \frac{\mathrm{B}(a-1,b+1-\frac{1}{\alpha})}{\mathrm{B}(a,b)} - \frac{b}{a-1}.
\end{aligned}
$$

If $a \to 1$, $\lim_{a\to 1} \mathrm{GI}_{Pareto}(a,b) = \dfrac{\psi(b+1) - \psi(b+1-\frac{1}{\alpha})}{\mathrm{B}(1,b)}$.

□

Proof of Result 10

Proof.

$$
\begin{aligned}
\textit{If } a \neq 1,\ IF_x\{\mathrm{GI}(a,b)\} &= -\int_0^1 g(p|a,b)\frac{IF_x\{L(p)\}}{p}dp \\
&= -\frac{1}{\mu}\int_0^1 g(p|a,b)\frac{x\mathbb{1}(x \le Q(p)) + Q(p)[p - \mathbb{1}(x \le Q(p))] - L(p)x}{p}dp \\
&= -\frac{x}{\mu}\int_0^1 g(p|a,b)\frac{\mathbb{1}(x \le Q(p))}{p}dp - \frac{1}{\mu}\int_0^1 g(p|a,b)\frac{Q(p)p}{p}dp \\
&\quad +\frac{1}{\mu}\int_0^1 g(p|a,b)\frac{Q(p)\mathbb{1}(x \le Q(p))}{p}dp + \frac{x}{\mu}\int_0^1 g(p|a,b)\frac{L(p)}{p}dp \\
&= -\frac{x}{\mu}\int_{F(x)}^1 \frac{g(p|a,b)}{p}dp - \frac{1}{\mu}\int_0^1 g(p|a,b)Q(p)dp \\
&\quad +\frac{1}{\mu}\int_{F(x)}^1 g(p|a,b)\frac{Q(p)}{p}dp + \{1 - \mathrm{GI}(a,b)\}\frac{x}{\mu} \\
&= -\frac{x}{\mu}\frac{-\mathrm{B}(F(x);a-1,b) + \mathrm{B}(a-1,b)}{\mathrm{B}(a,b)} - \frac{1}{\mu}\int_0^1 g(p|a,b)Q(p)dp \\
&\quad +\frac{1}{\mu}\int_0^1 g(p|a,b)\frac{Q(p)}{p}dp - \frac{1}{\mu}\int_0^{F(x)} g(p|a,b)\frac{Q(p)}{p}dp + \{1 - \mathrm{GI}(a,b)\}\frac{x}{\mu} \\
&= \frac{x}{\mu}\frac{a+b-1}{a-1}I_{F(x)}(a-1,b) - \frac{x}{\mu}\frac{a+b-1}{a-1} - \frac{1}{\mu}\int_0^1 g(p|a,b)Q(p)dp \\
&\quad +\frac{1}{\mu}\int_0^1 g(p|a,b)\frac{Q(p)}{p}dp - \frac{1}{\mu}\int_0^{F(x)} g(p|a,b)\frac{Q(p)}{p}dp + \{1 - \mathrm{GI}(a,b)\}\frac{x}{\mu} \\
&= \frac{x}{\mu}\frac{a+b-1}{a-1}I_{F(x)}(a-1,b) + \frac{1}{\mu}\int_0^1 g(p|a,b)\frac{(1-p)Q(p)}{p}dp \\
&\quad -\frac{1}{\mu}\int_0^{F(x)} g(p|a,b)\frac{Q(p)}{p}dp - \left\{\frac{b}{a-1} + \mathrm{GI}(a,b)\right\}\frac{x}{\mu} \\
&= \frac{x}{\mu}\frac{a+b-1}{a-1}I_{F(x)}(a-1,b) + \frac{1}{\mu\mathrm{B}(a,b)}\int_0^\infty y\{F(y)\}^{a-2}\{1-F(y)\}^b dF(y)
\end{aligned}
$$

$$-\frac{1}{\mu \mathrm{B}(a,b)} \int_0^x y\{F(y)\}^{a-2}\{1-F(y)\}^{b-1} dF(y) - \left\{ \frac{b}{a-1} + \mathrm{GI}(a,b) \right\} \frac{x}{\mu}.$$

If $a = 1$, $IF_x\{\mathrm{GI}(a,b)\} = -\int_0^1 g(p|a,b) \dfrac{IF_x\{L(p)\}}{p} dp$

$$\begin{aligned}
&= -\frac{1}{\mu} \int_0^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{x\mathbb{1}(x \le Q(p)) + Q(p)[p - \mathbb{1}(x \le Q(p))] - L(p)x}{p} dp \\
&= -\frac{x}{\mu} \int_{F(x)}^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{1}{p} dp - \frac{1}{\mu} \int_0^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} Q(p) dp \\
&\quad + \frac{1}{\mu} \int_{F(x)}^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{Q(p)}{p} dp + \{1 - \mathrm{GI}(1,b)\} \frac{x}{\mu} \\
&= \frac{x}{\mu} b \{\gamma + [F(x) - bF(x)]_3 F_2(1,1,2-b;2,2;F(x)) + \ln F(x) + \psi(b)\} \\
&\quad + \frac{1}{\mu} \int_0^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{(1-p)Q(p)}{p} dp - \frac{1}{\mu} \int_0^{F(x)} \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{Q(p)}{p} dp \\
&\quad + \{1 - \mathrm{GI}(1,b)\} \frac{x}{\mu} \\
&= \frac{x}{\mu} b \{\gamma + [F(x) - bF(x)]_3 F_2(1,1,2-b;2,2;F(x)) + \ln F(x) + \psi(b)\} \\
&\quad + \frac{1}{\mu} \int_0^1 \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{(1-p)Q(p)}{p} dp - \frac{1}{\mu} \int_0^{F(x)} \frac{(1-p)^{b-1}}{\mathrm{B}(1,b)} \frac{Q(p)}{p} dp \\
&\quad + \{1 - \mathrm{GI}(1,b)\} \frac{x}{\mu} \\
&= \frac{x}{\mu} b \{\gamma + [F(x) - bF(x)]_3 F_2(1,1,2-b;2,2;F(x)) + \ln F(x) + \psi(b)\} \\
&\quad + \frac{1}{\mu \mathrm{B}(1,b)} \int_0^\infty \frac{y(1-F(y))^b}{F(y)} dF(y) - \frac{1}{\mu \mathrm{B}(1,b)} \int_0^x \frac{y(1-F(y))^{b-1}}{F(y)} dF(y) \\
&\quad + \{1 - \mathrm{GI}(1,b)\} \frac{x}{\mu}.
\end{aligned}$$

□

4 Simultaneous Confidence Bands for the Lorenz Curve and the Bonferroni curve in a Finite Population

Abstract

The Lorenz curve or the Bonferroni curve is a curve that includes errors in explanatory variables, and is treated completely differently from ordinary regression curves. The construction of simultaneous confidence bands (SCBs) for such curves seems to be a theme that has not been studied so far. Chapter 4 researches the construction of SCBs for the Lorenz curve and the Bonferroni curve, which are two crucial curves for illustrating the concentration of income. The methodology developed could be extended for building SCBs for similar curves when point estimators on such curves are established. The Gini index and the Bonferroni index derived respectively from the two curves are quantified instantly by the contructed SCBs.

Keywords: Bonferroni curve, linearisation, Lorenz curve, simultaneous confidence band.

4.1 Propaedeutics

Consider again a finite population of size N. Suppose y is the variable of interest. Let $y_1, y_2, y_3, \ldots, y_k, \ldots, y_N$ be the values of y, the income variable, assumed by the N population units sorted in ascending order, i.e., $\forall i \leq j$, $y_i \leq y_j$. For simplicity, denote U as a set of identification numbers corresponding to the variables y_k's, in other words, $U = \{1, 2, 3, \ldots, k, \ldots, N\}$. The total, the partial total and the rank of y are respectively defined by

$$Y = \sum_{i \in U} y_i, \qquad Y_k = \sum_{i \in U} y_i \mathbb{1}_{[y_i \leq y_k]}, \qquad N_k = \sum_{i \in U} \mathbb{1}_{[y_i \leq y_k]}.$$

For simplicity, only samples selected without replacement with a fixed sample size are considered. Define a random sample S of size n ($n \leq N$) selected from U as a random variable whose realisations are the samples. A sampling design without replacement for the selection of the sample s with fixed sample size n is a probability mass function $p(\cdot)$ such that $\forall s \subset U$, $p(s) = \mathrm{P}(S = s)$.

Similarly in Chapter 2, let $a_1, a_2, a_3, \ldots, a_k, \ldots, a_N$ be the Bernoulli random variables

indicating the presence of the units in the random sample S, that is to say, $\forall k \in U$,

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

Let π_k be the first-order inclusion probability of unit k and π_{kl} be the second-order inclusion probability that units k and l are both selected in S. In other words, $\forall k, l \in U$,

$$\pi_k = \mathrm{P}(k \in S) = \sum_{\substack{s \subset U \\ k \in s}} p(s) = \mathbb{E}\left[a_k\right] \text{ and } \pi_{kl} = \mathrm{P}(k \in S \,\text{and}\, l \in S) = \sum_{\substack{s \subset U \\ \{k,l\} \subset s}} p(s) = \mathbb{E}\left[a_k a_l\right].$$

By defining the sample indicator variables a_k's, the values of the variable(s) of interest can be separated clearly from the source of the randomness a_k's. The use of the sample indicator variables a_k's was introduced by Cornfield (1944), which plays an essential role in the variance-covariance matrix estimation for the construction of the simultaneous confidence bands (SCBs).

