Université de Neuchâtel
Institut de Statistique

# Measuring inequality in finite population sampling

Thèse présentée à la Faculté des sciences économiques
pour l'obtention du grade de Docteur en Statistique

par

Matti Langel

Acceptée sur proposition du jury :

| | | |
|---|---|---|
| Prof. Catalin STARICA | Université de Neuchâtel | Président du jury |
| Prof. Yves TILLÉ | Université de Neuchâtel | Directeur de thèse |
| Prof. Anne RUIZ-GAZEN | Université Toulouse I | Rapporteur |
| Prof. Chris SKINNER | London School of Economics | Rapporteur |

Thèse soutenue le 27 février 2012.

UNIVERSITÉ DE
NEUCHÂTEL

**IMPRIMATUR POUR LA THÈSE**

Measuring inequality in finite population sampling

**Matti LANGEL**

---

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
Prof. Catalin Starica (président du jury, Université de Neuchâtel)
Prof. Anne Ruiz-Gazen (Université de Toulouse 1)
Prof. Chris Skinner (London School of Economics)

Autorise l'impression de la présente thèse.

Neuchâtel, le 15 mars 2012

Le doyen

Gerald Reiner

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

*To the memories of my mother Françoise*
*and my friend Nicolas.*

# Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Yves Tillé, for giving me the opportunity to work with him and benefit from his knowledge and experience, not only in the field of survey sampling but also more generally as a researcher. His support, guidance and enthusiasm have been a constant source of motivation. Thank you, Yves, for teaching me so much.

I would also like to thank Prof. Catalin Starica for chairing the thesis jury, as well as Prof. Anne Ruiz-Gazen and Prof. Chris Skinner for kindly accepting to act as reviewers for this thesis. In addition, I would like to thank the Swiss National Science Foundation for having funded this research for two and a half years (grant no. 200021-121604). Gérard Geiser (Service statistique, Canton de Neuchâtel) should also be deeply thanked for the data set he made available to us.

For providing such a pleasant work environment during these years, I would like to thank all my colleagues and former colleagues at the Institute of Statistics, University of Neuchâtel and in particular to Anthea Monod, Erika Antal and Lionel Qualité for all their help and friendship. I am also grateful to all those who helped me improve different parts of this work by providing useful comments and suggestions in informal discussions, referee reports or conference sessions.

Many thanks go to my family and friends for their long-lasting support. Finally, I wish to thank my wife Claude for her love, friendship, constant encouragement and most importantly, for being who she is.

Neuchâtel, January 19, 2012.

# Abstract and keywords


### Measuring inequality in finite population sampling

**Abstract** This document focuses on the estimation of inequality measures for complex survey data. The proposed methodology takes into account both the complexity of these generally non-linear functions of interest and the complexity of the sampling strategy. The first chapter is dedicated to the presentation and definition of the main concepts used in both inequality and survey sampling theory. In the second chapter, a variety of inequality indices are compared in an empirical study on a real set of income data. Research is then directed towards three specific inequality measures: the Quintile share ratio (QSR), the Gini index and Zenga's new inequality index. The third chapter shows that the variance of the QSR can be estimated by means of the linearization approach without applying a kernel smoothing, and that a simple transformation enhances the coverage rate of the confidence interval. The two following chapters discuss the work of Corrado Gini from an unusual angle. For instance, both balanced sampling (of which he is a pioneer) and variance estimation for the inequality measure that bears his name are discussed in a historical perspective. Zenga's new inequality index is presented in the last chapter and a variance estimator is proposed.



**Keywords** survey sampling, variance estimation, Gini index, Lorenz curve, linearization, income, balanced sampling.



### Mesures d'inégalité et échantillonnage en population finie

**Résumé** Ce document se concentre sur l'estimation des mesures d'inégalité à l'aide de données d'enquête. La méthodologie proposée permet de tenir compte du caractère non-linéaire des mesures d'inégalité ainsi que de la complexité de la stratégie d'échantillonnage. Le premier chapitre est dédié à la présentation et à la définition des concepts principaux de l'étude quantitative des inégalités et de la théorie des sondages. Dans le second chapitre, plusieurs indices d'inégalité sont comparés au sein d'une étude empirique réalisée à l'aide de données réelles. La recherche se centre ensuite vers trois mesures d'inégalités spécifiques : le Quintile share ratio (QSR), l'indice de Gini et l'indice de Zenga. Ainsi, dans le troisième chapitre, nous montrons que la variance du QSR peut être estimée par linéarisation sans avoir recours à un lissage par noyau et qu'une simple transformation permet d'améliorer le taux de couverture de l'intervalle de confiance. Les deux chapitres suivants abordent les travaux de Corrado Gini sous un angle particulier, notamment à travers des réflexions historiques sur l'échantillonnage équilibré dont il a été l'un des pionniers, et sur l'estimation de variance de l'indice d'inégalité qui porte son nom. L'ultime chapitre est dédié à la présentation d'une mesure moins connue, l'indice de Zenga, pour laquelle nous proposons un estimateur de variance.



**Mots-clés** sondage, estimation de variance, indice de Gini, courbe de Lorenz, linéarisation, revenu, échantillonnage équilibré.

# Contents

# List of Figures

# List of Tables

# Introduction

The term *inequality* is often used as a shortcut for the unequal allocation of *income*. Although the main concern of this thesis is income inequality, one may note that from a statistical point of view, the field of application of inequality measures is not restricted to income or monetary matters. Indeed, an inequality measure is nothing more than an indicator of dispersion or heterogeneity and can thus be used in plenty of applications. The reader should therefore keep in mind that whenever the term income is used hereafter, the latter could also be replaced by any other quantitative characteristic allocated to any type of statistical unit.

One of the key objectives of measuring inequality is the production of synthetic indicators which summarize in a single scalar value the level of inequality of a population. These indicators are often complex, non-linear functions of interest. In practice, the level of inequality in a population is estimated by means of a sample. This sample is generally selected from a finite population using a (potentially complex) random sampling design. Inferring from a random sample to a finite population involves a specific methodology which takes the sampling design into account and in which classical statistical results may not be valid. While research on inequality measures is too often conducted without regard for its applicability in an official statistic framework, the goal of this thesis is to link inequality theory and survey sampling and propose methods of inference that account for both the complexity of the functions of interest and the complexity of the sampling strategy.

This document focuses mainly on variance estimation, which is one of the main challenges when working with both complex statistics and sampling designs. Obviously, a point estimate has only little relevance if it is not associated with a measure of accuracy, such as a variance estimator, a standard error or a confidence interval. In addition to an overview of many measures, three inequality indices will be studied in depth in the following chapters: the Gini index, the Quintile Share Ratio and Zenga's new inequality index. This choice is the result of two distinct objectives: (1) improve methodology for already widespread measures and (2) propose new measures or bring recent ones to light. Working on the first two measures cited above is in line with the first objective. Both the Gini index and the Quintile Share Ratio were indeed selected by the European Council in 2001 as the two income inequality indicators to be estimated by

all member states of the European Union (see for example Atkinson et al., 2004). Zenga's new inequality index, on the other hand, is a recently proposed measure which is not yet well established among practitioners despite its very good properties.

Before presenting the structure of the thesis, let us start by briefly discussing what this research is not about. The issue of defining the variable or characteristic of interest itself is not discussed. The example of income is intricate because it can be defined in many ways and can be difficult to construct. An income variable can indeed be the result of a combination of different income sources. A lot of questions thus arise: Should income be measured at the individual or household level? Should one measure gross or net income? Should welfare benefits be taken into account? How should the presence of dependent children in the household be dealt with? Should one allow the presence of negative incomes in the data? Despite the obvious importance of this issue in the process of measuring inequality, this document focuses on methodology which comes into play once the data has been processed. Thus, incomes are considered as given and non-response issues are also eluded. Non-response treatment is nonetheless a crucial subject as well, especially when dealing with such a sensitive variable as income because it seems reasonable to think that some categories of income earners are less prone to respond to surveys than others. Finally, although many simulations and examples throughout the document introduce real sets of income data from the Canton of Neuchâtel, Switzerland, this research is by no means an analysis of the level of inequality in that region.

The document is structured as follows. Chapter 1 introduces the main concepts, issues and notations that are needed when estimating inequality measures in finite population sampling. Both inequality theory tools and survey sampling notions are defined. Moreover, the main issues and motivations regarding the topic are raised. The chapter serves as an introduction to the further chapters where all the original research is contained.

Chapters 2 to 6 are self-contained papers published or submitted in peer-reviewed international journals. Chapter 2 is an empirical study and is therefore not very methodologically intensive. Nevertheless, it is the starting point of most research directions undertaken in this thesis. In this chapter, no fewer than seventeen different inequality measures are presented, including two new propositions. Assuming that detecting changes in the level of inequality of a population or distribution is one of the major roles of an inequality index, a wide simulation study was conducted on a real set of data to assess the ability of different indices to detect perturbations in the income distribution. This empirical research has shown that the Zenga index as well as one of the new propositions deserve particular attention and that some well-known measures are too insensitive to account for changes.

In Chapter 3, finite population inference for the Quintile Share Ratio (QSR) is proposed. Existing methodology for this important measure is improved, allowing for straightforward point and variance estimates under complex sampling designs. Specifically, we show that the linearization approach to variance estimation does not require smoothing of probability density functions. The issue of constructing reliable confidence intervals is also addressed. This chapter is a reprint of Langel & Tillé (2011d).

Chapter 4 is a reprint of Langel & Tillé (2011a) and is entirely dedicated to the work of Corrado Gini, who has made important contributions both to inequality theory with the famous Gini index and to survey sampling with the beginnings of balanced sampling. The motivation of this chapter is to present and link both ideas. Balanced sampling is discussed involving the controversial notions of randomness and representativeness in survey sampling theory. Moreover, a simulation study shows the benefits of balancing the sample on the Gini index of an anterior year instead of on a linear constraint such as the income.

Because of its leading position in income inequality theory, the Gini index has given rise to a large amount of research, especially papers regarding inference. In Chapter 5, we propose to go through the development of variance estimation for the Gini index. Particular attention is given to understanding some misconceptions which have led to erroneous methods such as the regression approach. The fact that the same result has been published several times is also discussed. Finally, a new result in the linearization approach noticeably simplifies the computation of the linearized variable for the Gini index.

While Chapters 3 to 5 are dedicated to enhancing methodology for well established measures, Chapter 6 follows from our findings from the empirical study discussed in Chapter 2. Indeed, the latter emphasized the interesting behavior of the Zenga index, for which we propose a methodology for finite population inference. In this chapter, the Zenga index is also compared to the Gini index with respect to different aspects such as robustness and skewness of the sampling distribution. Chapter 6 is a reprint of Langel & Tillé (2011c).

The document ends with a general conclusion and further lines of research. In the Appendix section, linearized variables of most inequality measures presented along the document are derived. These results allow for straightforward variance estimation of these measures when estimated by means of a sample.

# 1 An introduction to the measurement of inequality and survey sampling

**Abstract**

This chapter proposes a short overview of definitions and notions both from the field of inequality measures and survey sampling. Section 1.1 is dedicated to the basics of income inequality. First, the income distribution is defined as a probability distribution and the Lorenz curve, an essential notion of the field, is discussed. The different approaches to estimating income distribution are then presented. In Section 1.2, important notions of survey sampling are introduced. The section starts with a brief discussion of the two main approaches to finite population sampling. The framework of design-based inference is then presented, along with well-known results regarding the estimation of linear functions of interest. The chapter ends with a short presentation of the data used throughout the document (Section 1.3).

**Keywords:** complex sampling, income distribution, inequality, Lorenz curve

## 1.1 Measuring inequality

### 1.1.1 Income distribution and the Lorenz curve

What is meant by income inequality is directly dependent on what is meant by income. Defining income is not an easy matter since income can be the result of a combination of different sources. Moreover, when the objective is to compare different finite populations (e.g. countries), one should make sure that the standards used to define income in the different populations are the same. In the following document, the actual composition or construction of the income variable is not discussed. Instead, emphasis is placed on the allocation of income among the recipients, which is summarized by the notion of income distribution.

Before considering the finite population framework, let us first describe briefly what measuring inequality means from a classical statistical standpoint. For a more detailed overview of the main concepts and approaches to income inequality theory, the reader may refer to Cowell (2000). We shall let $Y$ denote a continuous random variable, generally representing income. For the sake of simplicity, we assume $Y$ to be positive. This assumption is quite strong because in practice income can be null or negative and while adjustments can sometimes easily be made to incorporate non-positive values, some inequality measures can only deal with positive incomes. The income distribution can be viewed as the probability distribution, or probability density function $f(y)$ of random variable $Y$. Let now $F(y)$ denote the strictly increasing cumulative distribution function such that

$$F(y) = \int_0^y f(u)du = \Pr(Y \leq y).$$

The quantile function $Q(\alpha)$ is defined as the inverse of $F(.)$:

$$Q(\alpha) = F^{-1}(\alpha), \alpha \in (0,1).$$

Quantiles are widely used in inequality theory. They help to define a central tool of inequality theory, the Lorenz (1905) curve given by

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

where

$$\mu = \int_0^{\infty} yf(y)dy = \int_0^1 Q(p)dp.$$

It is therefore assumed that $\mu$, the mean of random variable $Y$, exists and is non-null. Empirically, the Lorenz curve can be seen as the share of total income earned altogether by the poorest $100\alpha\%$ in the population. Under the assumptions made on $Y$ and $\mu$, the Lorenz curve is both increasing and convex in $\alpha \in (0,1)$. Indeed,

$$\frac{dL(\alpha)}{d\alpha} = \frac{Q(\alpha)}{\mu} > 0,$$

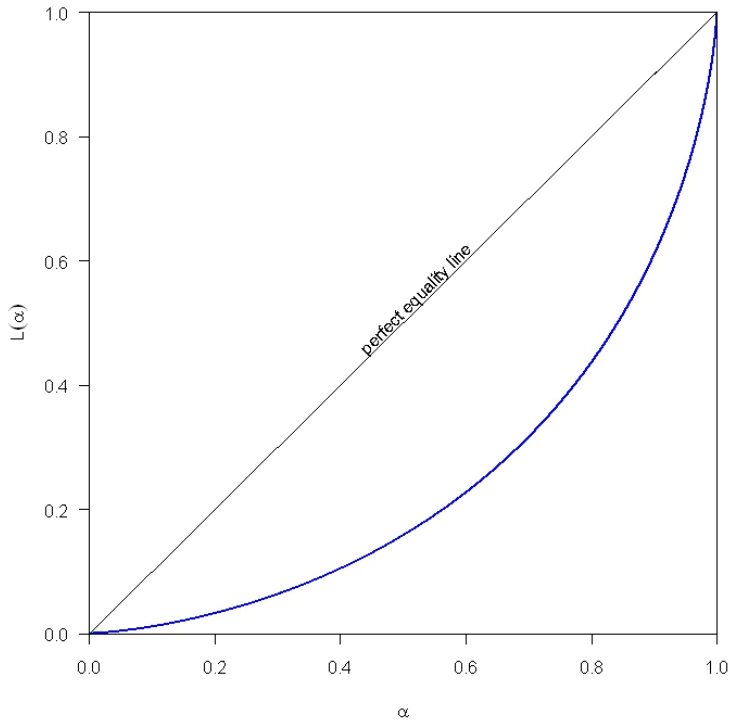
and

$$\frac{d^2L(\alpha)}{d^2\alpha} = \frac{1}{\mu f[Q(\alpha)]} > 0.$$

Many inequality measures, including the three measures detailed in the following chapters, can be expressed by means of the Lorenz curve. Graphically (see Figure 1.1), the Lorenz curve is usually compared with the $45^\circ$ perfect equality line, which itself can be viewed as the Lorenz Curve of a perfectly equal income distribution expressed by $L(\alpha) = \alpha$. In-

terpretation is straightforward: the closer the Lorenz curve is to the perfect equality line, the lower the level of inequality of the distribution.

Figure 1.1 – *Example of a Lorenz curve (data generated from a Log-normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$).*




The Lorenz curve is used abundantly in the following chapters when synthetic inequality measures such as the Gini index are discussed. As an illustration, we show that the Gini index can in fact be defined as twice the area between the perfect equality line and the Lorenz curve:

$$\mathrm{G} = 2\int_0^1 [\alpha - L(\alpha)]\, d\alpha = 1 - 2\int_0^1 L(\alpha) d\alpha.$$

### 1.1.2 Parametric and non-parametric approaches

In empirical applications the income distribution $f$ and the cumulative distribution function $F$ are of course unknown and must be estimated. Income inequality theory provides no consensus on how this estimation should be handled. The income distribution can be modelled or on the contrary, estimated in a non-parametric manner.

Income modelling involves fitting a particular distribution to the full empirical data (parametric approach) or solely to the tails of the data (semi-parametric approach). Expressions for the Lorenz curve and inequality measures can then be derived explicitly as functions of the parameters. Some frequently used distributions are the Dagum distributions,

the Generalized Beta family of distributions, Pareto distributions and others. Research on the parametric approach focuses on finding probability distributions that are capable of describing the complexity of income data through only a few parameters, deriving parametric expressions of known inequality measures and estimating these parameters. Kleiber & Kotz (2003), Chotikapanich (2008) and Graf et al. (2011) present comprehensive overviews as well as new developments on the subject.

On the contrary, the methodology proposed in this document is essentially non-parametric. Thus, no prior assumption on the shape of the income distribution is made and both Lorenz curves and inequality measures are estimated without referring to a model. Both approaches, of course, have advantages and drawbacks. Inference in the parametric approach is probably more convenient and, once the data is modelled, enables easy implementation of further analysis. Moreover, robustness issues can be reduced by assuming a model for the tails of the distribution where extreme values often occur and where reliable data is difficult to collect. The benefits of the non-parametric approach, on the other hand, are found in the absence of dependence to model assumptions such as restrictions on the shape of the income distribution. These restrictions can be serious as inequality measure estimates can be rather sensitive to the choice of the statistical distribution used to model the income structure (Chotikapanich & Griffiths, 2006).

## 1.2 Finite population sampling

### 1.2.1 Design-based and model-based approaches

In addition to being non-parametric with regards to the income distribution, the approach proposed in this document is design-based, meaning that the finite population is not regarded itself as a sample drawn from an infinite super-population model (model-based approach). The stochastic component in the design-based approach is thus entirely encompassed in the sampling design. Also, inference is performed with respect to the sampling design rather than to a model. Fundamental paradigmatic divergences between design- and model-based approaches to survey sampling are discussed in Hansen et al. (1983), Royall (1988), Brewer (1999), Little (2004) or Nedyalkova & Tillé (2008). Note that design-based inference can be *model-assisted* (Särndal et al., 1992). Design-based methods are sometimes viewed as the *traditional* approach to survey sampling. In official statistics, the design-based approach is often preferred because national statistical institutes want to avoid the risk of model misspecification. Official statistics producers prefer to rely on large sample asymptotic results rather than on model assumptions, which are seen as both more hazardous and more normative (or less impartial). This seems to be es-

pecially the case in European countries (see van den Brakel & Bethlehem, 2008), but although model-based techniques are more and more present in official statistics, for instance in small area estimation or non-response treatments, the predominance of design-based methods in national statistical institutes is not threatened (see Rao, 2011). Some definitions and basic notions of design-based inference are presented in the next section.

### 1.2.2 Definitions and notation

The goal of survey sampling is to use data collected from a fraction (a sample) of the finite population to perform inference on the latter. Sampling can be opposed to conducting a census, which involves collecting data among all population units. Obvious advantages of sampling over censuses are feasibility in terms of cost and time.

Formally, we call $U = \{1, \ldots, k, \ldots, N\}$ a finite population of $N$ units where $k$ denotes the unit's (unique) label and $N < \infty$. A sample without replacement $s \subset U$ is a non-empty subset of the finite population. Let $\mathcal{S}$ denote the collection of all possible samples from $U$.

**Definition 1.1** *A sampling design is a pair $(\mathcal{S}, p)$ where $p(.)$ is the probability distribution on $\mathcal{S}$ of a random variable $S$ such that $\Pr(S = s) = p(s) \geq 0, s \in \mathcal{S}$ and $\sum_{s \in \mathcal{S}} p(s) = 1$.*

The size of sample $s$ is denoted $n(s)$. A sampling design is of fixed size if $\text{var}[n(s)] = 0$ and the shorter notation $n$ may then be used to denote the size of the sample. From the sampling design, the first order inclusion probability of unit $k$, i.e. the probability that unit $k$ is selected into the sample, can be deduced by

$$\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s) \quad \text{for all} \quad k \in U.$$

Second order inclusion probabilities are defined by

$$\pi_{k\ell} = \Pr(k \in S, \ell \in S) = \sum_{s \ni k,\ell} p(s) \quad \text{for all} \quad k, \ell \in U, k \neq \ell.$$

In the following chapters, empirical illustrations will often be given with simple random sampling without replacement designs (*srswor*) although the methodology is directly applicable to more complex sampling designs.

**Definition 1.2** *A simple random sampling design without replacement (*srswor*) is a design of fixed size in which the probability of selecting sample $s$ is*

$$p(s) = \binom{N}{n}^{-1},$$

*if $s$ is of size $n$ and $p(s) = 0$ otherwise.*

Moreover, the first order inclusion probability of unit $k$ in an *srswor* design is $\pi_k = n/N$ for all $k \in U$. In design-based survey sampling theory, it is assumed that for all units $k \in U$ a particular characteristic of interest $y$ is observable and unknown. The value taken by unit $k$ on the characteristic of interest is denoted $y_k$. In the design-based approach, $y_k$ is not the realization of a random variable but a fixed quantity. In this document, the character of interest is generally income.

The purpose of a sampling procedure is to observe $y_k$ for all sampled units in order to estimate a function of interest (or finite population parameter) $\theta$ which is a function of the characteristic of interest:

$$\theta = \theta(y_k, k \in U).$$

In many applications, one is interested in estimating the population total $Y = \sum_{k \in U} y_k$ or a function of the total, e.g. the mean. The function of interest $\theta$ may nevertheless be a more complex quantity. In a situation where the characteristic of interest is income, one certainly wants to estimate the total income or the mean income of a finite population, but might as well be interested in estimating quantiles (e.g. the median income) or inequality measures such as the Gini index. An estimator of $\theta$ is a statistic denoted by $\widehat{\theta}$. Because for all $k \in U$ quantities $y_k$ are fixed, the value $\widehat{\theta}$ is only dependent upon the sampling design. Thus, an objective of finite population design-based inference is to propose an estimator of $\theta$ that is design-unbiased or approximately design-unbiased. Design-unbiasedness is achieved when the expectation with respect to the sampling design of $\widehat{\theta}$ is equal to $\theta$, i.e. when

$$\mathrm{E}(\widehat{\theta}) = \sum_{s \in \mathcal{S}} p(s)\widehat{\theta}(s) = \theta.$$

Also, the variance of $\widehat{\theta}$ with respect to the sampling design is defined by

$$\mathrm{var}(\widehat{\theta}) = \sum_{s \in \mathcal{S}} p(s)[\widehat{\theta}(s) - \mathrm{E}(\widehat{\theta})]^2.$$

Variance estimation is one of the most important topics in survey sampling because variance estimators are needed in order to evaluate accuracy of the point estimation and construct confidence intervals. When $\theta$ is a complex, non-linear function of interest, estimating $\mathrm{var}(\widehat{\theta})$ is not straightforward. Indeed, the sampling variance of a statistic depends on both the complexity of the statistic and the complexity of the sampling design. Wolter (2007, p.2) splits survey sampling theory into four categories as in Table 1.1. The simplest cases are regrouped in category *a*. This document focuses on proposing methodologies that are applicable to case *d*, although most empirical illustrations in the following chapters are done on case *c*. In both cases, the function of interest is complex and specific

methods are needed for variance estimation. However, results from cases *a* and *b* turn out to be useful for the purpose of estimating non-linear functions of interest.

Table 1.1 – *Four levels of complexity in survey sampling (Wolter, 2007, p.2)*

| | simple design | complex design |
|---|---|---|
| linear estimators | *a* | *b* |
| non-linear estimators | *c* | *d* |

As an illustration, we show below some well-known results for the estimation of a total $Y$. Each result is presented in its general form (case *b* of Table 1.1), as well as in the specific case of an *srswor* design of size $n$ (case *a* of Table 1.1). A well-known estimator of $Y = \sum_{k \in U} y_k$ is the Horvitz & Thompson (1952) estimator

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

The Horvitz-Thompson estimator is design-unbiased regardless of the sampling design as long as $\pi_k > 0$ for all $k \in U$. In the Horvitz-Thompson estimator, each sampled unit is weighted by the inverse of its inclusion probability. For a *srswor* design, $\pi_k = n/N$ for all $k \in U$ and

$$\widehat{Y}_\pi = \frac{N}{n} \sum_{k \in S} y_k.$$

If $\pi_k > 0$ for all $k \in U$, the variance of $\widehat{Y}_\pi$ involves second-order inclusion probabilities and is given by

$$\text{var}(\widehat{Y}_\pi) = \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{y_k y_\ell}{\pi_k \pi_\ell}.$$

Similarly, $\text{var}(\widehat{Y}_\pi)$ can be estimated by the Horvitz & Thompson (1952) variance estimator

$$\widehat{\text{var}}(\widehat{Y}_\pi) = \sum_{k \in S} \frac{y_k^2}{\pi_k^2}(1 - \pi_k) + \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}}. \tag{1.1}$$

Expression (1.1) is unbiased if $\pi_{k\ell} > 0$ for all $k, \ell \in U$, but can unfortunately take negative values. For fixed size designs, Sen (1953), Yates & Grundy (1953) have proposed another expression for the variance and variance estimator of $\widehat{Y}_\pi$. The Sen-Yates-Grundy estimator is given by

$$\widehat{\text{var}}_{SYG}(\widehat{Y}_\pi) = \frac{1}{2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \left( \frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}.$$

This estimator is unbiased for designs of fixed size and is positive when $\pi_k\pi_\ell - \pi_{k\ell} \geq 0$ for all $k, \ell \in U, k \neq \ell$. Under an *srswor* design, the variance is expressed by

$$\text{var}(\widehat{Y}_\pi) = \frac{N(N-n)}{n(N-1)} \sum_{k \in U} (y_k - \overline{Y})^2,$$

where $\overline{Y} = N^{-1} \sum_{k \in U} y_k$ and the estimator simplifies to

$$\widehat{\text{var}}(\widehat{Y}_\pi) = \frac{N(N-n)}{n(n-1)} \sum_{k \in S} (y_k - \bar{y})^2,$$

where $\bar{y} = n^{-1} \sum_{k \in S} y_k$ is the Horvitz-Thompson estimator of the mean for the *srswor* design.

### 1.2.3 Auxiliary information

Commonly, the survey practitioner has some knowledge of the structure of the finite population. For example, a census may be available and include demographic information on every unit in the population. In survey sampling, this type of information is referred to as auxiliary information and can be used in order to improve the accuracy of estimation. A very large share of survey sampling research focuses on the use of auxiliary information (see for example Särndal et al., 1992; Kish, 1995; Tillé, 2001). Auxiliary variables can be both quantitative or qualitative. Formally, let $x_1, ..., x_j, ...x_p$, denote a set of $p$ auxiliary variables and let $x_{kj}$ denote the value taken by unit $k \in U$ on auxiliary variable $x_j$. The population total of $x_j$ is defined by

$$X_j = \sum_{k \in U} x_{kj}.$$

While quantity $x_{kj}$ is sometimes known for all $k \in U$, total $X_j$ may also be the only available information for variable $x_j$. Auxiliary information can be used both at the design phase and at the estimation phase, applying methods such as unequal probability sampling, stratification, balanced sampling, multi-stage designs, calibration and others. A sound use of such information can greatly enhance the quality of estimation in comparison with applying a more naive sampling strategy.

## 1.3 Income data sets

Most of the empirical applications in the following chapters use real sets of income data from the Canton of Neuchâtel, Switzerland, for recent years, namely years 2005 through 2007. We are thankful to the *Service statistique* of the Canton of Neuchâtel for making this large set of data available. In this data set, the characteristic of interest used as the income is in fact the

taxable income. Unfortunately, it is not always collected at the individual level, because the taxing structure is such that, for instance, the income of a married couple is taxed all at once and dependent children are not directly taxed. Thus, the statistical units are income earners which may represent an individual, a family or other and should not be mistaken for households as one household can contain any number of income earners. In 2005 for instance, around 74,800 households (Swiss Federal Statistical Office, 2008) were recorded in the Canton of Neuchâtel while, in the same year 96,575 tax returns (including returns with null incomes) were recorded in our data set.

It should be mentioned that this research is by no means a study of the level of inequality in the Canton of Neuchâtel. The data has been made available as a training data set for academic purposes only. No results of estimates or true values of inequality measures are printed in this document. Moreover, no descriptive statistics such as mean, median or extreme values are disclosed. Instead, the data is treated as a full finite population from which samples can be drawn repeatedly in a simulation setup. The results are evaluated in terms of quality of estimation, mainly using relative biases. Note also that only non-null income earners are kept in the data for analysis. Moreover, since only very little auxiliary information is available, empirical applications in the following chapters are mostly done with simple random sampling designs.

The advantages of having such a data set to test the proposed methodology are twofold. Firstly, one can compare sample estimates to true finite population parameters, since the latter can be computed on the whole finite population. Secondly, the fact that the data is not artificial is a huge asset when working with income data because extreme observations are well captured. It is not an easy task to construct an artificial yet credible income structure and the level of skewness as well as the presence of extreme values are often undervalued in artificial data sets. As an illustration, a comparison of the quality of inference for the Zenga inequality index for both artificial and real data is proposed in Chapter 6. Clearly, because it is a topic where extreme observations are of great concern, income inequality measurement benefits greatly from being placed in a realistic framework.