A sampling weight w_k is associated to each population unit $k \in U$. Under the Horvitz-Thompson (1952) formulation, w_k equals $1/\pi_k$, i.e., the inverse of the first-order inclusion probability. The weights can also be determined by calibration. In principle, w_k can be interpreted as the number of units that unit k represents in the population.

The estimators of the population size, the total, the partial total and the rank of y are respectively defined by

$$\widehat{N} = \sum_{i \in U} w_i a_i, \qquad \widehat{Y} = \sum_{i \in U} w_i y_i a_i, \qquad \widehat{Y}_k = \sum_{i \in U} w_i y_i a_i \mathbb{1}_{[y_i \leq y_k]}, \qquad \widehat{N}_k = \sum_{i \in U} w_i a_i \mathbb{1}_{[y_i \leq y_k]}.$$

4.2 Lorenz curve and Bonferroni curve

Lorenz (1905) proposes the Lorenz curve as a method for measuring income concentration: "Plot along one axis cumulated per cents. of the population from poorest to richest, and along the other the per cent. of the total wealth held by these per cents. of the population." The Lorenz curve is defined by

$$\mathcal{L} = (L_1, L_2, L_3, \ldots, L_k, \ldots, L_N),$$

where $L_k = \left(\frac{N_k}{N}, \frac{Y_k}{Y}\right)$ is the Cartesian coordinates of the k^{th} point in a two-dimensional Euclidean space. The Lorenz curve $\mathcal{L}$ is therefore uniquely defined by the N points, $L_1, \ldots, L_k, \ldots, L_N$, within the unit square. Denote the abscissa and ordinate of L_k, for all $k \in U$, by $L_{k(1)}$ and $L_{k(2)}$ respectively. The plug-in estimator of $L_{k(1)}$ is

$$\widehat{L}_{k(1)} = \frac{\widehat{N}_k}{\widehat{N}} = \frac{\sum\limits_{i \in U} w_i a_i \mathbb{1}_{[y_i \leq y_k]}}{\sum\limits_{i \in U} w_i a_i},$$

while the plug-in estimator of $L_{k(2)}$ is

$$\widehat{L}_{k(2)} = \frac{\widehat{Y}_k}{\widehat{Y}} = \frac{\sum\limits_{i \in U} w_i y_i a_i \mathbb{1}_{[y_i \leq y_k]}}{\sum\limits_{i \in U} w_i y_i a_i}$$

Bonferroni (1930) proposes the Bonferroni curve:

$$\mathcal{B} = (B_1, B_2, B_3, \ldots, B_k, \ldots, B_N),$$

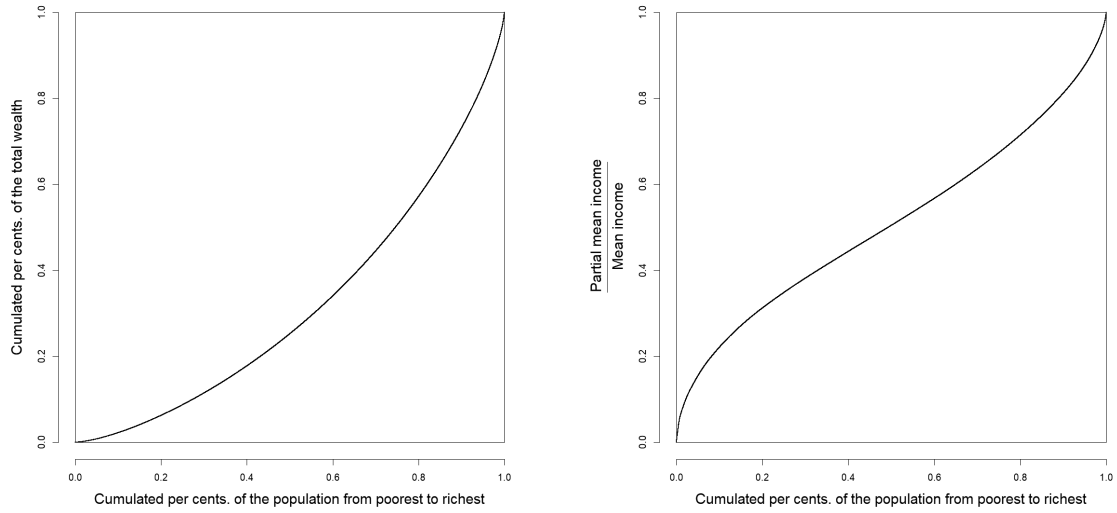
where $B_k = \left(\frac{N_k}{N}, \frac{\overline{Y}_k}{\overline{Y}}\right)$ is the Cartesian coordinates of the k^{th} point in a two-dimensional Euclidean space. The Bonferroni curve $\mathcal{B}$ is therefore uniquely defined by the N points, $B_1, \ldots, B_k, \ldots, B_N$, within the unit square. Denote the abscissa and ordinate of B_k, for all $k \in U$, by $B_{k(1)}$ and $B_{k(2)}$ respectively. The plug-in estimator of $B_{k(1)}$ is

$$\widehat{B}_{k(1)} = \frac{\widehat{N}_k}{\widehat{N}} = \frac{\sum\limits_{i \in U} w_i a_i \mathbb{1}_{[y_i \leq y_k]}}{\sum\limits_{i \in U} w_i a_i},$$

while the plug-in estimator of $B_{k(2)}$ is

$$\widehat{B}_{k(2)} = \frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}} = \left(\frac{\sum\limits_{i \in U} w_i y_i a_i \mathbb{1}_{[y_i \leq y_k]}}{\sum\limits_{i \in U} w_i a_i \mathbb{1}_{[y_i \leq y_k]}}\right) \Bigg/ \left(\frac{\sum\limits_{i \in U} w_i y_i a_i}{\sum\limits_{i \in U} w_i a_i}\right).$$

Figure 4.1 illustrates the Lorenz curve and the Bonferroni curve based on the 2015 IT-SILC data.



(a) Lorenz curve. (b) Bonferroni curve.

Figure 4.1: Lorenz curve and Bonferroni curve based on the 2015 IT-SILC data.

4.3 Linearisation and variance estimation of the Lorenz curve points

The Graf (2011) method introduced in Chapter 2 is applied as the linearisation method for variance estimation.

Result 11. *The sample linearised variable of* $\widehat{L}_{k(1)}$ *is*

$$\hat{z}_{k(1)i} := \frac{\partial \widehat{L}_{k(1)}}{\partial(w_i a_i)} = \frac{\mathbb{1}_{[y_i \leq y_k]}}{\widehat{N}} - \frac{1}{\widehat{N}}\widehat{L}_{k(1)},\ \forall i \in S,$$

and the population linearised variable of $\widehat{L}_{k(1)}$ *is*

$$z_{k(1)i} := \left.\frac{\partial \widehat{L}_{k(1)}}{\partial(w_i a_i)}\right|_{\substack{(w_1 a_1, w_2 a_2, \cdots, w_N a_N) \\ =(1,\,1,\,\cdots,\,1)}} = \frac{\mathbb{1}_{[y_i \leq y_k]}}{N} - \frac{1}{N}L_{k(1)},\ \forall i \in U.$$

Result 12. *The sample linearised variable of* $\widehat{L}_{k(2)}$ *is*

$$\hat{z}_{k(2)i} := \frac{\partial \widehat{L}_{k(2)}}{\partial(w_i a_i)} = \frac{y_i \mathbb{1}_{[y_i \leq y_k]}}{\widehat{Y}} - \frac{y_i}{\widehat{Y}}\widehat{L}_{k(2)},\ \forall i \in S,$$

and the population linearised variable of $\widehat{L}_{k(2)}$ *is*

$$z_{k(2)i} := \left.\frac{\partial \widehat{B}_{k(2)}}{\partial(w_i a_i)}\right|_{\substack{(w_1 a_1, w_2 a_2, \cdots, w_N a_N) \\ =(1,\,1,\,\cdots,\,1)}} = \frac{y_i \mathbb{1}_{[y_i \leq y_k]}}{Y} - \frac{y_i}{Y}L_{k(2)},\ \forall i \in U.$$

The proofs of Result 11 and Result 12 are given in the appendix.