# 2

# An evaluation of the performance of inequality measures for the detection of changes in an income distribution

**Abstract**

A simulation study on real data is performed in order to evaluate the capacity of seventeen inequality indices to detect changes in an income distribution. First, a finite population expression is proposed for each measure. The usual axioms and properties of inequality measures are then presented and discussed. Simulation results reveal interesting characteristics of indices based on quantile shares as well as new or recent measures such as the Average Share Ratio (ASR) and the Zenga index. The performance of the two Laeken indicators (Gini and QSR) is also discussed.

**Keywords:** income, inequality measures, Gini, QSR, simulations

## 2.1 Introduction

Dozens of indicators that measure the level of inequality of income or capital earners have been proposed in the literature. The most well-known are probably the Gini index, the Quintile Share Ratio (QSR), the Atkinson index and the Theil index. These indices and their intrinsic properties have been largely studied (see for example Cowell, 1977; Coulter, 1989; Silber, 1999; Atkinson & Bourguignon, 2000; Cowell, 2003).

An important property that has not yet been studied in a comparative study among a large set of inequality measures is the ability of an

index to detect a change in the income distribution. This feature is crucial when one wants to evaluate the increase or decrease of income inequality through time, or the effects of a given political or social action on inequality.

In this paper, we present a simulation study performed on the income distribution of the canton of Neuchâtel, Switzerland to evaluate seventeen different indices, and how they react to perturbations in the distribution. Besides the most common measures, some new or recent indices are also tested. The two Laeken inequality indicators, The Gini index and the QSR, are naturally included in the simulations. Laeken indicators are a set of European statistical indicators on poverty, inequality and social exclusion, established at the European Council in December 2001 in Laeken, Belgium in order to allow for comparisons among European countries. These measures are part of the EU-SILC (European Union Statistics on Income and Living Conditions) program of Eurostat (Eurostat, 2005).

Although a meticulous analysis of the properties of the proposed measures can give good insights on their reaction to changes in the income distribution, we believe that a comparative empirical study on real data can be of major interest to complement a more theoretical axiomatic approach. Moreover, measuring inequality involves dealing with income distributions which have often untypical densities with very heavy skewness issues, extreme values as well as potential clusters that could result from social policies such as minimum wage or social welfare. Thus, an exploratory and detailed analysis on real data seems particularly useful.

In order to evaluate the general performances of the whole set of indices for the detection of changes in the distribution, four different types of perturbations have been affected to the distribution : changes implying the whole distribution and changes concerning only a part of the distribution (high incomes, low incomes and intermediate incomes respectively). The reaction of each inequality measure to each perturbation is then established. Particular attention is paid to Laeken indicators.

In the next section, a finite population expression for all the indices used in the simulation study is presented. The presentations precede a succinct discussion on the properties of the indices, related to the axiomatic approach to inequality theory (Section 2.3). For most measures, detailed analysis have already been successfully performed (Sen, 1973; Coulter, 1989; Silber, 1999). Section 2.4 details the different perturbations that are applied to the distribution in the simulation setup and Section 2.5 presents the simulation method. The article ends with a discussion of the simulation results and a conclusion with suggestions for subsequent research.

## 2.2 Inequality indices

With the impressive number of different indices existing in the literature on inequality theory, it is not easy to draw up an exhaustive list of these measures. However, it is possible to outline different types of indices and group them with respect to the paradigm on which they are based. Also, it is important to observe that different type of indices measure different aspects of income inequality. For this simulation case, we have focused on seventeen indices containing some of the most widely used inequality measures, as well as new measures. They are detailed below in expressions (2.1) to (2.17) for an income distribution from a finite population $U$ of size $N$. The income of unit $k$ is defined as $y_k$, and the distribution is assumed to be ordered. To lighten the notation, we assume with no loss of generality that all $y_k$'s are distinct, and that the rank of income $y_k$ can be expressed simply by $k$. Let $Q_\alpha$ denote the quantile of order $\alpha$. Let us also consider:

$$Y = \sum_{k \in U} y_k,$$

$$\overline{Y} = \frac{Y}{N},$$

$$\text{var}(y) = \frac{1}{N} \sum_{k \in U} (y_k - \overline{Y})^2,$$

$$Y_\alpha = \sum_{k \in U} y_k \mathbb{1}\left(y_k \leq Q_\alpha\right),$$

where $\mathbb{1}(A) = 1$ if $A$ is true and 0 otherwise. Finally, let quantile $Q_\alpha$ be defined for a finite population by linear interpolation (Hyndman & Fan, 1996):

$$Q_\alpha = y_{k-1} + (y_k - y_{k-1}) \left[\alpha N - (k-1)\right],$$

where $\alpha N < k \leq \alpha N + 1$.

### 2.2.1 Gini index

The Gini (1921) index is by far the most popular measure of inequality. It has motivated a great amount of literature (see for example Shalit, 1985; Cowell, 2000; Xu, 2004; Berger, 2008) and is also one of the Laeken indicators. In the finite expression below, the index can take values between 0 (perfect equality) and $(N-1)/N$ (perfect inequality):

$$\text{G} = \frac{2}{NY} \sum_{k \in U} k y_k - \frac{N+1}{N}. \tag{2.1}$$

### 2.2.2 Coefficient of variation

As a measure of relative dispersion, the coefficient of variation may also be used as a simple inequality index. Perfect equality is achieved when CV$=0$ which implies $\text{var}(y)=0$, and large values of CV denote a high level of inequality:

$$\text{CV}=\frac{\sqrt{\text{var}(y)}}{\overline{Y}}. \tag{2.2}$$

### 2.2.3 Atkinson index

The Atkinson (1970) indices are a widely used class of inequality measure (see also Allison, 1978), defined by the general expression:

$$\text{A}_\epsilon=\begin{cases} 1-\left[\frac{1}{N}\sum_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^{1-\epsilon}\right]^{\frac{1}{1-\epsilon}}, & \text{if} \quad \epsilon\geq 0, \epsilon\neq 1, \\ 1-\prod_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^{\frac{1}{N}}, & \text{if} \quad \epsilon=1. \end{cases}$$

Using the generalized mean,

$$M_\epsilon=\begin{cases} \sqrt[\epsilon]{\frac{1}{N}\sum_{k\in U}y_k^\epsilon}, & \text{if } \epsilon\neq 0, \\ \left(\prod_{k\in U}y_k\right)^{\frac{1}{N}}, & \text{if } \epsilon=0, \end{cases}$$

the Atkinson index can also be written

$$\text{A}_\epsilon=1-\frac{M_{1-\epsilon}}{M_1},$$

with $\epsilon\geq 0$. The Atkinson index is thus a function of the income distribution and of a parameter $\epsilon$. The latter is defined as an inequality aversion parameter. As the value of $\epsilon$ increases, the index becomes less sensitive to transfers at the top of the distribution. The choice of the value of $\epsilon$ implies a normative dimension to the measurement of inequality and has been discussed in the literature (Atkinson, 1970; Moore, 1996; Lambert et al., 2003). In this study, the measure has been tested using $\epsilon=1/2$ and $\epsilon=1$:

$$\text{A}_{1/2}=1-\frac{\left(\frac{1}{N}\sum_{k\in U}\sqrt{y_k}\right)^2}{\overline{Y}}, \tag{2.3}$$

$$\text{A}_1=1-\prod_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^{\frac{1}{N}}. \tag{2.4}$$

Both $\text{A}_{1/2}$ and $\text{A}_1$ take value 0 in case of perfect equality and the highest inequality level is achieved when $\text{A}_{1/2}=(1-1/N)$ and $\text{A}_{1/2}=1$ respectively.

### 2.2.4 Generalized entropy index

The entropy-based indices form yet another category of inequality measures. The best-known measure in this class is the Theil (1967) index T. The latter is a specific case of both the Generalized entropy measure and the Rényi divergence measure (see Section 2.2.5). The Generalized entropy index can be expressed as (Shorrocks, 1980; Mussard et al., 2003; Cowell, 2006):

$$\mathrm{GE}_\varphi = \begin{cases} \dfrac{1}{\varphi(\varphi-1)}\left[\dfrac{1}{N}\displaystyle\sum_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^\varphi - 1\right], & \text{if} \quad \varphi \neq 0, \varphi \neq 1, \\ \dfrac{1}{N}\displaystyle\sum_{k\in U}\log\frac{Y}{Ny_k}, & \text{if} \quad \varphi = 0. \\ \displaystyle\sum_{k\in U}\frac{y_k}{Y}\log\frac{Ny_k}{Y}, & \text{if} \quad \varphi = 1. \end{cases}$$

As for $\epsilon$ in the Atkinson index, the sensitivity of the index to either tails of the income distribution can be tuned with parameter $\varphi$. Note however, that contrarily to $\epsilon$, higher values of $\varphi$ increase the sensitivity at the top of the distribution. We have tested three indices in this class, using parameter $\varphi = 1/2$, 1 and 2:

$$\mathrm{GE}_{1/2} = 4\left(1 - \sum_{k\in U}\sqrt{\frac{y_k}{NY}}\right), \tag{2.5}$$

$$\mathrm{GE}_1 = \mathrm{T} = \sum_{k\in U}\frac{y_k}{Y}\log\frac{Ny_k}{Y}, \tag{2.6}$$

$$\mathrm{GE}_2 = \frac{1}{2N}\sum_{k\in U}\frac{y_k^2}{\overline{Y}^2} - \frac{1}{2}. \tag{2.7}$$

The first case ($\mathrm{GE}_{1/2}$) is closely related to the Hellinger Distance (Gibbs & Su, 2002), differing only by a constant, the second case (T) is simply the Theil index (Theil, 1967; Coulter, 1989) also known as the Kullback & Leibler (1951) distance in information theory, whereas the third case ($\mathrm{GE}_2$) is similar to the $\chi^2$ Distance (Lindsay, 1994), also differing only by a constant, and to the Herfindahl-Hirschman index (Hirschman, 1964; Acar & Sankaran, 1999), one of the most popular measure for industrial concentration. All three measures above take value 0 when the income is equally distributed and have respective upper bounds of $4(1-1/\sqrt{N})$, $\log N$ and $(N-1)/2$.

### 2.2.5 Rényi divergence measure

The Rényi (1961) divergence measure is defined by:

$$R_\varphi = \begin{cases} \dfrac{1}{\varphi - 1} \log \left[ N^{\varphi - 1} \displaystyle\sum_{k \in U} \left( \frac{y_k}{Y} \right)^\varphi \right], & \text{if} \quad \varphi \neq 1, \\ \displaystyle\sum_{k \in U} \frac{y_k}{Y} \log \frac{N y_k}{Y}, & \text{if} \quad \varphi = 1. \end{cases}$$

As specified above, the Theil index (2.6) or Kullback & Leibler (1951) distance can also be expressed as a specific case of the Rényi Divergence with $\varphi = 1$. For this simulation study, two other Rényi divergence measures have been used, $\varphi = 1/2$ and $\varphi = 2$:

$$R_{1/2} = -2 \log \sum_{k \in U} \sqrt{\frac{y_k}{NY}}, \tag{2.8}$$

$$R_2 = \log \left( N \sum_{k \in U} \frac{y_k^2}{Y^2} \right). \tag{2.9}$$

Both $R_{1/2}$ and $R_2$ take values in the interval $[0; \log N]$, a small value indicating a low level of inequality. The role of the tuning parameter $\varphi$ is similar to that of the Generalized entropy index.

### 2.2.6 Zenga index

Zenga (2007)'s new inequality index, not to be mistaken for other measures proposed earlier in Zenga (1984), is based on the ratio between lower and upper arithmetic means. A finite population expression is given in Zenga (2007) by

$$Z = 1 - \frac{1}{N} \sum_{\ell \in U} \frac{\overline{Y}_\ell^-}{\overline{Y}_\ell^+}, \tag{2.10}$$

where

$$\overline{Y}_\ell^- = \frac{1}{\ell} \sum_{k \in U} y_k \mathbb{1}(k \leq \ell) \quad \text{and} \quad \overline{Y}_\ell^+ = \frac{1}{N - \ell + 1} \sum_{k \in U} y_k \mathbb{1}(k \geq \ell).$$

The highest level of inequality is reached when $Z = (1 - 1/N^2)$, while perfect equality is achieved for $Z = 0$.

### 2.2.7 Quantile share ratios

The quantile share ratios constitute another important class of inequality measures. Their construction is simple in the continuous case but requires a definition for the finite population quantile in the discrete case. They are expressed as the ratio of the total income earned by observations *above*

quantile of order $(1-\alpha)$ to that earned by observations *below* quantile of order $\alpha$. Thus, a general finite population expression is given by:

$$\text{SR} = \frac{Y - Y_{1-\alpha}}{Y_{\alpha}}.$$

The case where $\alpha = 0.2$, is called the Quintile Share Ratio (QSR) and is the most common example in this class of measures as it is part of the set of Laeken indicators (Leiten & Traat, 2005; Osier, 2006), but other similar measures can be computed. In this paper, the Decile Share Ratio (DSR, with $\alpha = 0.1$) and the Median Share Ratio (MSR with $\alpha = 0.5$) are also presented:

$$\text{QSR} = \frac{Y - Y_{0.8}}{Y_{0.2}}, \tag{2.11}$$

$$\text{DSR} = \frac{Y - Y_{0.9}}{Y_{0.1}}, \tag{2.12}$$

$$\text{MSR} = \frac{Y - Y_{0.5}}{Y_{0.5}}. \tag{2.13}$$

A new measure proposed here is the Average Share Ratio (ASR), which is defined as the harmonic mean of all quantile share ratios with $\alpha \leq 0.5$. A finite population expression for the ASR is given by

$$\text{ASR} = \begin{cases} \dfrac{N}{2}\left[\displaystyle\sum_{\ell=1}^{N/2} \frac{\sum_{k\in U} y_k \mathbb{1}(k \leq \ell)}{\sum_{k\in U} y_k \mathbb{1}(k > N-\ell)}\right]^{-1}, & \text{if } N \text{ is even,} \\ \dfrac{N+1}{2}\left[\displaystyle\sum_{\ell=1}^{(N+1)/2} \frac{\sum_{k\in U} y_k \mathbb{1}(k \leq \ell)}{\sum_{k\in U} y_k \mathbb{1}(k > N-\ell)}\right]^{-1}, & \text{if } N \text{ is odd.} \end{cases} \tag{2.14}$$

By using the harmonic mean, the ASR allows for the presence of null incomes. Expressions (2.11), (2.12), (2.13) and (2.14) take 1 as a minimum value when all $y_k$'s are equal. They have no upper bound.