$\forall \xi \in \{1,2\}$, the linearisation of $\widehat{L}_{k(\xi)}$ is:

$$\begin{aligned}\widehat{L}_{k(\xi)(w_1a_1,w_2a_2,\cdots,w_Na_N)} &\approx \widehat{L}_{k(\xi)\underbrace{(1,1,\cdots,1)}_{\mathrm{N}}} + \sum_{i\in U} \frac{\partial \widehat{L}_{k(\xi)} \overbrace{(1,1,\cdots,1)}^{\mathrm{N}}}{\partial(w_ia_i)}(w_ia_i-1) \\ &= L_{k(\xi)} + \sum_{i\in U} w_i z_{k(\xi)i}a_i - \sum_{i\in U} z_{k(\xi)i} \\ &= L_{k(\xi)} + \sum_{i\in S} w_i z_{k(\xi)i} - \sum_{i\in U} z_{k(\xi)i}. \end{aligned} \tag{46}$$

Rearrange Expression (46): $\widehat{L}_{k(\xi)} - L_{k(\xi)} \approx \sum_{i\in S} w_i z_{k(\xi)i} - \sum_{i\in U} z_{k(\xi)i}$. Thus,

$$\mathrm{Var}\left[\widehat{L}_{k(\xi)}\right] \approx \mathrm{Var}\left[\sum_{i\in S} w_i z_{k(\xi)i}\right].$$

By estimating $z_{k(\xi)i}$ by $\hat{z}_{k(\xi)i}$, $\widehat{\mathrm{Var}}\left[\widehat{L}_{k(\xi)}\right] := \widehat{\mathrm{Var}}\left[\sum_{i\in S} w_i \hat{z}_{k(\xi)i}\right]$.

Let $\zeta_{k(\xi)}$ be the total of the linearised variables' estimators of $\widehat{L}_{k(\xi)}$ and $\widehat{\zeta}_{k(\xi)}$ be its estimator, that is to say, $\zeta_{k(\xi)} = \sum_{i\in U} \hat{z}_{k(\xi)i}$ and $\widehat{\zeta}_{k(\xi)} = \sum_{i\in U} w_i \hat{z}_{k(\xi)i}a_i$. In the Horvitz-Thompson set-up, the weight $w_i = 1/\pi_i, \forall i \in U$. For a general sampling design whose first-order (π_i) and second-order (π_{ij}) inclusion probabilities are all positive, a generalised Horvitz-Thompson formulation of the variance estimator of the total estimator $\widehat{\zeta}_{k(\xi)}$ is

$$\widehat{\mathrm{Var}}\left[\widehat{\zeta}_{k(\xi)}\right] = \sum_{i\in S}\sum_{j\in S} \frac{\hat{z}_{k(\xi)i}}{\pi_i}\frac{\hat{z}_{k(\xi)j}}{\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}. \tag{47}$$

Consider simple random sampling without replacement with fixed sample size (SRSWOR): $p(s) = \binom{N}{n}^{-1}$. Its first-order inclusion probabilities are:

$$\pi_k = \sum_{\substack{s\subset U \\ k\in s}} p(s) = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N},\ \forall k \in U,$$

and its second-order inclusion probabilities are:

$$\pi_{kl} = \sum_{\substack{s\subset U \\ \{k,l\}\subset s}} p(s) = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)},\ \forall k,l \in U.$$

Simplification of Expression (12) under SRSWOR results in the Horvitiz-Thompson variance estimator of the total estimator being:

$$\widehat{\mathrm{Var}}\left[\widehat{\zeta}_{k(\xi)}\right] = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{i\in S}\left(\hat{z}_{k(\xi)i} - \widehat{\bar{\zeta}}_{k(\xi)}\right)^2, \tag{48}$$

where $\widehat{\bar{\zeta}}_{k(\xi)}$ is the sample mean of the linearised variables' estimators $\hat{z}_{k(\xi)i}$'s.

Consider stratified simple random sampling without replacement with fixed sample size (StrSRS): $p(s) = \prod_{h=1}^{H} p_h(s_h)$, where $p_h(s_h) = \binom{N_h}{n_h}^{-1}$ and $h = 1, 2, \ldots, H$. Under StrSRS, the population U is stratified into H strata with SRSWOR applied in each stratum. The sample selection within stratum h is independent of the sample selection in all other strata:

$$\widehat{\mathrm{Var}}\left[\widehat{\zeta}_{k(\xi)}\right] = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h - 1}\sum_{i \in S_h}\left(\hat{z}_{k(\xi)i} - \widehat{\bar{\zeta}}^{h}_{k(\xi)}\right)^2, \tag{49}$$

where $\widehat{\bar{\zeta}}^{h}_{k(\xi)}$ is the sample mean calculated within stratum h: $\widehat{\bar{\zeta}}^{h}_{k(\xi)} = \frac{1}{n_h}\sum_{i \in S_h} \hat{z}_{k(\xi)i}$.

4.4 Covariance matrix estimation of the Lorenz curve

Further define $\mathcal{L}_{(1)} := \left[L_{1(1)}, L_{2(1)}, \ldots, L_{N(1)}\right]^\top$ and $\mathcal{L}_{(2)} := \left[L_{1(2)}, L_{2(2)}, \ldots, L_{N(2)}\right]^\top$, of which the plug-in estimators are respectively defined by $\widehat{\mathcal{L}}_{(1)} := \left[\widehat{L}_{1(1)}, \widehat{L}_{2(1)}, \ldots, \widehat{L}_{N(1)}\right]^\top$ and $\widehat{\mathcal{L}}_{(2)} := \left[\widehat{L}_{1(2)}, \widehat{L}_{2(2)}, \ldots, \widehat{L}_{N(2)}\right]^\top$. Note that $\widehat{\mathcal{L}}_{(1)}$ and $\widehat{\mathcal{L}}_{(2)}$ are vector functions with the input (random) vector $\mathbf{w} \circ \mathbf{a} := [w_1 a_1, w_2 a_2, \ldots, w_N a_N]^\top$, where $\circ$ is the Hadamard product, $\mathbf{w} := [w_1, w_2, \ldots, w_N]^\top$ is the non-random (by assumption) vector composed of sampling weights and $\mathbf{a} := [a_1, a_2, \ldots, a_N]^\top$ is the random vector composed of the sample indicator variables. $\forall \xi \in \{1, 2\}$, let $Z_{k(\xi)}$ be the total of the linearised variables of $\widehat{L}_{k(\xi)}$ and $\widehat{Z}_{k(\xi)}$ be its estimator, that is to say, $Z_{k(\xi)} = \sum_{i \in U} z_{k(\xi)i}$ and $\widehat{Z}_{k(\xi)} = \sum_{i \in U} w_i z_{k(\xi)i} a_i$. Define $\mathbf{Z}_{(\xi)} := \left[Z_{1(\xi)}, Z_{2(\xi)}, \ldots, Z_{N(\xi)}\right]^\top$ and $\widehat{\mathbf{Z}}_{(\xi)} := \left[\widehat{Z}_{1(\xi)}, \widehat{Z}_{2(\xi)}, \ldots, \widehat{Z}_{N(\xi)}\right]^\top$. Let $\mathbf{1}$ be the colomn vector of 1's, in other words, $\mathbf{1} = \underbrace{[1,\ 1,\ \cdots,\ 1]^\top}_{\mathrm{N}}$.

Using numerator layout notation, the derivative of $\widehat{\mathcal{L}}_{(\xi)}$ with respect to $\mathbf{w} \circ \mathbf{a}$ is the NxN Jacobian matrix $\mathbf{J}_\xi$ whose (i, j)th entry is the partial derivative:

$$\left[\mathbf{J}_\xi\right]_{ij} = \frac{\partial \widehat{L}_{i(\xi)}}{\partial(w_j a_j)},\ \forall i, j \in U \text{ and } \forall \xi \in \{1, 2\},$$

which by definition equals $\hat{z}_{i(\xi)j}$. The derivative of $\widehat{\mathcal{L}}_{(\xi)}$ at $\mathbf{1}$ is represented by $\mathbf{J}_\xi(\mathbf{1})$ whose (i, j)th entry, by definition, equals $z_{i(\xi)j}$:

$$\left[\mathbf{J}_\xi(\mathbf{1})\right]_{ij} = \left.\frac{\partial \widehat{L}_{i(\xi)}}{\partial(w_j a_j)}\right|_{\substack{\mathbf{w}\circ\mathbf{a} \\ =\mathbf{1}}} = z_{i(\xi)j},\ \forall i, j \in U \text{ and } \forall \xi \in \{1, 2\}.$$

The linearisation of $\widehat{\mathcal{L}}_{(\xi)}$ relies on the generalisation of Taylor's theorem :

$$\widehat{\mathcal{L}}_{(\xi)}(\mathbf{w}\circ\mathbf{a}) \approx \widehat{\mathcal{L}}_{(\xi)}(\mathbf{1}) + \mathbf{J}_{\xi}(\mathbf{1})\Big(\mathbf{w}\circ\mathbf{a}-\mathbf{1}\Big) \tag{50}$$

$$= \mathcal{L}_{(\xi)} + \widehat{\mathbf{Z}}_{(\xi)} - \mathbf{Z}_{(\xi)} \tag{51}$$

Rearrange Expression (50): $\widehat{\mathcal{L}}_{(\xi)} - \mathcal{L}_{(\xi)} \approx \mathbf{J}_{\xi}(\mathbf{1})\Big(\mathbf{w}\circ\mathbf{a}-\mathbf{1}\Big)$. Thus,

$$\mathrm{Var}\Big[\widehat{\mathcal{L}}_{(\xi)}\Big] \approx \mathrm{Var}\Big[\mathbf{J}_{\xi}(\mathbf{1})\Big(\mathbf{w}\circ\mathbf{a}\Big)\Big] = \mathbf{J}_{\xi}(\mathbf{1})\mathrm{Var}\Big[\Big(\mathbf{w}\circ\mathbf{a}\Big)\Big]\mathbf{J}_{\xi}^{\top}(\mathbf{1}).$$

$\forall k, i \in U$, estimating $z_{k(\xi)i}$ by $\hat{z}_{k(\xi)i}$ yields $\widehat{\mathrm{Var}}\Big[\widehat{\mathcal{L}}_{(\xi)}\Big] := \mathbf{J}_{\xi}\,\mathrm{Var}\Big[\Big(\mathbf{w}\circ\mathbf{a}\Big)\Big]\,\mathbf{J}_{\xi}^{\top}$.