### 2.2.8 Interquantile ratios

The final type of indices used in this paper regroups all the interquantile ratios. This class has to be clearly distinguished from the quantile share ratios class in Section 2.2.7. An interquantile ratio is a plain ratio of two quantiles of order $(1-\alpha)$ and $\alpha$ repsectively. A general finite population expression is given by

$$\text{QR} = \frac{Q_{1-\alpha}}{Q_{\alpha}},$$

with $\alpha \leq 0.5$. The particular cases using quartiles (IQR) and deciles (IDR) are expressed here because of their frequent occurrence in applications (for example Stoetzel, 1980; Merle & Andreyev, 2002):

$$\text{IQR} = \frac{Q_{0.75}}{Q_{0.25}}, \tag{2.15}$$

$$\text{IDR} = \frac{Q_{0.9}}{Q_{0.1}}. \tag{2.16}$$

In addition, we propose hereafter a new measure, the Mean Quantile Ratio (MQR) which is defined as the harmonic mean of all interquantile ratios:

$$\text{MQR} = \begin{cases} \dfrac{N}{2}\left(\displaystyle\sum_{k=1}^{N/2}\frac{y_k}{y_{N-k+1}}\right)^{-1}, & \text{if } N \text{ is even,} \\ \dfrac{N+1}{2}\left(\displaystyle\sum_{k=1}^{(N+1)/2}\frac{y_k}{y_{N-k+1}}\right)^{-1}, & \text{if } N \text{ is odd.} \end{cases} \tag{2.17}$$

Expressions (2.15), (2.16) and (2.17) take 1 as a minimum value when all $y_k$'s are equal. They have no upper bound.

## 2.3 Properties of inequality measures

### 2.3.1 The axiomatic approach to inequality measurement

A large portion of the literature on income inequality places emphasis on a series of principles (mostly axioms and properties) which define the backbone of inequality measures. Most of these principles have been extensively discussed (see for example Pigou, 1912; Dalton, 1920; Sen, 1973; Shorrocks, 1980, 1984; Cowell & Kuga, 1981b; Cowell, 1988, 2000, 2006; Zheng, 1994). Although this study is essentially empirical, it is of interest to know what these principles are and for which measures they are fulfilled. Moreover, a better understanding of the performance of each index in the simulation study can be obtained by analyzing these properties.

In the following, an inequality measure $I$ denotes a real valued continuous function defined on the set of all possible income distributions. $I(\mathbf{y})$ is the value of $I$ computed on the $N$-component income vector $\mathbf{y} = (y_1, ..., y_k, ..., y_N)$. Moreover, the polarization of inequality measure $I$ is assumed to be such that if $\mathbf{y}_A$ and $\mathbf{y}_B$ are the respective income vectors of two finite populations $A$ and $B$, the relation $I(\mathbf{y}_A) > I(\mathbf{y}_B)$ implies that population $A$ is more unequal than population $B$ with respect to index $I$.

**Definition 2.1** *(Anonimity) An inequality measure $I$ satisfies the anonymity principle if $I(\mathbf{y}) = I(\tilde{\mathbf{y}})$ where $\tilde{\mathbf{y}}$ denotes any permutation of vector $\mathbf{y}$.*

The anonymity principle is important but very straightforward. It implies that the level of inequality does not depend on who earns the income. This property is fulfilled by all measures proposed in Section 2.2. However, this principle may not be satisfied when the character of interest is multivariate (Cowell, 2000).

**Definition 2.2** *(Normalization) An inequality measure $I$ satisfies the normalization principle if $I(\bar{\mathbf{y}}) = 0$, where $\bar{\mathbf{y}} = \overline{Y}\mathbf{1}$ and where $\mathbf{1}$ is an $N$-component vector of 1's.*

A normalized inequality measure yields an inequality level of 0 when income is perfectly equally distributed among units in the population. This principle is violated by interquantile ratios and quantile share ratios.

**Definition 2.3** *(Population principle) An inequality measure $I$ satisfies the population principle if $I(\mathbf{y}) = I(\mathbf{y}_\lambda)$ where $\mathbf{y}_\lambda$ is a $\lambda N$-component vector denoting the $\lambda$-fold replication of vector $\mathbf{y}$, $\lambda \in \mathbb{N}$.*

The population principle implies invariance of the inequality level to a replication of the finite population. For instance, if two identical populations are aggregated, the inequality level obtained on the aggregated population will be the same as for the initial population. This principle may be violated by measures that involve finite population quantiles, depending upon the chosen finite population expression of the latter. ASR and MQR as expressed in (2.14) and (2.17) respectively, do not strictly satisfy the population principle if $N$ is odd.

**Definition 2.4** *(Principle of transfers) An inequality measure $I$ satisfies the principle of transfers (Pigou, 1912; Dalton, 1920) if $I(\mathbf{y}) \geq I(\tilde{\mathbf{y}})$ where*

$$\tilde{\mathbf{y}} = (y_1, ..., y_k + \delta, ..., y_\ell - \delta, ..., y_N),$$

*with $y_k \leq y_\ell$ and $0 < \delta < (y_\ell - y_k)/2$.*

The principle of transfers thus states a simple idea: an inequality measure should report a decrease in the level of inequality arising from an income transfer from a richer to a poorer unit as long as the transfer does not result in the poorer becoming the richer.

The Gini index, Coefficient of variation, Atkinson index, Zenga index, Generalized entropy index and Rényi divergence satisfy the principle of transfers. On the other hand, the principle can be violated by interquantile ratios and quantile share ratios. Indeed, some transfers are simply not detected by these indices, either because they occur out of the scope of the index (for example a transfer between two units near the median does not affect IQR) or because their effect is cancelled out (for example a transfer between two units belonging to the same quantile share does not affect QSR).

The principle of transfers has been studied widely and has prompted

the development of other criteria, such as the principle of diminishing transfers (Kolm, 1999) which states that the higher the income difference between two actors prior to the transfer, the larger the decrease in inequality. The principle of transfers can also be found in its weak form which suggests that the transfer results in a non increase of the inequality level. Moreover, in the Atkinson index, the Generalized entropy index and the Rényi divergence, the sensitivity of different types of transfers can be tuned by using different values of parameters $\epsilon$ and $\varphi$.

**Definition 2.5** *(Scale invariance) An inequality measure $I$ is scale invariant (or homogeneous of degree 0) if $I(\lambda \mathbf{y}) = I(\mathbf{y})$ with $\lambda \in \mathbb{R}_+^*$.*

Scale invariance allows for a measure to be unit-free, which is important when income is measured in several different currencies. All measures proposed in this paper are unit-free and satisfy this property. The class of indices satisfying scale invariance are usually denoted as *relative* inequality measures to distinguish them from the class of translation invariant indices denoted as *absolute* inequality measures (Chakravarty, 1999).

**Definition 2.6** *(Translation invariance) An inequality measure $I$ is translation invariant if $I(\mathbf{y} + b\mathbf{1}) = I(\mathbf{y})$ with $b \in \mathbb{R}$.*

Absolute inequality measures are not part of this study, but the debate on whether inequality measures should be scale- or translation-invariant is a long-lasting one (Kolm, 1976a,b; Foster & Shorrocks, 1991; Bosmans & Cowell, 2010). Zheng (1994) has shown that if $I$ satisfies the principle of transfers, $I$ cannot be both relative and absolute. All measures proposed in Section 2.2 are thus not translation invariant. On the contrary, they are all affected by translation since adding a positive constant to all incomes decreases the level of inequality.

In economic literature, Definitions 2.1 to 2.5 are often considered as axioms forming the basis of what an appropriate inequality measure should be. This approach excludes quantile share ratios as well as interquantile ratios. Hence, these measures have gained less attention than measures such as the Gini index, Atkinson index or the Generalized entropy index. Measures that are beyond the scope of this axiomatic approach to inequality are nonetheless useful in terms of interpretability and, in some cases, robustness. The fact that the QSR was chosen as the primary income inequality indicator by the European Council in 2001 (Eurostat, 2005) demonstrates that the axiomatic approach is not the only paradigm to inequality measurement.

### 2.3.2 Decomposability

Finally, an appreciated property for an inequality measure is additive decomposability between subgroups. An additive decomposable measure can be expressed by an addition of two separate terms. The first (within-subgroup) term is a weighted sum of the inequality index computed on each subgroup. The second term assesses the inequality level between subgroups.

Bourguignon (1979) and Shorrocks (1980) have shown that the only indices simultaneously allowing for direct additive decomposability and the other above properties (translation invariance excepted) may be found in the Generalized entropy class. However, the Atkinson index has been shown to be decomposable in a multiplicative manner (Das & Parikh, 1981; Blackorby et al., 1981; Dayioglu & Baslevent, 2006) and the decomposition of the Gini index has been widely studied (Bhattacharya & Mahalanobis, 1967; Pyatt, 1976; Cowell & Kuga, 1981a; Dagum, 1997; Yitzhaki, 1994, 1998). Recently, decomposition of the Zenga index has also been tackled (Radaelli, 2010).

## 2.4 Perturbations of the distribution

### 2.4.1 The data

The power to detect changes was studied in a simulation study on the seventeen measures presented in Section 2.2 for nine different situations, involving changes at different levels of the income distribution: at the top, at the bottom, at the center and among all units. The data set used for the simulation study is taxable income distribution for the Canton of Neuchâtel, Switzerland in 2007. It contains $N = 88,011$ non null income earners, with incomes expressed in Swiss francs (CHF) and is denoted hereafter as population $A$. Population $A$ is heavily skewed and contains very extreme observations (see Section 1.3 for more details on the data).

In the simulations, samples are drawn from the original population $A$ and from population $B$ which is a modified version of population $A$, also of size $N$. The nine different modifications (or perturbations) applied to $A$ in order to obtain $B$ are detailed below. Incomes in population $A$ and $B$ are respectively denoted $y_k^A$ and $y_k^B$ and $Q_\alpha^A$ stands for the quantile of order $\alpha$ in population $A$.

### 2.4.2 Changes at the top of the distribution

1. *High incomes set to an upper limit.* For this perturbation, incomes of the richer 10% have been levelled off to the ninth decile $Q^A_{0.9}$:

$$y^B_k = \begin{cases} Q^A_{0.9}, & \text{if } y^A_k \geq Q^A_{0.9}, \\ y^A_k, & \text{otherwise.} \end{cases}$$

2. *Doubling very high incomes.* The income of the richest 5% is doubled:

$$y^B_k = \begin{cases} 2y^A_k, & \text{if } y^A_k \geq Q^A_{0.95}, \\ y^A_k, & \text{otherwise.} \end{cases}$$

3. *Dividing very high incomes by two.* The income of the richest 5% is divided by two:

$$y^B_k = \begin{cases} \frac{1}{2}y^A_k, & \text{if } y^A_k \geq Q^A_{0.95}, \\ y^A_k, & \text{otherwise.} \end{cases}$$

### 2.4.3 Changes at all levels of the distribution

4. *Translation.* A fixed additional income of $10,000$ (CHF) has been attributed to every observation:

$$y^B_k = y^A_k + 10000, k \in U.$$

5. *Income after taxation.* One important question of income inequality consists in studying the effect of taxation on the inequality level. The amount of tax $t_k$ owed by taxpayer $k \in U$ is available in the data. Thus the distribution of income after taxation is obtained simply by

$$y^B_k = y^A_k - t_k, k \in U.$$

### 2.4.4 Changes at the bottom of the distribution

6. *Guaranteed minimum income.* Another relevant issue of social policies is the notion of a guaranteed minimum income. For the present simulation study, the minimum income was set to 30% of median income and allocated to all units under this threshold:

$$y^B_k = \begin{cases} 0.3Q^A_{0.5}, & \text{if } y^A_k \leq 0.3Q^A_{0.5}, \\ y^A_k, & \text{otherwise.} \end{cases}$$

7. *Increasing very low incomes.* Due to the low level of the small, non-null incomes, a heavy perturbation of the data is necessary here to induce

significant changes in the values of the different indices. Therefore, the income of the poorest 5% have been multiplied by 10:

$$y_k^B = \begin{cases} 10y_k^A, & \text{if } y_k^A \leq Q_{0.05}^A, \\ y_k^A, & \text{otherwise.} \end{cases}$$

8. *Decreasing very low incomes.* For the same reason as above, a large change in the distribution is also needed for this situation. To be consistent with the previous case, incomes of the poorest 5% of the distribution here are divided by ten:

$$y_k^B = \begin{cases} \frac{1}{10}y_k^A, & \text{if } y_k^A \leq Q_{0.05}^A, \\ y_k^A, & \text{otherwise.} \end{cases}$$

### 2.4.5 Changes at the center of the distribution

9. *Reduction of the dispersion around the center of the distribution.* For this isolated situation, the income of 40% of the population is modified (20% above median, 20% below) as follows:

$$y_k^B = \begin{cases} 0.1y_k^A + 0.9\overline{Y}_A, & \text{if } Q_{0.3}^A \leq y_k^A \leq Q_{0.7}^A, \\ y_k^A, & \text{otherwise,} \end{cases}$$

where $\overline{Y}_A = N^{-1}\sum_{k\in U} y_k^A$.

## 2.5 Simulation method

For each case presented in the previous section, 10,000 replications were performed. Each step consists in drawing two independent samples of size $n = 1,000$ observations. The first sample is drawn from population $A$ and is denoted $S_A$; the second sample is drawn from population $B$ and is denoted $S_B$. A simple random sampling design without replacement is used to draw both samples. All seventeen inequality measures presented in Section 2.2 are then estimated on both samples.

As samples $S_A$ and $S_B$ are independent and of equal size, it is reasonable to focus on the following statistic $z$ for each index :

$$z = \frac{\overline{\widehat{I}_B} - \overline{\widehat{I}_A}}{\sqrt{\text{var}(\widehat{I}_A) + \text{var}(\widehat{I}_B)}},$$

where $\widehat{I}_A$ and $\widehat{I}_B$ is the inequality index of interest calculated respectively for sample $S_A$ and sample $S_B$. $\overline{\widehat{I}_A}$ and $\text{var}(\widehat{I}_A)$ are respectively the mean and variance of index $\widehat{I}$ computed over the 10,000 samples drawn from

population $A$. A similar notation is used for samples selected from population $B$. Because of independence between the two samples the variance of $\widehat{I_B} - \widehat{I_A}$ is:

$$\text{var}(\widehat{I_B} - \widehat{I_A}) = \text{var}(\widehat{I_A}) + \text{var}(\widehat{I_B}).$$

A positive (respectively negative) value of $z$ indicates that the index reports an increase (respectively a decrease) of the inequality level. A large absolute value of $z$ indicates that the index has a substantial power of detection of changes for the particular case of interest. These statistics shed light on how well each of the seventeen indices are able to detect changes between the two distributions. Note that a decrease in the level of inequality is expected for all perturbations except Changes 2 and 8 where inequality is expected to rise.

## 2.6 Results and discussion

From the results reported in Table 2.1, we first notice that some categories of measures do not detect the proposed changes in a satisfying manner: the Generalized entropy measures, the Rényi divergence measures, as well as the coefficient of variation and the interquantile ratios (MQR excluded). One of the most relevant results here, is that despite its good properties (additive decomposability) the ability of the Generalized entropy measure (including the Theil index) and Renyi divergence measure to account for the proposed changes is weak unless a small value of parameter $\varphi$ is chosen. Indeed these measures appear to be generally more sensitive to changes for smaller values of $\varphi$, even when perturbation is applied only at the top of the distribution. As expected, interquantile ratios are not very sensitive to changes. Because they are built with quantiles, interquantile ratios are robust against extreme values but are, on the other hand, insensitive to many types of changes. This brings to light one crucial issue when measuring inequality: robustness is both desired and feared. While a robust measure will have the advantage of not depending on a small number of influential observations, it will fail to report evolutions in the income structure unless the changes are spectacular.

One important result of this study is the quality of indices based on quantile share ratios. All quantile share ratios seem to perform well. The QSR and DSR detect most perturbations in a satisfying manner, except, unsurprisingly, the reduction of dispersion around the center of the distribution. Because of their intrinsic construction (only the tails of the distribution are taken into account), the QSR and DSR are unable to detect the middle-class reduction of dispersion. The performance of the QSR provides good news for official statistical institutions using Laeken indicators. However, one shall note that the MSR, which is seldom used in

| | Perturbation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | top | | | all levels | | bottom | | | center |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| G | −3.49 | 2.92 | −2.56 | −2.69 | −0.76 | −0.75 | −1.01 | 0.13 | −1.47 |
| CV | −1.01 | 0.80 | −0.65 | −0.27 | −0.10 | −0.04 | −0.06 | 0.01 | −0.10 |
| $A_{1/2}$ | −2.35 | 2.35 | −1.74 | −2.20 | −0.53 | −0.97 | −1.04 | 0.51 | −0.48 |
| $A_1$ | −2.27 | 2.53 | −1.80 | −4.25 | −0.76 | −2.56 | −2.44 | 2.34 | −0.41 |
| $GE_{1/2}$ | −2.30 | 2.30 | −1.72 | −2.17 | −0.52 | −0.96 | −1.03 | 0.50 | −0.47 |
| T | −1.69 | 1.67 | −1.23 | −0.99 | −0.29 | −0.30 | −0.38 | 0.10 | −0.31 |
| $GE_2$ | −0.47 | 0.40 | −0.34 | −0.14 | −0.05 | −0.01 | −0.03 | 0.01 | −0.05 |
| $R_{1/2}$ | −2.26 | 2.24 | −1.69 | −2.14 | −0.52 | −0.94 | −1.01 | 0.49 | −0.47 |
| $R_2$ | −1.18 | 1.05 | −0.74 | −0.32 | −0.12 | −0.05 | −0.08 | 0.01 | −0.12 |
| Z | −4.16 | 2.92 | −2.59 | −3.94 | −0.87 | −2.08 | −1.80 | 0.34 | −1.34 |
| QSR | −2.20 | 2.18 | −1.83 | −4.13 | −0.95 | −2.29 | −2.37 | 0.43 | 0.01 |
| DSR | −2.09 | 1.97 | −1.53 | −3.77 | −0.70 | −3.52 | −2.82 | 0.93 | 0.00 |
| MSR | −2.50 | 2.39 | −2.14 | −3.58 | −1.00 | −0.90 | −1.46 | 0.15 | −3.14 |
| ASR | −2.41 | 2.32 | −1.99 | −4.32 | −0.99 | −2.24 | −2.23 | 0.42 | −1.36 |
| IQR | 0.01 | 0.00 | −0.39 | −3.45 | −0.95 | −0.01 | −1.39 | 0.00 | −0.01 |
| IDR | −0.08 | −0.01 | −0.73 | −3.71 | −0.85 | −0.83 | −2.25 | 0.00 | −0.01 |
| MQR | −0.12 | 0.07 | −0.47 | −4.73 | −1.18 | −0.78 | −2.05 | 0.12 | −6.06 |

Table 2.1 – *Simulation results: value of statistic z for each inequality measure under each perturbation of the data. Values of z indicate the ability of the index to detect the modification between distributions A and B. A positive value (or negative value) of z indicates that the index reports an increase (or a decrease) of the inequality level. A large absolute value of z indicates that the index is well suited for detecting changes for the particular case of interest.*

applied statistics, is the most versatile of the three tested basic quantile share ratios.

In our simulations, the other Laeken indicator and leading inequality measure, the Gini index G is one of the most sensitive measures for changes at the top of the distribution. However, it reports only average results for other types of changes. In general, our results show that the choice of Laeken inequality indices (QSR and Gini index) is perhaps not optimal, but is far better than expected. For the specific purpose of detecting changes, the QSR seems more suitable than the Gini index. However, as we have seen earlier, the QSR does not satisfy all basic properties of the axiomatic approach to inequality and has thus not been extensively studied in the literature. With its recent promotion as a Laeken indicator, the odds are that the QSR will receive more attention in the near future.

The two different Atkinson indices perform somewhat differently from one another. Indeed, $A_{1/2}$ is globally less competitive than $A_1$. The latter seems to be one of the best measures for this objective. However, it has a major drawback: as shown above in Section 2.2, $A_1$ is expressed with the

Geometric mean. As a result, the value of the index is 1 (perfect inequality) as soon as one observation has a null income. Moreover, and for the same reasons, very small incomes have an excessive influence over the value of the statistic.

Another outcome of this simulation study is that there seems to be room for new or recent measures such as the Zenga index, the MQR and the ASR. The Zenga index is one the most versatile measure, detecting changes successfully in practically all different situations. Moreover it is the most efficient index for the detection of changes with high incomes. Consequently, it is a measure that deserves more attention in future inequality research. The MQR is not a robust measure like the plain interquantile ratios but has improved sensitivity for perturbations among the whole population. The ASR is the most versatile measures in this set of inequality indices. It is well-suited for detecting changes in all types of perturbation. Unlike usual quantile share ratios, the ASR is also able to detect changes around the center of the data.

## 2.7 Conclusion

Measuring changes in income distribution is an important goal of inequality theory. Throughout these simulations on real data, we were able to provide some insights on which measure should or should not be recommended for this purpose. We have shown that quantile share ratios are a reliable class of indices. The good behavior of the latter measures and the weaknesses of the Generalized entropy class also shows that inequality theory cannot rely only on the reasoning of the axiomatic approach. The Atkinson index with $\epsilon = 1$ has also proved to be a very competitive measure as long as strong assumptions are satisfied (no null income in the distribution).

The Gini index gives passable results but is less efficient than its competitor (the QSR) among the Laeken indicators. Finally, the new (ASR and MQR) or recent (Zenga index) measures are very promising. It is indeed interesting to note that the two best measures, on average, for this purpose (ASR and Zenga) have been developed very recently. A future challenge is the expression of estimators and variance estimators within complex sampling designs for these new indicators, as well as an in-depth discussion of their properties.

## Acknowledgments

The authors would like to thank the Cantonal Statistical Office of Neuchâtel, Switzerland and especially Gérard Geiser for providing the data set used in the simulations. The authors are also grateful to Monique Graf

(Swiss Federal Statistical Office), Lionel Qualité and Anthea Monod (University of Neuchâtel) for their useful comments and discussions.

# Statistical inference for the quintile share ratio

# 3


## Abstract

In recent years, the Quintile Share Ratio (or QSR) has become a very popular measure of inequality. In 2001, the European Council decided that income inequality in European Union member states should be described using two indicators: the Gini Index and the QSR. The QSR is generally defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. Thus, it can be expressed using quantile shares, where a quantile share is the share of total income earned by all of the units up to a given quantile. The aim of this paper is to propose an improved methodology for the estimation and variance estimation of the QSR in a complex sampling design framework. Because the QSR is a non-linear function of interest, the estimation of its sampling variance requires advanced methodology. Moreover, a non-trivial obstacle in the estimation of quantile shares in finite populations is the non-unique definition of a quantile. Thus, two different conceptions of the quantile share are presented in the paper, leading us to two different estimators of the QSR. Regarding variance estimation, Osier (2006, 2009) proposed a variance estimator based on linearization techniques. However, his method involves Gaussian kernel smoothing of cumulative distribution functions. Our approach, also based on linearization, shows that no smoothing is needed. The construction of confidence intervals is discussed and a proposition is made to account for the skewness of the sampling distribution of the QSR. Finally, simulation studies are run to assess the relevance of our theoretical results. [1]



**Keywords:** Inequality measure, sampling, variance, estimation, quantile, confidence intervals


[1] This chapter is a reprint of: Langel, M. and Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference* **141**, 2976–2985. http://dx.doi.org/10.1016/j.jspi.2011.03.023

## 3.1 Introduction

Nowadays, the Quintile Share Ratio (QSR) is a widely used measure of inequality. Together with the Gini index, it is one of the two Laken indicators of inequality selected at the European Council in Laken, Belgium in 2001. Laken indicators are used in the European Statistics on Income and Living Conditions (EU-SILC) program run by Eurostat (Eurostat, 2005; Traat, 2006). The QSR is a function of quantile shares, where a quantile share is the share of total income earned by all of the units up to a given quantile.

This paper focuses on conducting statistical inference for the QSR in a complex random sampling framework. However, the proposed method can be applied to other quantile share-based measures. Because the QSR is a non-linear function of the incomes, variance estimation is not straightforward and requires specific techniques. The variance estimators proposed here are based on the linearization approach by Deville (1999). Inference for the QSR using this approach has already been conducted by Osier (2006, 2009) and similar work has been done for the Gini index (Deville, 1996, 1999; Berger, 2008; Barrett & Donald, 2009). However, Osier's approach is intricate because it requires kernel smoothing of cumulative distribution functions. With the improvement proposed in this paper, smoothing is no longer required. Moreover, an alternative estimator and variance estimator are presented, and a set of simulations advocate in favor of the latter.

The paper is organized as follows. Section 3.2 starts with a presentation of key concepts, namely quantiles, quantile shares and partial sums. The continuous case is discussed to begin with, but emphasis is placed on finite population expressions as well as on estimators under complex sampling designs. In this section, it is also stressed that the partial sum centralizes the main issues in conducting valid inference for the QSR. Thus, two distinct finite population expressions and estimators of the partial sum are presented. The first one is based on quantiles and leads to a natural expression of the QSR, while the second is an alternative expression that gets around the finite population quantile issue and leads to another finite population expression of the inequality measure of interest. Both are described in Section 3.3 and their respective estimators are given.

Section 3.4 is a succinct description of the linearization technique using influence functions for variance estimation, as initially proposed by Deville (1999). The approach is then applied in parallel to both estimators of the QSR, providing us with two distinct variance estimators. Firstly, the influence functions of both expressions of the partial sum are derived in Sections 3.5 and 3.6. In these sections, we point out that, unlike in the approach by Osier (2006, 2009), no smoothing is needed. The two resulting variance estimators are then derived in Sections 3.7 and 3.8. A discussion on confidence intervals and skewness issues is proposed in Section 3.9.

Finally, two sets of simulations on real data are presented in Section 3.10, preceding some concluding remarks.

## 3.2 Estimation of quantile shares

Quantile share-based measures constitute a very interesting class of inequality indices in their capacity to detect perturbations at different levels of an income distribution (Langel & Tillé, 2011b). However, inference on these measures is not straightforward, especially when dealing with complex sampling designs. Consider a continuous strictly increasing cumulative distribution function $F(y)$ and $F'(y)$, its derivative and probability density function. Also, let us denote $Q_\alpha$, the quantile of order $\alpha$, such that $F(Q_\alpha) = \alpha$. The quantile function can be written as the inverse of the cumulative distribution function $Q_\alpha = F^{-1}(\alpha)$. A quantile share is the share of total income earned by all the income earners up to quantile of order $\alpha$. The definition for the continuous case is

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

This expression is also frequently referred to as the Lorenz function or Lorenz curve (Lorenz, 1905; Gastwirth, 1972; Cowell, 1977; Kovacevic & Binder, 1997), which is a central tool of inequality theory.

Let $U$ denote a finite population of $N$ identifiable units $u_1$,...,$u_k$,...,$u_N$. For the sake of simplicity, we will hereafter denote unit $u_k$ by its identifier $k$. Associated with each unit $k$ is the value $y_k$ of some characteristic of interest, for example income. To lighten the notation, we will assume with no loss of generality that all $y_k$'s are distinct and sorted. The finite population quantile share is $L(\alpha) = Y_\alpha / Y$, where $Y = \sum_{k \in U} y_k$ and

$$Y_\alpha = \sum_{k \in U} y_k \mathbb{1}\left(y_k \leq Q_\alpha\right), \tag{3.1}$$

with $Q_\alpha$, the quantile of order $\alpha$ and with $\mathbb{1}(A) = 1$ if $A$ is true and 0 otherwise. Expression (3.1) is thereafter denoted as the *partial sum* of income $y$. The finite population quantile share $L(\alpha)$ is the cumulative sum of income up to a given quantile $Q_\alpha$ over the total income. Or, in other words, the share of total income earned by the $\alpha N$ poorer units. In the following, we will mainly focus on the partial sum $Y_\alpha$, because it embodies the complex part of the quantile share.

A classical notation from survey sampling theory is used hereafter. Thus, let us denote $S$, a random sample of size $n$, and the function $p(s) = \Pr(S = s)$, which gives the probability of selecting the particular sample $s \subset U$. The inclusion probability of unit $k$ is denoted $\pi_k$ and defined such that $\pi_k = \Pr(k \in S)$. Also, $w_k$ stands for the weight of unit $k$. Weights

can simply be the inverse of the inclusion probability $w_k = 1/\pi_k$, but can also result from a calibration procedure (Deville & Särndal, 1992) or non-response adjustments (Särndal & Lundström, 2005). In the following, it is assumed that the sampling design used to draw sample $S$ is associated with a known expression of the variance of the estimated total

$$\widehat{Y} = \sum_{k \in S} w_k y_k. \tag{3.2}$$

From a sample, $Y_\alpha$ can be estimated in a similar fashion using the plug-in estimator

$$\widehat{Y}_\alpha = \sum_{k \in S} w_k y_k \mathbb{1}\left(y_k \leq \widehat{Q}_\alpha\right), \tag{3.3}$$

where $\widehat{Q}_\alpha$ is an estimator of quantile $Q_\alpha$. Both the quantile and its estimator have to be precisely defined. While the definition of a quantile in a continuous distribution is clear and unique, it is not so in the finite population context, where $F$, the cumulative distribution of income $y$, is a step function. Accordingly, obtaining a univocal definition of quantile $Q_\alpha$ is not possible. In the paper by Hyndman & Fan (1996), nine different definitions of sample quantiles are described, all of them existing in the literature and in statistical packages. In order to compute $\widehat{Y}_\alpha$ in the simulation study (Section 3.10), we will be using the fourth definition of the quantile of Hyndman & Fan (1996), which is based on a simple linear interpolation of the cumulative distribution function:

$$Q_\alpha = y_{k-1} + (y_k - y_{k-1})\left[\alpha N - (k-1)\right], \tag{3.4}$$

where $\alpha N < k \leq \alpha N + 1$. The quantile can be estimated from a sample by:

$$\widehat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1})\left(\frac{\alpha \widehat{N} - W_{k-1}}{w_k}\right),$$

where $W_k = \sum_{\ell \in S} w_\ell \mathbb{1}(y_\ell \leq y_k)$, $\widehat{N} = W_n$ and the value of $k$ is such that $W_{k-1} < \alpha \widehat{N} \leq W_k$.