Since $\mathrm{Var}\Big[\,(\mathbf{w}\circ\mathbf{a})\,\Big] = \mathbb{E}\Big[\,(\mathbf{w}\circ\mathbf{a})\,(\mathbf{w}\circ\mathbf{a})^{\top}\,\Big] - \mathbb{E}\Big[\,(\mathbf{w}\circ\mathbf{a})\,\Big]\mathbb{E}\Big[\,(\mathbf{w}\circ\mathbf{a})^{\top}\,\Big]$, $\mathrm{Var}\Big[\,(\mathbf{w}\circ\mathbf{a})\,\Big]$ can be represented by Λ whose (i,j)th entry is

$$\Big[\Lambda\Big]_{ij} = \mathbb{E}\Big[w_i w_j a_i a_j\Big] - \mathbb{E}\Big[w_i a_i\Big]\mathbb{E}\Big[w_j a_j\Big], \forall i, j \in U.$$

A Horvitz-Thompson formulation under SRSWOR is: $\forall i, j \in U$,

$$\Big[\Lambda\Big]_{ij} = \begin{cases} \dfrac{N-n}{n}, & \text{if } i = j, \\ -\dfrac{N-n}{n(N-1)}, & \text{if } i \neq j. \end{cases}$$

A Horvitz-Thompson formulation under StrSRS is: $\forall i, j \in U$,

$$\Big[\Lambda\Big]_{ij} = \begin{cases} \dfrac{N_h-n_h}{n_h}, & \text{if } i = j \text{ and } i, j \in U_h, \\ -\dfrac{N_h-n_h}{n_h(N_h-1)}, & \text{if } i \neq j \text{ and } i, j \in U_h, \\ 0, & \text{if else.} \end{cases}$$

In order to tackle problems under some more complex sampling design, where especially the first and second order inclusion probabilities are difficult to be obtained or where the weights rely on the random sample S, a brutal yet straightfoward method of estimating the covariances is as follows. Rearrange Expression (51): $\widehat{\mathcal{L}}_{(\xi)} - \mathcal{L}_{(\xi)} \approx \widehat{\mathbf{Z}}_{(\xi)} - \mathbf{Z}_{(\xi)}$. Thus, $\mathrm{Var}\Big[\widehat{\mathcal{L}}_{(\xi)}\Big] \approx \mathrm{Var}\Big[\widehat{\mathbf{Z}}_{(\xi)}\Big]$. Define the random vector $\widehat{\zeta}_{(\xi)} := \Big[\widehat{\zeta}_{1(\xi)}, \widehat{\zeta}_{2(\xi)}, \ldots, \widehat{\zeta}_{N(\xi)}\Big]^{\top}$. $\forall k, i \in U$, estimating $z_{k(\xi)i}$ by $\hat{z}_{k(\xi)i}$ yields $\widehat{\mathrm{Var}}\Big[\widehat{\mathcal{L}}_{(\xi)}\Big] := \widehat{\mathrm{Var}}\Big[\widehat{\zeta}_{(\xi)}\Big]$, which is an NxN matrix whose (k,l)th entry is:

$$\mathrm{Cov}\Big(\widehat{\zeta}_{k(\xi)}, \widehat{\zeta}_{l(\xi)}\Big) = \frac{1}{\widehat{N}-1}\sum_{i\in S} w_i\Big(\hat{z}_{k(\xi)i} - \widehat{\overline{\zeta}}_{k(\xi)}\Big)\Big(\hat{z}_{l(\xi)i} - \widehat{\overline{\zeta}}_{l(\xi)}\Big).$$

4.5 Cross-covariance estimation of the Lorenz curve

It is possible and not difficult to estimate the cross-covariance matrix of the Lorenz curve. For the capture of the full information of the covariances of the Lorenz curve, define $\mathcal{L}_{\frac{(1)}{(2)}} := \left[L_{1(1)}, L_{2(1)}, \ldots, L_{N(1)}, L_{1(2)}, L_{2(2)}, \ldots, L_{N(2)}\right]^\top$. The plug-in estimator of $\mathcal{L}_{\frac{(1)}{(2)}}$ is $\widehat{\mathcal{L}}_{\frac{(1)}{(2)}} := \left[\widehat{L}_{1(1)}, \widehat{L}_{2(1)}, \ldots, \widehat{L}_{N(1)}, \widehat{L}_{1(2)}, \widehat{L}_{2(2)}, \ldots, \widehat{L}_{N(2)}\right]^\top$ whose covariance matrix is:

$$\operatorname{Var}\left[\widehat{\mathcal{L}}_{\frac{(1)}{(2)}}\right] = \left| \begin{array}{cc} \overbrace{\operatorname{Cov}(\widehat{\mathcal{L}}_{(1)}, \widehat{\mathcal{L}}_{(1)})}^{\Delta_{11}} & \overbrace{\operatorname{Cov}(\widehat{\mathcal{L}}_{(1)}, \widehat{\mathcal{L}}_{(2)})}^{\Delta_{12}} \\ \underbrace{\operatorname{Cov}(\widehat{\mathcal{L}}_{(2)}, \widehat{\mathcal{L}}_{(1)})}_{\Delta_{21}} & \underbrace{\operatorname{Cov}(\widehat{\mathcal{L}}_{(2)}, \widehat{\mathcal{L}}_{(2)})}_{\Delta_{22}} \end{array} \right|.$$

Since Δ_{11} and Δ_{22} are estimable given the discussions in Section 4.4, the cross-covariance matrices Δ_{12} and Δ_{21} are left for evaluation. The linearisation of $\widehat{\mathcal{L}}_{\frac{(1)}{(2)}}$ is:

$$\widehat{\mathcal{L}}_{\frac{(1)}{(2)}}(\mathbf{w} \circ \mathbf{a}) \approx \widehat{\mathcal{L}}_{\frac{(1)}{(2)}}(\mathbf{1}) + \mathbf{J}_{\frac{(1)}{(2)}}(\mathbf{1}) \Big(\mathbf{w} \circ \mathbf{a} - \mathbf{1}\Big), \tag{52}$$

where $\mathbf{J}_{\frac{(1)}{(2)}}$ is a $2N$xN matrix represented as: $\mathbf{J}_{\frac{(1)}{(2)}} = \left[\frac{\mathbf{J}_1}{\mathbf{J}_2}\right]$. Following the same logic,

$$\widehat{\operatorname{Var}}\left[\widehat{\mathcal{L}}_{\frac{(1)}{(2)}}\right] := \mathbf{J}_{\frac{(1)}{(2)}} \operatorname{Var}\left[\Big(\mathbf{w} \circ \mathbf{a}\Big)\right] \mathbf{J}^\top_{\frac{(1)}{(2)}}.$$

For the Horvitz-Thompson formulation, an explicit calculation of $\widehat{\operatorname{Var}}\left[\widehat{\mathcal{L}}_{\frac{(1)}{(2)}}\right]$ under SRSWOR yields: $\forall i, j \in U$,

$$\Big(\widehat{\Delta}_{11}\Big)_{ij} = \frac{N-n}{n} \sum_{\tau \in U} \hat{z}_{j(1)\tau} \left[\hat{z}_{i(1)\tau} - \frac{1}{N-1}\Big(\widehat{\zeta}_{i(1)} - \hat{z}_{i(1)\tau}\Big)\right],$$

$$\Big(\widehat{\Delta}_{22}\Big)_{ij} = \frac{N-n}{n} \sum_{\tau \in U} \hat{z}_{j(2)\tau} \left[\hat{z}_{i(2)\tau} - \frac{1}{N-1}\Big(\widehat{\zeta}_{i(2)} - \hat{z}_{i(2)\tau}\Big)\right],$$

$$\Big(\widehat{\Delta}_{12}\Big)_{ij} = \Big(\widehat{\Delta}_{21}{}^\top\Big)_{ij} = \frac{N-n}{n} \sum_{\tau \in U} \hat{z}_{j(2)\tau} \left[\hat{z}_{i(1)\tau} - \frac{1}{N-1}\Big(\widehat{\zeta}_{i(1)} - \hat{z}_{i(1)\tau}\Big)\right].$$

4.6 Simultaneous confidence bands of the Lorenz curve

Suppose that $\widehat{L}_i := \Big(\widehat{L}_{i(1)}, \widehat{L}_{i(2)}\Big)^\top$, $i = 1, \ldots, n$, are jointly distributed as a bivariate Gaussian distribution with mean $\mu_i = \big(\mu_{i(1)}, \mu_{i(2)}\big)^\top$ and covariance structure

$$\Sigma_{i,j} = \widehat{\operatorname{Cov}}\left(\widehat{L}_i, \widehat{L}_j\right) = \left[\begin{array}{c:c} \Big(\widehat{\Delta}_{11}\Big)_{ij} & \Big(\widehat{\Delta}_{12}\Big)_{ij} \\ \hdashline \Big(\widehat{\Delta}_{21}\Big)_{ij} & \Big(\widehat{\Delta}_{22}\Big)_{ij} \end{array}\right]$$

where μ_i's are unknown and $\Sigma_{i,j}$'s are known.