As emphasized above, the value of $Q_\alpha$, and consequently of $Y_\alpha$, is dependent on how the discontinuities of the cumulative distribution function are dealt with. This issue fosters the use of another definition of the partial sum that is not directly dependent on the definition of the quantile:

$$\widetilde{Y}_\alpha = \sum_{k \in U} y_k H\left[\alpha N - (k-1)\right], \tag{3.5}$$

where

$$H(x) = \begin{cases} 0 & \text{if} \quad x < 0, \\ x & \text{if} \quad 0 \leq x < 1, \\ 1 & \text{if} \quad x \geq 1. \end{cases} \tag{3.6}$$

Here, $H$ is the cumulative distribution function of a uniform random variable. Under this definition, $\widetilde{Y}_\alpha$ is a strictly increasing function of $\alpha$. This is a desirable property for the estimation of the quantile share and its variance in sampling from a finite population. Partial sum $\widetilde{Y}_\alpha$ can be estimated from a sample $S$ using the plug-in estimator

$$\widehat{\widetilde{Y}}_\alpha = \sum_{k \in S} w_k y_k H\left(\frac{\alpha \widehat{N} - W_{k-1}}{w_k}\right). \tag{3.7}$$

In the following, both $Y_\alpha$ (3.1) and $\widetilde{Y}_\alpha$ (3.5) will be considered in order to define the QSR and provide variance estimators using influence functions.

## 3.3 The quintile share ratio

The QSR is defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. For the continuous case, we thus have

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

In finite populations, the QSR can be viewed as a function of quantile shares or as a function of partial sums. Because we have proposed two different finite population definitions of the partial sum, the QSR can be defined by

$$\text{QSR} = \frac{Y - Y_{0.8}}{Y_{0.2}}, \tag{3.8}$$

or by

$$\widetilde{\text{QSR}} = \frac{Y - \widetilde{Y}_{0.8}}{\widetilde{Y}_{0.2}}, \tag{3.9}$$

and can be respectively estimated from a sample by

$$\widehat{\text{QSR}} = \frac{\widehat{Y} - \widehat{Y}_{0.8}}{\widehat{Y}_{0.2}}, \tag{3.10}$$

or by

$$\widehat{\widetilde{\text{QSR}}} = \frac{\widehat{Y} - \widehat{\widetilde{Y}}_{0.8}}{\widehat{\widetilde{Y}}_{0.2}}. \tag{3.11}$$

## 3.4 Approximation of the variance by linearization

The estimators of the Quintile Share Ratio as defined in (3.10) and (3.11) are nonlinear statistics, and therefore a general expression for their sampling variance is not known. In the literature, a variety of methods such

as resampling techniques or linearization allow for variance estimation of complex statistics (for a survey of most existing methods see Wolter, 2007). Linearization methods have given rise to a lot of research using different approaches (Woodruff, 1971; Binder & Patak, 1994; Kovacevic & Binder, 1997; Deville, 1999; Demnati & Rao, 2004). This paper focuses on the linearization method developed by Deville (1999). His approach is based on the influence function, a predominant notion in the field of robust statistics (Hampel et al., 1985). The idea behind this method is to study the influence of unit $k$ on the population parameter of interest by adding an infinitesimal variation of the weight to this unit. A population parameter $\theta$ can be written as a functional $T(M)$, where measure $M$ allocates a unit mass to all $k \in U$. The influence function of $T$ is then defined as

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

where $\delta_k$ denotes the Dirac measure for unit $k$. The term *linearized variable* is used hereafter to denote $z_k$. This terminology is used by Deville (1999) and advocated by Skinner (2004). Under asymptotic conditions described in Deville (1999), it is shown that the variance of the estimated total of the linearized variable $z_k$ is an approximation to the variance of statistic $\widehat{\theta}$ (an estimator of $\theta$) :

$$\text{var}\left(\sum_{k \in S} z_k w_k\right) \approx \text{var}\left(\widehat{\theta}\right). \tag{3.12}$$

In practice, values $z_k$ at the left-hand side of Expression (3.12) are not known because they rely on unavailable information at the population level. Thus, the $z_k$'s are estimated from the sample by using the plug-in estimator $\hat{z}_k = I[T(\widehat{M})]_k$, where $\widehat{M}$ is the measure allocating a mass $w_k$ to all $k \in S$. With the proposed method, the variance of a complex statistic $\widehat{\theta}$ can be estimated under any sampling design for which the expression of the variance of the estimator of a total is available.

## 3.5 Linearization of $Y_\alpha$

The method described in the above section can be applied to obtain an expression for the influence function of both definitions of the partial sum. Derivation for the influence function of $Y_\alpha$ is shown in the present section, whereas Section 3.6 presents derivation for $\widetilde{Y}_\alpha$. For $Y_\alpha$, a solution is proposed in Osier (2006, 2009) :

$$I(Y_\alpha)_k = y_k \mathbb{1}(y_k \leq Q_\alpha) + \widetilde{S}'(Q_\alpha) I(Q_\alpha)_k, \tag{3.13}$$

where $\widetilde{S}'$ is the derivative of $\widetilde{S}$, a smoothed function of

$$S(y) = \sum_{k \in U} y_k \mathbb{1}(y_k \leq y).$$

Two issues arise from this expression: the smoothing of $S$ and the computation of $I(Q_\alpha)_k$, the influence function of quantile $Q_\alpha$. A solution to the latter issue is proposed by Deville (1999):

$$I(Q_\alpha)_k = -\frac{\mathbb{1}(y_k \leq Q_\alpha) - \alpha}{\widetilde{F}'(Q_\alpha) N}, \tag{3.14}$$

with $\widetilde{F}'(y)$, the derivative of $\widetilde{F}(y)$, a smoothed function of

$$F(y) = \frac{1}{N} \sum_{k \in U} \mathbb{1}(y_k \leq y).$$

At this point, computing $I(Y_\alpha)_k$ thus implies the smoothing of two discontinuous step functions, $S$ and $F$. Deville (1999) suggests kernel smoothing for $F$. In order to estimate the variance of the QSR estimated from a sample for various European Union member states, Osier (2006, 2009) has applied (3.13) using Gaussian kernel smoothing of $S$ and $F$. We propose hereafter a simpler solution that does not require smoothing of the latter functions.

Let us momentarily consider $\widetilde{S}(y)$ and $\widetilde{F}(y)$, the smoothed functions of $S$ and $F$ respectively. Let also $G(y) = \widetilde{S}(y)/Y$. Since $\widetilde{S}(y)$ is differentiable, $G(y)$ is also differentiable. If $G'(y)$ denotes the derivative of $G(y)$, then

$$\widetilde{S}'(y) = G'(y) Y. \tag{3.15}$$

Functions $G(y)$ and $F(y)$ are both cumulative distribution functions. $G(y)$ can also be defined by

$$G(y) = \frac{\int_0^y u d\widetilde{F}(u)}{\int_0^\infty u d\widetilde{F}(u)} = \frac{N \int_0^y u d\widetilde{F}(u)}{Y}.$$

Thus, $G'(y)$, can be written

$$G'(y) = \frac{N y \widetilde{F}'(y)}{Y}. \tag{3.16}$$

Let us now substitute (3.14) and (3.15) in Expression (3.13):

$$I(Y_\alpha)_k = y_k \mathbb{1}(y_k \leq Q_\alpha) - \frac{Y}{N} \frac{G'(Q_\alpha)}{\widetilde{F}'(Q_\alpha)} \left[\mathbb{1}(y_k \leq Q_\alpha) - \alpha\right]. \tag{3.17}$$

Moreover, from (3.16) we have

$$\frac{G'(Q_\alpha)}{\widetilde{F}'(Q_\alpha)} = \frac{Q_\alpha N}{Y}, \tag{3.18}$$

and thus, replacing (3.18) into (3.17), we finally obtain

$$I(Y_\alpha)_k = \alpha Q_\alpha - (Q_\alpha - y_k)\mathbb{1}(y_k \leq Q_\alpha), \tag{3.19}$$

where no smoothing is needed. Indeed, densities $\widetilde{F}'$ and $G'$ do not appear in Result (3.19), making the computation of the influence function of $Y_\alpha$ markedly more straightforward than the method initially proposed by Osier (2006, 2009).

## 3.6 Linearization of $\widetilde{Y}_\alpha$

The influence function of $\widetilde{Y}_\alpha$ can also be derived. First, the rule for the linearization of a product (Deville, 1999; Dell et al., 2002) is applied to Equation (3.5) :

$$I(\widetilde{Y}_\alpha)_k = y_k H\left(\alpha N - k + 1\right) + \sum_{j \in U} y_j I\left[H\left(\alpha N - j + 1\right)\right]_k. \tag{3.20}$$

The influence function of $H$ is:

$$I[H(\alpha N - j + 1)]_k = \mathbb{1}(0 < \alpha N - j + 1 \leq 1)[\alpha - \mathbb{1}(k < j)],$$

and because $\mathbb{1}(k < j) = \mathbb{1}(y_k < y_j)$, we obtain

$$\sum_{j \in U} y_j I[H(\alpha N - j + 1)]_k = [\alpha - \mathbb{1}(y_k < \widetilde{Q}_\alpha)]\widetilde{Q}_\alpha, \tag{3.21}$$

where $\widetilde{Q}_\alpha$ denotes the first definition of the finite population quantile in the paper by Hyndman & Fan (1996):

$$\widetilde{Q}_\alpha = y_i, \quad \text{where} \quad i - 1 < \alpha N \leq i.$$

Finally, the influence function of $\widetilde{Y}_\alpha$ is obtained by substituting (3.21) in (3.20):

$$I(\widetilde{Y}_\alpha)_k = y_k H(\alpha N - k + 1) + [\alpha - \mathbb{1}(y_k < \widetilde{Q}_\alpha)]\widetilde{Q}_\alpha. \tag{3.22}$$

No prior definition of the quantile in finite populations was required in order to derive $I(\widetilde{Y}_\alpha)_k$. However, quantile $\widetilde{Q}_\alpha$ appears in the final expression.

## 3.7 Linearization of QSR

The influence function of QSR is computed by applying the derivation rule for the linearization of a ratio (Deville, 1999; Dell et al., 2002) on Expression (3.8),

$$I(\text{QSR})_k = \frac{y_k - I(Y_{0.8})_k}{Y_{0.2}} - \frac{(Y - Y_{0.8})I(Y_{0.2})_k}{Y_{0.2}^2},$$

and by replacing $I(Y_{0.2})_k$ and $I(Y_{0.8})_k$ with the result obtained in (3.19). We therefore have

$$\begin{aligned} I(\text{QSR})_k &= \frac{y_k - [0.8Q_{0.8} - (Q_{0.8} - y_k)\mathbb{1}(y_k \leq Q_{0.8})]}{Y_{0.2}} \\ &\quad - \frac{(Y - Y_{0.8})[0.2Q_{0.2} - (Q_{0.2} - y_k)\mathbb{1}(y_k \leq Q_{0.2})]}{Y_{0.2}^2}. \end{aligned}$$

Our final aim here is to derive a sampling variance estimator for $\widehat{\text{QSR}}$. Linearization theory shows us that, with $z_k = I(\text{QSR})_k$,

$$\text{var}(\widehat{\text{QSR}}) \approx \text{var}\left(\sum_{k \in S} z_k w_k\right).$$

In practice, however, the linearized variable $z_k$ involves unavailable information at the population level and has to be estimated from the sample by its plug-in estimator

$$\begin{aligned} \hat{z}_k &= \frac{y_k - [0.8\widehat{Q}_{0.8} - (\widehat{Q}_{0.8} - y_k)\mathbb{1}(y_k \leq \widehat{Q}_{0.8})]}{\widehat{Y}_{0.2}} \\ &\quad - \frac{(\widehat{Y} - \widehat{Y}_{0.8})[0.2\widehat{Q}_{0.2} - (\widehat{Q}_{0.2} - y_k)\mathbb{1}(y_k \leq \widehat{Q}_{0.2})]}{\widehat{Y}_{0.2}^2}. \end{aligned}$$

The estimated variance of $\widehat{\text{QSR}}$ is obtained by estimating the variance of the weighted sum $\widehat{Z} = \sum_{k \in S} \hat{z}_k w_k$. The method is thus easily applicable to a whole variety of complex sampling designs. Under a simple random sampling design without replacement, the variance estimator for $\widehat{\text{QSR}}$ is:

$$\widehat{\text{var}}_{lin}(\widehat{\text{QSR}}) = \frac{N(N-n)}{n(n-1)} \sum_{k \in S} (\hat{z}_k - \bar{z})^2, \tag{3.23}$$

with $\bar{z} = n^{-1} \sum_{k \in S} \hat{z}_k$. In the following, the latter estimator is always referred to as $\widehat{\text{var}}_{lin}(\widehat{\text{QSR}})$ to emphasize the fact that it is obtained through a linearization technique and to clearly distinguish it from the Monte Carlo estimator used in the simulation studies.

## 3.8 Linearization of $\widetilde{\text{QSR}}$

A very similar derivation can be done for the influence function of $\widetilde{\text{QSR}}$. Indeed, we have

$$I(\widetilde{\text{QSR}})_k = \frac{y_k - I(\widetilde{Y}_{0.8})_k}{\widetilde{Y}_{0.2}} - \frac{(Y - \widetilde{Y}_{0.8})I(\widetilde{Y}_{0.2})_k}{\widetilde{Y}_{0.2}^2},$$

and applying (3.22), $I(\widetilde{\text{QSR}})_k$ can be rewritten

$$\begin{aligned} I(\widetilde{\text{QSR}})_k &= \frac{y_k - \{y_k H(0.8N - k + 1) + \widetilde{Q}_{0.8}[0.8 - \mathbb{1}(y_k < \widetilde{Q}_{0.8})]\}}{\widetilde{Y}_{0.2}} \\ &\quad - \frac{(Y - \widetilde{Y}_{0.8})\{y_k H(0.2N - k + 1) + \widetilde{Q}_{0.2}[0.2 - \mathbb{1}(y_k < \widetilde{Q}_{0.2})]\}}{\widetilde{Y}_{0.2}^2}. \end{aligned}$$

With $\tilde{z}_k = I(\widetilde{\text{QSR}})_k$, we have

$$\text{var}(\widehat{\widetilde{\text{QSR}}}) \approx \text{var}\left(\sum_{k \in S} \tilde{z}_k w_k\right),$$

and in practice, the (unknown) $\tilde{z}_k$'s are replaced by the plug-in estimator

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

where $\widehat{\widetilde{Q}}_\alpha = y_i$, with $W_{i-1} < \alpha N \leq W_i$. Finally, the variance estimator for a simple random sampling design without replacement is constructed similarly as in (3.23):

$$\widehat{\text{var}}_{lin}(\widehat{\widetilde{\text{QSR}}}) = \frac{N(N-n)}{n(n-1)} \sum_{k \in S} \left(\hat{\tilde{z}}_k - \bar{\tilde{z}}\right)^2 \tag{3.24}$$

where $\bar{\tilde{z}} = n^{-1} \sum_{k \in S} \hat{\tilde{z}}_k$.