Assume that the points $\{\mu_i\}_{i=1,\ldots,n}$ are the true Lorenz curve points and $\{\widehat{L}_i\}_{i=1,\ldots,n}$ are the point estimators for them. Suppose the true Lorenz curve is the piecewise linear curve connecting neighbors:

$$\mathcal{L}(t) = (1-\delta)\mu_i + \delta\mu_{i+1} \in \mathbb{R}^2, \quad t \in [1,n],$$

where

$$i = i(t) = [t], \quad \delta = \delta(t) = t - [t].$$

The estimator for $\mathcal{L}(t)$ is

$$\widehat{\mathcal{L}}(t) = (1-\delta)\widehat{L}_i + \delta\widehat{L}_{i+1} \in \mathbb{R}^2, \quad t \in [1,n].$$

As $\widehat{L}_i$ is approximately unbiased with sufficiently large sample size,

$$\widehat{\mathcal{L}}(t) \dot{\sim} N_2\big(\mathcal{L}(t), \Sigma(t)\big) \quad \text{and} \quad \Sigma(t) = \big((1-\delta)I_2, \delta I_2\big) \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,i+1} \\ \Sigma_{i+1,i} & \Sigma_{i+1,i+1} \end{pmatrix} \begin{pmatrix} (1-\delta)I_2 \\ \delta I_2 \end{pmatrix}.$$

Define a chi-square process

$$g(t) = \big(\widehat{\mathcal{L}}(t) - \mathcal{L}(t)\big)^\top \Sigma(t)^{-1} \big(\widehat{\mathcal{L}}(t) - \mathcal{L}(t)\big), \quad t \in [1,n].$$

For each t, $g(t) \sim \chi^2_2$. The tail probability formula for $\max_{t\in[1,n]} g(t)$ is proposed by Eq. (3.2) of Davies (1987). Let c_α be the point such that

$$\mathbb{P}\left(\max_{t\in[1,n]} g(t) < c_\alpha\right) = \mathbb{P}\Big(g(t) < c_\alpha,\ \forall t \in [1,n]\Big) = 1-\alpha.$$

Then, SCB for $\mathcal{L}(t)$ is obtained as

$$\bigcup_{t\in[1,n]} \left\{\mathcal{L}(t) \mid \big(\widehat{\mathcal{L}}(t) - \mathcal{L}(t)\big)^\top \Sigma(t)^{-1} \big(\widehat{\mathcal{L}}(t) - \mathcal{L}(t)\big) < c_\alpha\right\}.$$

Further note that this procedure can be applied as long as $\mathcal{L}(t)$ and $\widehat{\mathcal{L}}(t)$ are of the forms:

$$\mathcal{L}(t) = \phi(t)^\top \boldsymbol{\mu} \quad \text{and} \quad \widehat{\mathcal{L}}(t) = \phi(t)^\top \widehat{\boldsymbol{\mathcal{L}}}, \tag{53}$$

respectively, where

$$\boldsymbol{\mu} = \left(\mu_1^\top, \ldots, \mu_n^\top\right)^\top, \quad \widehat{\boldsymbol{\mathcal{L}}} = \left(\widehat{L}_1^\top, \ldots, \widehat{L}_n^\top\right)^\top,$$

and $\phi(t) \in \mathbb{R}^{2n\times 2}$ a known vector-valued function in t.

Recall that the assumed true curve $\mathcal{L}(t)$ and its estimator $\widehat{\mathcal{L}}(t)$ are given in (53). The variance of $\widehat{\mathcal{L}}(t)$ is $\Sigma(t)$. When t is fixed,

$$g(t) = (\widehat{\mathcal{L}}(t) - \mathcal{L}(t))\Sigma(t)^{-1}(\widehat{\mathcal{L}}(t) - \mathcal{L}(t))$$

is distributed as the chi-square distribution with 2 degrees of freedom

$$\mathbb{P}\big(g(t) > u\big) = \mathbb{P}(\chi_s^2 > u), \quad s = 2.$$

From Davies (1987) (see Theorem A.2),

$$\begin{aligned}\mathbb{P}\bigg(\sup_{t\in[1,n]} g(t) > u\bigg) &\sim \mathbb{P}(\chi_s^2 > u) + C \times u^{\frac{1}{2}(s-1)} e^{-\frac{1}{2}u} \pi^{-\frac{1}{2}} 2^{-\frac{1}{2}s} / \Gamma(\frac{1}{2}s + \frac{1}{2}) \\ &\sim \mathbb{P}(\chi_s^2 > u) + C \times \frac{1}{\sqrt{2\pi}} \mathbb{P}(\chi_{s+1}^2 > u), \quad s = 2,\end{aligned}$$

where

$$C = \int_1^n \mathbb{E}[\|\eta(t)\|]\mathrm{d}t, \quad \mathbb{E}[\|\eta(t)\|] = \mathbb{E}[|\partial g(t)^{\frac{1}{2}}/\partial t|]\pi^{\frac{1}{2}} \frac{\Gamma(\frac{1}{2}s + \frac{1}{2})}{\Gamma(\frac{1}{2}s)}.$$

Estimation of $\mathbb{E}[|\partial g(t)^{\frac{1}{2}}/\partial t|]$ can be obtained by Monte Carlo simulation.

A tentative graphical illustration of the aforementioned procedures of building SCBs for the Lorenz curve is shown in Figure 4.2 based on a sample of size 10 drawn from the 2015 IT-SILC data. The Lorenz curve of the population is the red curve, while the connected black line is the estimated Lorenz curve from the sample. The SCB of the estimated Lorenz curve is the union of the ellipses shown as the blue lines in the figure.

4.7 Extension to the Bonferroni curve

The aforementioned procedure for the construction of SCBs is not restricted solely to the application on the Lorenz curve. In fact, it can be applied whenever point estimators on the curve are established. Take the Bonferroni curve as an example.

Result 13. *The sample linearised variable of* $\widehat{B}_{k(1)}$ *is*

$$\hat{\phi}_{k(1)i} := \frac{\partial \widehat{B}_{k(1)}}{\partial(w_i a_i)} = \frac{\mathbb{1}_{[y_i \le y_k]}}{\widehat{N}} - \frac{1}{\widehat{N}}\widehat{B}_{k(1)}, \; \forall i \in S,$$

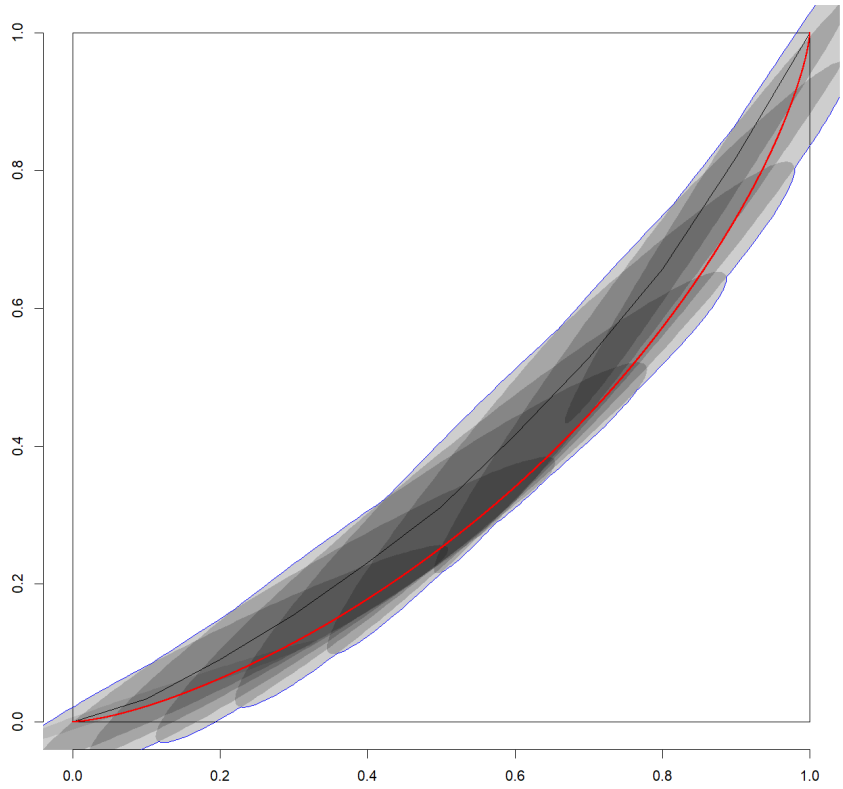


Figure 4.2: Preliminary illustration of the aforementioned procedures of building SCBs for the Lorenz curve based on the 2015 IT-SILC data for a sample of size 10.

and the population linearised variable of $\widehat{B}_{k(1)}$ *is*

$$\phi_{k(1)i} := \frac{\partial \widehat{B}_{k(1)}}{\partial(w_i a_i)}\bigg|_{\substack{(w_1a_1, w_2a_2, \cdots, w_Na_N) \\ =(1,\,1,\,\cdots,\,1)}} = \frac{\mathbb{1}_{[y_i \le y_k]}}{N} - \frac{1}{N} B_{k(1)},\ \forall i \in U.$$

Result 14. *The sample linearised variable of* $\widehat{B}_{k(2)}$ *is*

$$\hat{\phi}_{k(2)i} := \frac{\partial \widehat{B}_{k(2)}}{\partial(w_i a_i)} = \frac{y_i \mathbb{1}_{[y_i \le y_k]}}{\widehat{N}_k \widehat{\overline{Y}}} - \frac{\mathbb{1}_{[y_i \le y_k]}}{\widehat{N}_k} \widehat{B}_{k(2)} - \frac{y_i}{\widehat{Y}} \widehat{B}_{k(2)} + \frac{1}{\widehat{N}} \widehat{B}_{k(2)},\ \forall i \in S,$$

and the population linearised variable of $\widehat{B}_{k(2)}$ *is*

$$\phi_{k(2)i} := \frac{\partial \widehat{B}_{k(2)}}{\partial(w_i a_i)}\bigg|_{\substack{(w_1a_1, w_2a_2, \cdots, w_Na_N) \\ =(1,\,1,\,\cdots,\,1)}} = \frac{y_i \mathbb{1}_{[y_i \le y_k]}}{N_k \overline{Y}} - \frac{\mathbb{1}_{[y_i \le y_k]}}{N_k} B_{k(2)} - \frac{y_i}{Y} B_{k(2)} + \frac{1}{N} B_{k(2)},\ \forall i \in U.$$

The proofs of Result 13 and Result 14 are given in the appendix of this chapter.

Construction of SCBs for the Bonferroni curve can be done by simply replacing $\hat{z}_{k(1)i}$ with $\hat{\phi}_{k(1)i}$ and $\hat{z}_{k(2)i}$ with $\hat{\phi}_{k(2)i}$ and then replicating the procedure.

4.8 Conclusion

Since its advent, the Lorenz curve has been the research topic of numerous studies. It is one of the most important curves for income inequality measurement. Chapter 4 uses the original definition of Lorenz (1905). The Lorenz curve is estimated as the linear interpolation of the plug-in estimators of the Lorenz curve points. The variances and covariances of the point estimators of the Lorenz curve are again estimated by the Graf linearisation method. Relying on the Davies method, chapter 4 proposes a method of building SCBs for the estimation of the Lorenz curve, which fills a vacancy in the study of the subject. Furthermore, as the estimated Lorenz curve contains errors in explanatory variables, the proposed method differs from the standard method of constructing SCBs for regression curves. A preliminary graphical illustration of building SCBs for the Lorenz curve is presented in the current research. Numerical studies on the behaviors of the SCBs are in the direction of future research.

The procedure of building SCBs is not limited to the Lorenz curve. In chapter 2, the Bonferroni index is reexamined as an alternative index alongside with the Gini index for measuring inequality. The Bonferroni index is determined by the Bonferroni curve, of which the SCBs could also be constructed by adopting the same method. In fact, such procedure could be applied to all similar curves as long as point estimators on the curves are established. Therefore, chapter 4 provides a universal method of constructing SCBs for some of the most important curves in the field of income inequality measurement.

4.9 Appendix

Proof of Result 11 and Result 12

Proof. The proofs are a simple application of the quotient rule:

$$\begin{aligned}\forall i \in S,\ \hat{z}_{k(1)i} :&= \frac{\partial \widehat{L}_{k(1)}}{\partial(w_i a_i)} = \frac{\partial}{\partial(w_i a_i)} \left(\frac{\sum\limits_{j\in U} w_j a_j \mathbb{1}_{[y_j \le y_k]}}{\sum\limits_{j\in U} w_j a_j} \right) \\ &= \frac{\mathbb{1}_{[y_i \le y_k]} \left(\sum\limits_{j\in U} w_j a_j \right) - \left(\sum\limits_{j\in U} w_j a_j \mathbb{1}_{[y_j \le y_k]} \right)}{\left(\sum\limits_{j\in U} w_j a_j \right)^2}\end{aligned}$$

$$= \frac{\mathbb{1}_{[y_i \le y_k]}}{\left(\sum_{j \in U} w_j a_j\right)} - \frac{1}{\left(\sum_{j \in U} w_j a_j\right)} \widehat{L}_{k(1)}$$
$$= \frac{\mathbb{1}_{[y_i \le y_k]}}{\widehat{N}} - \frac{1}{\widehat{N}} \widehat{L}_{k(1)}.$$

Similarly,

$$\forall i \in S,\ \hat{z}_{k(2)i} := \frac{\partial \widehat{L}_{k(2)}}{\partial(w_i a_i)} = \frac{\partial}{\partial(w_i a_i)} \left(\frac{\sum_{j \in U} w_j y_j a_j \mathbb{1}_{[y_j \le y_k]}}{\sum_{j \in U} w_j y_j a_j} \right)$$
$$= \frac{\left(y_i \mathbb{1}_{[y_i \le y_k]}\right)\left(\sum_{j \in U} w_j y_j a_j\right) - \left(\sum_{j \in U} w_j y_j a_j \mathbb{1}_{[y_j \le y_k]}\right) y_i}{\left(\sum_{j \in U} w_j y_j a_j\right)^2}$$
$$= \frac{y_i \mathbb{1}_{[y_i \le y_k]}}{\left(\sum_{j \in U} w_j y_j a_j\right)} - \frac{y_i}{\left(\sum_{j \in U} w_j y_j a_j\right)} \widehat{L}_{k(2)}$$
$$= \frac{y_i \mathbb{1}_{[y_i \le y_k]}}{\widehat{Y}} - \frac{y_i}{\widehat{Y}} \widehat{L}_{k(2)}.$$

Obviously, $\forall i \in U$,

$$z_{k(1)i} := \hat{z}_{k(1)i} \Big|_{\substack{(w_1 a_1, w_2 a_2, \cdots, w_N a_N) \\ =(1,\, 1,\, \cdots,\, 1)}} = \frac{\mathbb{1}_{[y_i \le y_k]}}{N} - \frac{1}{N} L_{k(1)}$$

and

$$z_{k(2)i} := \hat{z}_{k(2)i} \Big|_{\substack{(w_1 a_1, w_2 a_2, \cdots, w_N a_N) \\ =(1,\, 1,\, \cdots,\, 1)}} = \frac{y_i \mathbb{1}_{[y_i \le y_k]}}{Y} - \frac{y_i}{Y} L_{k(2)}.$$

□

Proof of Result 13 and Result 14

The proof of Result 13 is the same as the proof of Result 11. For Result 14:

Proof. Since

$$\frac{\partial \widehat{\overline{Y}}_k}{\partial(w_i a_i)} = \frac{y_i \mathbb{1}_{[y_i \le y_k]} \left(\sum_{j \in U} w_j a_j \mathbb{1}_{[y_j \le y_k]} \right) - \left(\sum_{j \in U} w_j y_j a_j \mathbb{1}[y_j \le y_k] \right) \mathbb{1}_{[y_i \le y_k]}}{\left(\sum_{j \in U} w_j a_j \mathbb{1}[y_j \le y_k] \right)^2}$$
$$= \frac{\mathbb{1}[y_i \le y_k]}{\widehat{N}_k} \left(y_i - \widehat{\overline{Y}}_k \right)$$

and

$$\frac{\partial \widehat{\overline{Y}}}{\partial(w_i a_i)} = \frac{y_i\left(\sum\limits_{j\in U} w_j a_j\right) - \left(\sum\limits_{j\in U} w_j y_j a_j\right)}{\left(\sum\limits_{j\in U} w_j a_j\right)^2} = \frac{1}{\widehat{N}}\left(y_i - \widehat{\overline{Y}}\right),$$

then

$$\begin{aligned}
\forall i \in S,\ \hat{\phi}_{k(2)i} :&= \frac{\partial \widehat{B}_{k(2)}}{\partial(w_i a_i)} = \frac{\partial}{\partial(w_i a_i)}\left(\frac{\widehat{\overline{Y}}_k}{\widehat{\overline{Y}}}\right) \\
&= \frac{\frac{\partial \widehat{\overline{Y}}_k}{\partial(w_i a_i)}\widehat{\overline{Y}} - \widehat{\overline{Y}}_k \frac{\partial \widehat{\overline{Y}}}{\partial(w_i a_i)}}{\widehat{\overline{Y}}^2} \\
&= \frac{\frac{\mathbb{1}[y_i \le y_k]}{\widehat{N}_k}\left(y_i - \widehat{\overline{Y}}_k\right)\widehat{\overline{Y}} - \widehat{\overline{Y}}_k \frac{1}{\widehat{N}}\left(y_i - \widehat{\overline{Y}}\right)}{\widehat{\overline{Y}}^2} \\
&= \frac{y_i \mathbb{1}[y_i \le y_k]}{\widehat{N}_k \widehat{\overline{Y}}} - \frac{\mathbb{1}[y_i \le y_k]}{\widehat{N}_k}\widehat{B}_{k(2)} - \frac{y_i}{\widehat{\overline{Y}}}\widehat{B}_{k(2)} + \frac{1}{\widehat{N}}\widehat{B}_{k(2)}.
\end{aligned}$$

□

5 Final Remarks

Precisely measuring income inequality is an important yet difficult task. The Gini index has been widely implemented. However, solely relying on the Gini index is not sufficient to show the differences of the inequality levels between different economies. Recent studies suggest the use of more than one inequality index for better measuring income inequality. The Bonferroni inequality index could be suitable, which is the focus of the present research.