## 3.9 Construction of a confidence interval

As shown by the simulation results in Section 3.10 below, the variances are successfully and accurately estimated by using the linearization method. However, because of the skewness of the sampling distributions of the statistics $\widehat{\text{QSR}}$ and $\widehat{\widetilde{\text{QSR}}}$, the normality based confidence intervals built around these estimators are somewhat less convincing.

To account for the skewness issues in the interval estimation of $\widehat{\text{QSR}}$ and $\widetilde{\text{QSR}}$, we propose an alternative method for the construction of a more reliable confidence interval. The method, based on Box-Cox transformations (Box & Cox, 1964), aims to obtain a less skewed distribution after transformation, and consequently to build confidence intervals on the latter distribution. The Box-Cox transformations for a parameter $\theta$ are given by:

$$\theta^{(\lambda)} = \begin{cases} \dfrac{\theta^{\lambda} - 1}{\lambda} & \text{if} \quad \lambda \neq 0, \\ \log \theta & \text{if} \quad \lambda = 0. \end{cases} \tag{3.25}$$

The method consists of constructing confidence intervals for $\text{QSR}^{(\lambda)}$ and $\widetilde{\text{QSR}}^{(\lambda)}$. For that purpose, variance estimators of the latter statistics also need to be derived. With the linearized variable $z_k = I(\text{QSR})_k$, the influence function of any Box-Cox transformation $\text{QSR}^{(\lambda)}$ is

$$I[\text{QSR}^{(\lambda)}]_k = z_k \text{QSR}^{\lambda - 1}. \tag{3.26}$$

Thus, we also have

$$I[\widehat{\text{QSR}}^{(\lambda)}]_k = \hat{z}_k \widehat{\text{QSR}}^{\lambda - 1}, \tag{3.27}$$

and the variance estimator

$$\widehat{\text{var}}_{lin}[\widehat{\text{QSR}}^{(\lambda)}] = \widehat{\text{QSR}}^{2(\lambda-1)} \widehat{\text{var}}_{lin}(\widehat{\text{QSR}}). \tag{3.28}$$

Expression (3.28) shows that a variance estimator and confidence interval can be directly derived for any value of parameter $\lambda$. An alternative confidence interval for the untransformed QSR can thus be obtained by computing the inverse transformation of the lower and upper bounds of the confidence interval for $\text{QSR}^{(\lambda)}$. The procedure can be applied equivalently to $\widetilde{\text{QSR}}^{(\lambda)}$. The aim is to choose the value of $\lambda$ which yields the most symmetric distribution. The sampling distribution of the statistic is unknown. However, our simulation studies below (Section 3.10) show that $\lambda = -1$ seems to be an appropriate solution.

## 3.10 Simulation studies

Two simulation studies have been carried out on real data to evaluate the quality of the linearization variance estimators of the QSR proposed in this paper. The data used in the simulations is the household taxable income of the Canton of Neuchâtel, Switzerland for year 2007. It regroups a population of $N = 88,106$ non-null income earners. The data is highly positively skewed, which is not surprising for income data.

The first simulation is dedicated to $\widehat{\text{QSR}}$, while the second focuses on $\widetilde{\text{QSR}}$. For the first simulation study, 100,000 samples of size $n = 1,000$

are drawn using a simple random sampling design without replacement. Firstly, the value of $\widehat{\text{QSR}}$ is computed on each sample, which provides us with $\text{var}_{sim}(\widehat{\text{QSR}})$, a Monte Carlo estimator of the variance of $\widehat{\text{QSR}}$ under the simulations. The linearization variance estimator $\widehat{\text{var}}_{lin}(\widehat{\text{QSR}})$ (Equation 3.23) is then computed on each sample using the linearized variable $\hat{z}$. The main goal of this study is to analyze the quality of the latter estimator in terms of bias and variance with respect to $\text{var}_{sim}(\widehat{\text{QSR}})$. For this purpose, the Monte Carlo expected value and variance of the linearization variance estimators, respectively denoted $\text{E}_{sim}[\widehat{\text{var}}_{lin}(\widehat{\text{QSR}})]$ and $\text{var}_{sim}[\widehat{\text{var}}_{lin}(\widehat{\text{QSR}})]$, are computed.

Likewise, the second simulation study aims to compare $\text{var}_{sim}(\widehat{\widehat{\text{QSR}}})$ and $\widehat{\text{var}}_{lin}(\widehat{\widehat{\text{QSR}}})$. The method and sampling designs are exactly identical in both studies. Eventually, the joint analysis of the results from the two simulations allows for a brief comparison of both definitions of the Quintile Share Ratio, and of their capacity to provide a reliable estimation. The results can be viewed in Table 3.1.

Table 3.1 – *Results for both simulation studies (100,000 replications each). The middle column summarizes results from the simulation study for which estimator* $\widehat{\text{QSR}}$ *and linearized variable* $\hat{z}$ *are used. Respectively, results displayed on the right-hand side for the second simulation study were computed using* $\widehat{\widehat{\text{QSR}}}$ *and* $\hat{\hat{z}}$*. The coverage rate for a 95% normality based confidence interval for* QSR *is denoted* CR*.*

| | **Simulations on** $\widehat{\theta} = \widehat{\text{QSR}}$ | **Simulations on** $\widehat{\theta} = \widehat{\widehat{\text{QSR}}}$ |
|---|---|---|
| $\text{var}_{sim}(\widehat{\theta})$ | 0.9216 | 0.9280 |
| $\text{E}_{sim}[\widehat{\text{var}}_{lin}(\widehat{\theta})]$ | 0.9153 | 0.9273 |
| $\text{var}_{sim}[\widehat{\text{var}}_{lin}(\widehat{\theta})]$ | 1.7731 | 1.7793 |
| $\text{RB}[\widehat{\text{var}}_{lin}(\widehat{\theta})]$ | $-0.68\%$ | $-0.07\%$ |
| $\text{RB}(\widehat{\theta})$ | $-0.60\%$ | $0.12\%$ |
| CR | 92.0% | 93.1% |

The relative bias for $\widehat{\theta}$ and $\widehat{\text{var}}_{lin}(\widehat{\theta})$ are respectively defined by

$$\text{RB}(\widehat{\theta}) = [\text{E}_{sim}(\widehat{\theta}) - \theta]/\theta,$$

and

$$\text{RB}[\widehat{\text{var}}_{lin}(\widehat{\theta})] = \{\text{E}_{sim}[\widehat{\text{var}}_{lin}(\widehat{\theta})] - \text{var}_{sim}(\widehat{\theta})\}/\text{var}_{sim}(\widehat{\theta}).$$

The two definitions above are slightly different because while $\widehat{\theta}$ is compared to the true finite population value $\theta$, the variance estimator is compared to the Monte Carlo estimator $\text{var}_{sim}(\widehat{\theta})$ which approximates the true value.

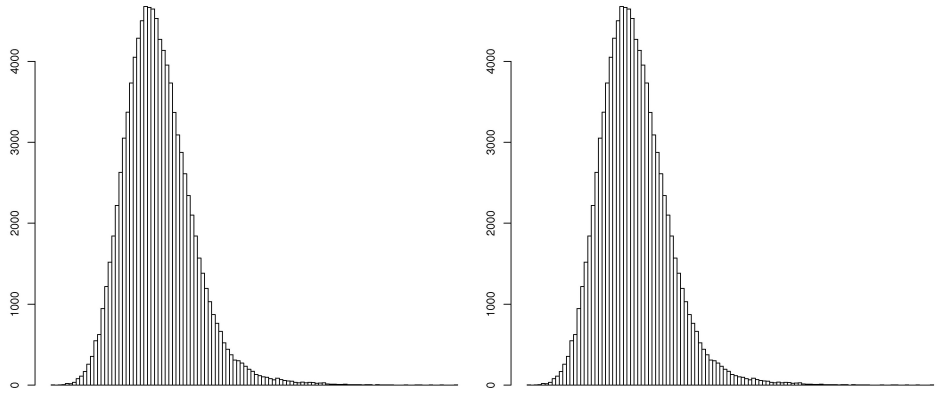
The results show that the variance estimators obtained via the linearization method are very close to the Monte Carlo variance estimator.

This is emphasized by the relative bias of the linearization variance estimators in both simulations ($-0.68\%$ and $-0.07\%$ respectively). Point estimation is also accurate with a relative bias of $-0.60\%$ for the estimator $\widehat{\text{QSR}}$ and of $0.12\%$ for $\widehat{\widehat{\text{QSR}}}$. Although these results show the validity and accuracy of both categories of estimators, they also emphasize that $\widehat{\widehat{\text{QSR}}}$ leads to a slightly better inference for point and variance estimation than $\widehat{\text{QSR}}$.

It is also to be noticed that 95% confidence intervals in both simulations are only partially satisfactory in terms of coverage rate (CR). The coverage rate of the confidence intervals constructed around $\widehat{\widehat{\text{QSR}}}$ is however distinctly better than for $\widehat{\text{QSR}}$ with 93.1% and 92.0%, respectively. This is one of the major reasons why inference on $\widehat{\widehat{\text{QSR}}}$ should be preferred. Also, a possible improvement of the coverage rate is discussed below.

As shown in Figure 3.1, a substantial level of skewness has been observed in the sampling distributions of $\widehat{\text{QSR}}$ and $\widehat{\widehat{\text{QSR}}}$. This seems to result from the skewness of the incomes and from the sensitivity of these statistics to extreme values. The skewness in the distributions of $\widehat{\text{QSR}}$ and $\widehat{\widehat{\text{QSR}}}$

Figure 3.1 – *Histograms of the distributions of* $\widehat{\text{QSR}}$ *and* $\widehat{\widehat{\text{QSR}}}$ *computed on* $100,000$ *simple random samples without replacement of size* $n = 1,000$.

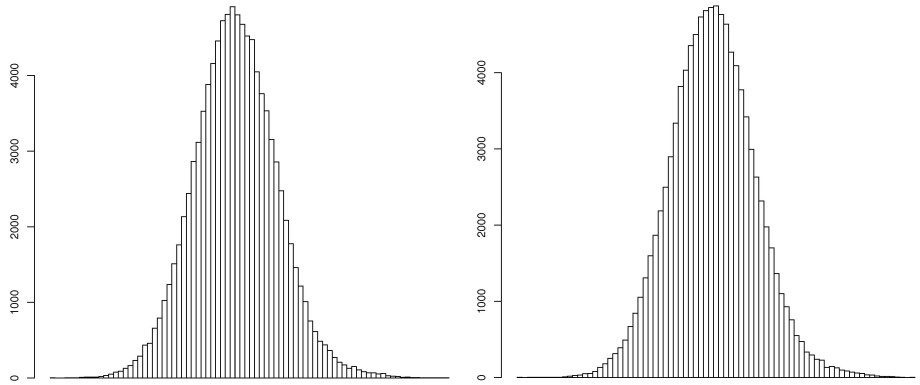



(a) $\widehat{\text{QSR}}$ (b) $\widehat{\widehat{\text{QSR}}}$

makes it difficult for us to achieve reliable confidence intervals. To account for this question, we have applied the procedure proposed in Section 3.9, and have constructed confidence intervals for two different Box-Cox transformations of $\widehat{\text{QSR}}$ and $\widehat{\widehat{\text{QSR}}}$: $\lambda = 0$ and $\lambda = -1$. Figure 3.2 displays the improvement in terms of skewness of the distribution provided by the Box-Cox $\lambda = -1$ transformation.

In order to produce confidence intervals for these transformations, linearization variance estimators for $\lambda = 0$ and $\lambda = -1$ are obtained from

Figure 3.2 – *Histograms of the distributions of* $\widehat{\text{QSR}}^{(-1)}$ *and* $\widehat{\widehat{\text{QSR}}}^{(-1)}$ *computed on* 100,000 *simple random samples without replacement of size* $n = 1,000$.




(a) $\widehat{\text{QSR}}^{(-1)}$ (b) $\widehat{\widehat{\text{QSR}}}^{(-1)}$

Expression (3.28):

$$\widehat{\text{var}}_{lin}[\widehat{\text{QSR}}^{(0)}] = \frac{1}{\widehat{\text{QSR}}^2}\widehat{\text{var}}_{lin}(\widehat{\text{QSR}}),$$

$$\widehat{\text{var}}_{lin}[\widehat{\text{QSR}}^{(-1)}] = \frac{1}{\widehat{\text{QSR}}^4}\widehat{\text{var}}_{lin}(\widehat{\text{QSR}}),$$

and similarly with $\widehat{\widehat{\text{QSR}}}$ for the second set of simulations. A confidence interval for QSR is then simply obtained by applying a back transformation on the confidence bounds. As shown in Table 3.2, the transformations result in a substantial improvement of the coverage rate for both simulation studies. Because they performed a more severe correction of the asymmetry, the $\lambda = -1$ transformations yield better results than the log-transformation ($\lambda = 0$), with coverage rates of 93.6% in the first study, and 94.0% in the second.

Table 3.2 – *Coverage rates (*CR*) for* 95% *confidence intervals for* QSR *using Box-Cox transformations. The initial coverage rate (untransformed) from Table 3.1 is reprinted for comparison purposes.*

| | **Simulations on** $\widehat{\text{QSR}}$ | **Simulations on** $\widehat{\widehat{\text{QSR}}}$ |
|---|---|---|
| CR (untransformed) | 92.0% | 93.1% |
| CR ($\lambda = 0$) | 92.9% | 93.8% |
| CR ($\lambda = -1$) | 93.6% | 94.0% |

## 3.11 Conclusion

The goal of this paper was to provide reliable tools for finite population inference for the Quintile Share Ratio. We have proposed two distinct estimators for the latter inequality measure, as well as two corresponding variance estimators. Although presented in the simulation studies for simple random sampling, all proposed estimators can be used with complex sampling designs. Indeed, because it focuses on linearization techniques, the method holds for any sampling design as long as the expression for the variance of an estimated total is known.

Next, the two variance estimators proposed in this paper do not require the smoothing of any cumulative distribution function or density. Accordingly, variance estimation for the Quintile Share Ratio is not only faster and simpler with our technique than with past methodology, but it also avoids issues that are inherent to non-parametric smoothing, such as the choice of the type of kernel or the size of the bandwidth. This is thus a sensible improvement to the method proposed by Osier (2006, 2009).