Because the level of income inequality is usually estimated through surveyed samples, the precision of the income inequality indices estimators should not be overlooked. Two estimators of the Bonferroni index are studies based on RB, NRMSE and empirical coverage rates for assessing the accuracies of the estimation. Furthermore, estimators of the income inequality indices are often non-linear statistics, of which the variances could be difficult for estimation. Instead of bootstrapping, which is a possible and well-known method for variance estimation, a linearisation method is discussed. In the present research, the Graf linearised variables are implemented for approximating the variances of the two estimators of the Bonferroni index. The simplicity and accuracy of the Graf method is demonstrated through simulation studies.

The scope of the doctoral research is not restricted to the Gini and the Bonferroni indices. In fact, there are many other income inequality indices which have been proposed in the literature. Although researchers understand the differences among the several prevalent inequality indices, there is no unified theory for providing a comprehensive understanding of them. By defining a generalised income inequality index, the present research provides a unification of multiple inequality indices in the same framework. As a result, the qualities of the various inequality indices associated with the generalised index could be analysed concisely. In addition, two methods (viz. rectangular rule and trapezoidal rule) of defining and estimating the income inequality indices in finite populations could be generalised. The trapezoidal rule of estimating the indices are less biased than the rectangular rule for all sample sizes studied.

The doctoral thesis extends to the studies on the influence functions of the income inequality indices. By computing the influence function of the generalised income in-

equality index, the influence functions of the various income inequality indices in the same framework could be easily derived. Interpreting the Graf linearised variables as discrete analogues of the influence functions, the influence of the different levels of the income distribution on the income inequality indices could be analysed by a simple visual inspection. As each index has its own sensitivity to different levels of the income distribution, a mature analysis of the level of income inequality should be based on a set of carefully chosen indices to have a global view of inequality.

Finally, a method of constructing SCBs for the Lorenz curve and the Bonferroni curve is proposed. Since the Gini index and the Bonferroni index are defined through the two curves respectively, quantification of the uncertainty from the estimation of the two curves is a non-trivial problem. Using again the Graf linearisation method, the variances and as well as the covariances of all the point estimators on the curves could be estimated. Adjusting the critical value according to a specified confidence level by the Davies method, SCBs could be constructed by making the unions of the confidence ellipses of the point estimators on the curves. The procedure could be applied for constructing SCBs for curves of the same kind, such as the Zenga curve.

All in all, this doctoral research contributes to the field of income inequality measurement by defining a generalised income inequality index both in a continuous population and in a finite population. It studies the estimation and variance estimation of the inequality indices and uses influence function for illustrating and comparing the sensitivity of different indices to different levels of the income distribution. It provides a method for the construction of SCBs for essential curves in the area of income inequality measurement. As written in the very beginning, income inequality is a profound subject. It is impossible to include every aspect of it in one manuscript. However, this thesis marks a step for further explorations in the domain.

Bibliography

Aaberge, R. (2000). Characterizations of Lorenz curves and income distributions. *Social Choice and Welfare 17*(4), 639–653.

Aaberge, R. (2007). Gini's nuclear family. *The Journal of Economic Inequality 5*(3), 305–322.

Arbel, Y., C. Fialkoff, A. Kerner, and M. Kerner (2022). Do population density, socio-economic ranking and gini index of cities influence infection rates from coronavirus? israel as a case study. *The Annals of regional science*, 1–26.

Arestis, P. (2018). Importance of tackling income inequality and relevant economic policies. *Inequality: Trends, Causes, Consequences, Relevant Policies*, 1–42.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory 2*, 244–263.

Bárcena, E. and L. J. Imedio (2008). The Bonferroni, Gini, and De Vergottini indices. Inequality, welfare, and deprivation in the European union in 2000. In *Inequality and Opportunity: Papers from the Second ECINEQ Society Meeting*, Bingley (UK), pp. 231–257. Emerald Group Publishing Limited.

Bárcena-Martín, E. and J. Silber (2011). On the concepts of Bonferroni segregation index and curve. *Rivista Italiana di Economia, Demografia e Statistica 62*(2), 57–74.

Bárcena-Martín, E. and J. Silber (2013). On the generalization and decomposition of the Bonferroni index. *Social Choice and Welfare 41*(4), 763–787.

Bárcena-Martín, E. and J. Silber (2017). The Bonferroni index and the measurement of distributional change. *Metron 75*(1), 1–16.

Benedetti, C. (1980). Di alcuni indici di disuguaglianza del benessere. *Statistica 40*(1), 7–12.

Benedetti, C. (1986). Sulla interpretazione benesseriale di noti indici di concentrazione e di altri. *Metron 44*(1), 421–429.

Berger, Y. and İ. Gedik Balay (2020). Confidence intervals of gini coefficient under unequal probability sampling. *Journal of Official Statistics 36*(2), 237–249.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review 5*(3), 279–292.

Binder, D. A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the ASA Survey Research Methods Section 1991*, 34–42.

Binder, D. A. and Z. Patak (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association 89*(427), 1035–1043.

Bonferroni, C. E. (1930). *Elementi di statistica generale*. Firenze: Libreria Seber.

Bonferroni, C. E. (1933). *Elementi di statistica generale*. Torino: Litografia Felice Gili.

Chakravarty, S. R. (2007). A deprivation-based axiomatic characterization of the absolute Bonferroni index of inequality. *The Journal of Economic Inequality 5*(3), 339–351.

Chakravarty, S. R. and P. Muliere (2004). Welfare indicators: a review and new perspectives. 2. measurement of poverty. *Metron 62*(2), 247–281.

Cornfield, J. (1944). On samples from finite populations. *Journal of the American Statistical Association 39*(226), 236–239.

Dabla-Norris, M. E., M. K. Kochhar, M. N. Suphaphiphat, M. F. Ricka, and M. E. Tsounta (2015). *Causes and consequences of income inequality: A global perspective*. International Monetary Fund.

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal 30*(119), 348–361.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 74*(1), 33–43.

De, S. K. and B. Chattopadhyay (2017). Minimum risk point estimation of Gini index. *Sankhya B 79*(2), 247–277.

De Vergottini, M. (1940). Sul significato di alcuni indici di concentrazione. *Giornale degli Economisti e Annali di Economia 2*(5-6), 317–347.

De Vergottini, M. (1950). Sugli indici di concentrazione. *Statistica 10*(4), 445–454.

Decancq, K. and M. A. Lugo (2012). Inequality of wellbeing: A multidimensional approach. *Economica 79*(316), 721–746.

Demnati, A. and J. N. K. Rao (2004). Linearization variance estimators for survey data. *Survey Methodology 30*(1), 17–26.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology 25*(2), 193–203.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Dong, Z., Y. Tille, G. M. Giorgi, and A. Guandalini (2023). Generalised income inequality index. *International Statistical Review*.

Dong, Z., Y. Tillé, G. M. Giorgi, and A. Guandalini (2021). Linearization and variance estimation of the bonferroni inequality index. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(3), 1008–1029.

Essama-Nssah, B. and P. Lambert (2011). Influence functions for distributional statistics. *Society for the Study of Economic Inequality working paper. ECINEQ WP 236*, 2011.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica 39*(6), 1037–1039.

Gastwirth, J. L. (2017). Is the Gini index of inequality overly sensitive to changes in the middle of the income distribution? *Statistics and Public Policy 4*(1), 1–11.

Giaccardi, F. (1950). Indici di concentrazione. *Giornale degli Economisti e Annali di Economia anno IX (nuova serie),*(5-6), 282–290.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto veneto di scienze, lettere ed arti 73*, 1203–1248. English traslation In *Metron* 2005, 63(1):3–38.

Giordani, P. and G. M. Giorgi (2010). A fuzzy logic approach to poverty analysis based on the Gini and Bonferroni inequality indices. *Statistical Methods and Applications 19*(4), 587–607.

Giorgi, G. M. (1992). *Il rapporto di concentrazione di Gini: Genesi, evoluzione ed una bibliografia commentata*. Siena: Libreria Ticci.

Giorgi, G. M. (1998). Concentration index, Bonferroni. In S. Kotz, H. L. Johnson, and C. B. Read (Eds.), *Encyclopedia of Statistical Sciences, Update 2*, Volume 2, pp. 141–146. New York: Wiley.

Giorgi, G. M. (2005). Gini's scientific work: an evergreen. *Metron 63*(3), 299–315.

Giorgi, G. M. (2019). The Gini concentration ratio: Back to the future. *Rivista Italiana di Economia, Demografia e Statistica 73*(2), 5–14.

Giorgi, G. M. (2020). *Gini coefficient*. SAGE Publications Limited.

Giorgi, G. M. and M. Crescenzi (2001a). Bayesian estimation of the Bonferroni index from a Pareto-type I population. *Statistical Methods and Applications 10*(1-3), 41–48.

Giorgi, G. M. and M. Crescenzi (2001b). A look at the Bonferroni inequality measure in a reliability framework. *Statistica 61*(4), 571–583.

Giorgi, G. M. and C. Gigliarano (2017). The Gini concentration index: A review of the inference literature. *Journal of Economic Surveys 31*(4), 1130–1148.

Giorgi, G. M. and A. Guandalini (2013). A sampling estimator of the Bonferroni inequality index. *Rivista Italiana di Economia, Demografia e Statistica 67*(3-4), 151–158.

Giorgi, G. M. and A. Guandalini (2018). Decomposing the Bonferroni inequality index by subgroups: Shapley value and balance of inequality. *Econometrics 6*(2), 1–18.

Giorgi, G. M. and R. Mondani (1994). The exact sampling distribution of the Bonferroni concentration index. *Metron 52*(3-4), 5–41.

Giorgi, G. M. and R. Mondani (1995a). Sampling distribution of the Bonferroni inequality index from exponential population. *Sankhyā: The Indian Journal of Statistics, Series B 57*(1), 10–18.

Giorgi, G. M. and R. Mondani (1995b). Sampling distribution of the Bonferroni inequality index from exponential population. *Sankhyā B57*, 10–18.

Giorgi, G. M. and S. Nadarajah (2010). Bonferroni and Gini indices for various parametric families of distributions. *Metron 68*(1), 23–46.

Giorgi, G. M. and A. Pallini (1990). Inequality indices: Theoretical and empirical aspects of their asymptotic behaviour. *Statistical Papers 31*(1), 65–76.

Graf, E. and Y. Tillé (2014). Variance estimation using linearization for poverty and social exclusion indicators. *Survey Methodology 40*(1), 61–79.

Graf, M. (2011). *Use of survey weights for the analysis of compositional data*, pp. 114–127. Chichester: Wiley.

Greselin, F., L. Pasquazzi, and R. Zitikis (2010). Zenga's new index of economic inequality, its estimation, and an analysis of incomes in Italy. *Journal of Probability and Statistics 2010, ID 718905*.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*(346), 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1985). *Robust statistics: the approach based on influence functions*, Volume 196. John Wiley & Sons.

Hasisi, B., S. Perry, Y. Ilan, and M. Wolfowicz (2020). Concentrated and close to home: the spatial clustering and distance decay of lone terrorist vehicular attacks. *Journal of quantitative criminology 36*, 607–645.

Hoover, E. M. j. (1936). The measurement of industrial localization. *Review of Economics and Statistics 18*, 162–171.

Hoover, E. M. j. and F. Giarratani (1984). *An Introduction to Regional Economics*. New York: Alfred A. Knopf, Inc.

Hörcher, D. and D. J. Graham (2021). The gini index of demand imbalances in public transport. *Transportation 48*(5), 2521–2544.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Imedio-Olmedo, L. J., E. M. Parrado-Gallardo, and E. Bárcena-Martín (2012). Income inequality indices interpreted as measures of relative deprivation/satisfaction. *Social Indicators Research 109*(3), 471–491.

Istat (2015). *Indagine sulle condizioni di vita (UDB IT-SILC)*. Available on-line: https://www.istat.it/it/archivio/4152, (accessed on 4 December 2017).

Jihui Tu, Haigang Sui, W. F. K. S. C. X. and Q. Han (2017). Detecting building façade damage from oblique aerial images using local symmetry feature and the gini index. *Remote Sensing Letters 8*(7), 676–685.

Kalton, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society: Series A (General) 142*(2), 210–222.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics 22*(1), 79–86.

Langel, M. and Y. Tillé (2012). Inference by linearization for Zenga's new inequality index: a comparison with the Gini index. *Metrika 75*(8), 1093–1110.

Langel, M. and Y. Tillé (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(2), 521–540.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association 9*(70), 209–219.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R*, Volume 565. John Wiley & Sons.

Mehran, F. (1976). Linear measures of income inequality. *Econometrica 44*(4), 805–809.

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica 51*(4), 561–580.

Nygård, F. and A. Sandström (1981). *Measuring income inequality*. Stockholm: Almqvist & Wiksell International.

Nygård, F. and A. Sandström (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.

Nygård, F. and A. Sandström (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics 1*, 399–412.

Nygård, F. and A. Sandström (1989). Income inequality measures based on sample surveys. *Journal of Econometrics 42*(1), 81–95.

Osberg, L. (2017). On the limitations of some current usages of the Gini index. *Review of Income and Wealth 63*(3), 574–584.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. In *Survey Research Methods*, Volume 3, pp. 167–195.

Pareto, V. (1897). Cours d'economique politique (vol. 2, part i, chapter 1). *Lausanne, Switzerland*.

Pernot, P. and A. Savin (2021). Using the gini coefficient to characterize the shape of computational chemistry error distributions. *Theoretical Chemistry Accounts 140*, 1–11.

Pickett, K. and R. Wilkinson (2010). *The spirit level: Why equality is better for everyone*. Penguin UK.

Piesch, W. (1975). *Statistische Konzentrationsmaße*. Tübingen: J.B.C. Mohr (Paul Siebeck).

Pietra, G. (1915). Delle relazioni tra gli indici di variabilità. *Atti del Reale Istituto veneto di scienze, lettere ed arti 74*, 775–804. English traslation In *Metron* 2014, 72(1):5–16.

Pigou, A. C. (1912). *Wealth and welfare*. Macmillan and Company, limited.

Piketty, T. (2015). About capital in the twenty-first century. *American Economic Review 105*(5), 48–53.

Pizzetti, E. (1951). Relazioni tra indici di concentrazione. *Statistica 11*(3-4), 294–316.

Pundir, S., S. Arora, and K. Jain (2005). Bonferroni curve and the related statistical inference. *Statistics and Probability Letters 75*(2), 140–150.

Sandström, A., J. H. Wretman, and B. Waldén (1985). Variance estimators of the Gini coefficient: Simple random sampling. *Metron 43*, 41–70.

Sandström, A., J. H. Wretman, and B. Waldén (1988). Variance estimators of the Gini coefficient: Probability sampling. *Journal of Business and Economic Statistics 6*, 113–120.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*(2), 99–119.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. Hoboken: John Wiley & Sons.

Schutz, R. R. (1951). On the measurement of income inequality. *The American Economic Review 41*(1), 107–122.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, 613–625.

Silber, J. (1999). Introduction: Thirty years of intensive research on income inequality measurement. *Handbook of income inequality measurement*, 1–18.

Silber, J. and H. Son (2010). On the link between the Bonferroni index and the measurement of inclusive growth. *Economics Bulletin 30*(1), 421–428.

Souma, W. (2001). Universal structure of the personal income distribution. *Fractals 9*(04), 463–470.

Stigilitz, J. E. (2012). The price of inequality.

Tarsitano, A. (1990). The Bonferroni index of income inequality. In C. Dagum and M. Zenga (Eds.), *Income and Wealth Distribution, Inequality and Poverty*, pp. 228–242. Berlin: Springer.

Theil, H. (1967). *Economics and Information Theory*. Amsterdam: North-Holland.

Vallée, A.-A. and Y. Tillé (2019). Linearisation for variance estimation by means of sampling indicators: Application to non-response. *International Statistical Review 87*(2), 347–367.

Wolter, K. (2007). *Introduction to variance estimation*. Springer Science & Business Media.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association 66*(334), 411–414.

Xu, K. (2003). How has the literature on Gini's index evolved in the past 80 years? *Dalhousie University, Economics Working Paper*.

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling Gini. *Research on Economic Inequality 8*, 13–30.

Zardetto, D. (2015). Regenesees: an advanced R system for calibration, estimation and sampling error assessment in complex sample surveys. *Journal of Official Statistics 31*(2), 177–203.

Zenga, M. (1984). Proposta per un indice di concentrazione basato sui rapporti tra quantili di popolazione e quantili di reddito. *Giornale degli Economisti e Annali di Economia 43*, 301–326.

Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica e Applicazioni 4*, 3–27.