Also, while the paper focuses on the Quintile Share Ratio, the main contribution involves more specifically the partial sum and the quantile share. Thus, the method is not restricted to the Quintile Share Ratio and can be applied to other quantile share-based functions of interest. Our simulation studies on real income data confirm the theoretical findings and show that the method is accurate and straightforward to apply. As in depth analysis of the simulations shows, estimating the quintile share ratio with $\widehat{\widehat{\text{QSR}}}$ seems to be slightly more favorable in terms of bias and coverage rate than with $\widehat{\text{QSR}}$. Using real data also reminds us that skewness issues as well as the sensitivity of the statistic to extreme values are obstacles to reliable inference. We have shown that Box-Cox transformations can help address the problem of skewness. In particular, studying the $\lambda = -1$ transformation instead of the Quintile Share Ratio itself can be a valuable alternative. Finally, studies on the robustness of inequality measures (Hulliger & Münnich, 2006) can provide insights on the issue of sensitivity to outliers.

## Acknowledgements

The authors are grateful to the Office Cantonal de la Statistique (Canton de Neuchâtel, Switzerland) and especially to Gérard Geiser for the dataset. The authors would also like to thank an anonymous referee for useful comments and suggestions. This research is supported by Grant no. 200021-121604 of the Swiss National Science Foundation.

# 4 Corrado Gini, a pioneer in balanced sampling and inequality theory

**Abstract**

This paper attempts to make the link between two of Corrado Gini's contributions to statistics: the famous inequality measure that bears his name and his work in the early days of balanced sampling. Some important notions of the history of sampling such as representativeness, randomness, and purposive selection are clarified before balanced sampling is introduced. The Gini index is described, as well as its estimation and variance estimation in the sampling framework. Finally, theoretical grounds and some simulations on real data show how some well used auxiliary information and balanced sampling can enhance the accuracy of the estimation of the Gini index. [1]

**Keywords:** balanced sampling, Gini, inequality, linearization, purposive selection, representativeness

## 4.1 Introduction

Undoubtedly, the most famous contribution of Corrado Gini to the field of statistics is his work on inequality measures, prominently the Gini index, which is still today the leading inequality measure. It is used in various domains outside statistics such as economics, sociology, demography and is a strong political tool. A probably less known contribution of Gini is to be found in the field of survey statistics. Indeed, in the 1920s, Gini took part in the committee that advocated the use of samples in official statistics. Moreover, some years later he sampled Italian circumscriptions

[1] This chapter is a reprint of: Langel, M. and Tillé, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. *Metron* **69**, 45–65.

in a way that is often referred to as the first example of a balanced sample. The goal of this article is to provide an understanding of the evolution of the notion of balanced sampling across the twentieth century and how this notion has participated in the debate between random and purposive selection methods. Eventually, we propose to link both of the above contributions of Corrado Gini and show how some well used auxiliary information and balanced sampling can enhance the accuracy of the estimation of the Gini index.

The next section is dedicated to a short overview of the history of sampling and particularly of the debate on purposive and random methods. In Section 4.3, the controversial notion of representativeness is widely discussed. Balanced sampling and Gini's contribution to the topic are considered in Section 4.4. Sections 4.5 and 4.6 define the Gini index, its expression in the sampling framework as well as the estimation of its sampling variance. The two further sections are dedicated to the linkage between the Gini index and balanced sampling: Section 4.7 shows how balanced sampling can enhance the accuracy of the estimation of the Gini index by means of a sample, followed by a confirmatory simulation study on real data operated and discussed in Section 4.8. Finally, the last section of the paper is dedicated to concluding remarks.

## 4.2 The early stages of sampling

### 4.2.1 Acceptance in the scientific community

The common use of sampling in statistics is recent. Many papers report the historical background of sampling in statistics (for example Hansen et al., 1985; Hansen, 1987; Fienberg & Tanur, 1995, 1996). First, its struggle to be accepted as a valid method in a field that is accustomed to full-coverage methods (census); Second, the confrontation between the two different paradigms inside the field, random selection and purposive selection.

A notorious and early proposition of the use of a sample instead of a full census was done by A.N. Kiaer (1896, 1899, 1903, 1905) at the 5th International Statistical Institute (ISI) meeting in Bern in 1895. Sampling methods would be more widely accepted thirty years later, in 1925, when some results are presented at the 16th ISI Sessions in Rome. These results are those of the Reports on the representative method in statistics, proposed by A. Jansen and a committee of five other statisticians (Jensen, 1926). Corrado Gini is one of them.

The history of sampling theory has been driven by several concepts, often associated to unclear meanings. There is frequent confusion and misunderstanding between the notions of representativeness, purposive selection, random selection or balanced samples. These concepts, as well

as their evolution across the history of survey methods, are discussed in the following sections.

### 4.2.2 Purposive and random selection

In addition to eventually giving credit to the idea of sampling, the report by Jensen (1926) already opposes two methods of sampling: purposive selection and random selection. This separation has been the vector of a large amount of researches on sampling.

When speaking of random selection it is firstly of importance to distinguish between the population of interest and the sampling frame. The sampling frame is the list of units from which the sample is to be drawn. For example, if a survey on male adults of a given country (the population of interest) is to be done, a list of all male adults in the country has to be generated before sampling can be effective. Ideally, the sampling frame corresponds fully to the population of interest, but for various reasons under-coverage (units that are in the population but not in the frame) or over-coverage (units that should not be in the frame) is possible.

The selection process is referred to as random when all units in the sampling frame have a non-null probability of being selected in the sample and that this inclusion probability can be precisely established. Moreover, the scheme is such that every possible subset (e.g. sample) $s$ of the population $U$ has a probability of selection, denoted $p(s)$. Unlike in some purposive methods, the interviewer does not take any part in the selection process of the sample, which is operated by an algorithm. Based on mathematical validation and allowing for the construction of confidence intervals, random sampling is soon preferred in the scientific community, as well as in official statistics. However, purposive selection is still used nowadays, mostly via quota sampling methods.

Purposive selection (or non-probability sampling), regroups any sampling method in which the inclusion probabilities are not known or in which some units have no chance to be selected in the sample. These methods are often used by private polling organizations because these organizations generally do not have access to the sampling frame from which a sample can be drawn (a census for example) and also because the total cost of the survey can be lowered when using non-probability methods.

## 4.3 Representativeness

### 4.3.1 A polysemic term

The idea and concept of *representativeness* was already used in Kiaer's work (Kiaer, 1896, 1899, 1903, 1905). Because the idea of a *representative*

*sample* is reassuring for an uninitiated audience as it provides an illusion of scientific validity, it has been an important notion in sampling ever since. However, the multiplicity of definitions to which it can be associated has been at the core of many debates and misunderstandings in the history of sampling. Thus, the term is much less used in modern survey sampling literature and in our opinion it is a term best to avoid in survey methodology.

Kruskal & Mosteller (1979a,b,c, 1980) have written several survey papers in which they typologize the different definitions of representativeness in and out of the field of statistics. They have listed nine different views of the word and illustrated them by numerous excerpts from the literature. From their work, one can also see that the definition of representativeness and the question of purposive or random selection are often directly linked.

### 4.3.2 Representativeness of the sample

One point of interest which is seldom clarified is the statistical object to which the quality of representativeness is attributed. It could be the statistical unit to be sampled (a representative individual, a representative district or firm), the sample itself or the strategy used for the sample selection procedure. Indeed, for example, using a method of sampling which supposedly provides a representative sample and analyzing a posteriori the level of representativeness of an accomplished sample are two very different conceptions of the problem. Also considering the question of randomness and representativeness, Hájek (1959, 1981) invokes the term of representative strategy (and not representative sample) when talking about sampling methods like balancing or calibration, a strategy being a pair composed of a sampling design and an estimator. When used hereafter, the term of representativeness is applied to the sample, except where specified.

### 4.3.3 Representativeness as a miniature of the population

Although many decisive differences occur between the authors, the underlying idea of representativeness in most cases is that to be considered representative a sample should be a miniature of the population of interest. Obviously, this is a purely theoretical construct, which creates many divergences in its application. Indeed, in order to select a sample which would be a downsized replica of the population, the whole complexity of the population structure has to be known, which is impossible. Generally, only a small aspect of this structure is known, turning representativeness into a much more relative notion. Moreover, the more information one has in the population, the more dissimilar are each of its units. In the

end, the only perfectly representative sample would thus be the population itself. Also, as discussed further in this section, the results of Neyman (1934) in optimal stratification show that, in fact, expecting a sample to be a miniature of the population is an erroneous ideal.

In practice, it is therefore only possible to come close to this view on representativeness and many different solutions have been proposed to achieve it, some of which being totally antagonistic. These methods depend obviously on whether purposive selection or random selection is preferred, but also on the different interpretations of representativeness. An example of these divergences is well described by the question of the presence or absence of selective forces, raised by Kruskal & Mosteller (1979a,b,c, 1980) and discussed below.

### 4.3.4 Selective forces

It can be advocated that the complete absence of a selective process involving any kind of auxiliary information is a guarantee of representativeness. This idea fosters the use of simple random sampling, because in that framework every unit in the population has the same probability of being selected in the sample, regardless of any of the unit's characteristics. Here, selective forces are viewed as a curb to representativeness, and the definition of the latter is partially confused with randomness. Moreover, the notion of randomness itself seems bounded to simple random sampling, because other designs like stratification or unequal probability sampling rely on selective forces.

On the contrary, other definitions see in selective forces a necessary requirement to guarantee representativeness. For instance, this can be illustrated by the widespread idea that the more a sample presents the same characteristics (for example gender ratio, age groups, mean income) as the population of interest, the more it can be declared representative. For the latter, perfect representativeness is not accessible in practice, but can nonetheless be understood as a miniature of the population and approached by using the available information as controls.

Likewise, some have an intermediate position. More than two decades after his work with Corrado Gini (Gini & Galvani, 1929), which is detailed in later sections of this paper, Galvani (1951) distinguishes three different procedures of sampling: random selection, purposive selection and stratified selection.

In his paper, random selection is what is more often known as simple random sampling. It uses no auxiliary information, and therefore requires no prior knowledge of the population to be sampled. For the author, it is also the procedure which is closest to the idea of reduction of the totality, where the totality would be the whole population and all its characteristics. According to Galvani, this statement does not imply

that any achieved sample from a simple random sampling design is effectively representative for all characteristics of the population. It could be therefore argued that Galvani is unclear whether representativeness is an attribute of the accomplished sample or of the method of selection.

Purposive selection, instead, requires some auxiliary knowledge on the whole population which are used to guarantee the representativeness of the sample in relation to these auxiliary variables. However, Galvani notes that, unlike in the random selection, the purposive procedure can by no means be considered representative for characteristics that are not used in the selection process. If we summarize Galvani's point of view, he considers on one side simple random sampling which, as a *method*, is representative for all characteristics in the population, and on the other side, purposive selection which yields a *sample* that can be described as representative for only a closed number of characteristics. This leads him to oppose absolute representativeness (random selection) and relative representativeness (purposive selection), the latter being described as inferior to the former.

Stratified sampling is coherently considered by Galvani as a random selection procedure which makes use of some knowledge of the heterogeneity of the population regarding some characteristic. The property of "impartiality", or reduction of the totality, can also be credited to the stratified sampling method, with presumably a better accuracy than for simple random sampling. In other words, Galvani thinks that stratification is not a type of selective force which can jeopardize representativeness. He states that random selection methods should be preferred because they satisfy fully to the conditions of representativeness. Moreover, probability theory can be applied to the random case.

### 4.3.5 Representativeness as a purpose

In the Kruskal & Mosteller (1979a,b,c, 1980) papers emphasis is not put on one essential question: should representativeness be a purpose for survey sampling? From our point of view, the goal of a survey is to provide good estimations (in terms of bias and variance) of some characteristics of the population by means of a sample. In that framework, representativeness can possibly be a means for achieving good estimations, not a goal in itself.

Probability sampling theory and an adequate use of auxiliary information has shown that using unequal probabilities of inclusion or over-representing some groups of the population can enhance the accuracy of the estimation. In that sense, a sample can be obviously very far from being a miniature of the population and estimate efficiently the characteristics of interest. The famous paper by Neyman (1934) stresses two major assets of probability sampling theory: only probability sampling methods

can be theoretically validated because they are the outcome of a random experiment; the results on optimal stratification prove that a representative sample is not optimal in terms of accuracy. The latter result, which can be viewed as counter-intuitive is a major defeat for non-probability sampling, as well as for the idea of representativeness as a purpose in the field of survey sampling.

## 4.4 Balanced sampling

### 4.4.1 A broad definition

Representativeness and randomness are both major debates of sampling theory and the development of balanced sampling has certainly changed their physiognomy. Balanced sampling, as defined below, has put the conception of the representative sample as a miniature of the population in practice, first under the purposive selection paradigm, than in the random sampling framework. A balanced sampling strategy is a method of selection which uses auxiliary information at the design phase. Moreover, a sample is said to be balanced if its natural estimators of total on some auxiliary variables $X$ will match (or approximately match) the known population totals of these variables. What is meant here by a natural estimator is an estimator such that the weight of any given unit does not change from a sample to another. The main challenge of balanced sampling is the selection process, because the sample has to be selected with respect to the balancing constraints.

This broad definition is consistent with purposive as well as random methods. A more formal definition for the random case is given in Section 4.7.2 when the Cube method is introduced. Coherently, balanced sampling is at first classified as a purposive selection method. One of the first known application of balanced sampling, proposed in Italy by Gini & Galvani (1929) has enhanced this idea. Later, it will be shown that *balanced* and *random* are not two mutually exclusive concepts. These findings have modified the definition of randomness in sampling, which in the beginning of sampling theory was mostly reserved to simple random sampling.

### 4.4.2 Gini and the beginnings of balanced sampling

Not long after the ISI Session in Rome in 1925, a new census is to be run in Italy and room has to be made for the new data. The Italian statistical office wants, however, to keep a *representative* sample of the previous census of 1921. To do so, Gini & Galvani (1929) use a method referred to as purposive selection by Neyman (1952), but also the beginnings of the idea of balanced sampling (Yates, 1960). At that time, Italy is separated into 214 circumscriptions. The authors propose to keep a sample containing

all units inside 29 circumscriptions. The 29 circumscriptions are not selected at random, but instead they are selected in order to match the population means of some important variables. Indeed, the authors selected seven control variables and selected the sample of 29 circumscriptions in order for the sample means to be as close as possible to the population means. As a result, they realize that for most other variables that were not included in the balancing procedure, the match between the population mean and the sample mean is very poor.

Both Neyman (1952) and Yates (1960) discuss the paper of Gini & Galvani (1929) to condemn the purposive selection method. In a different way, they both stress out that, from a sampling point of view, the selection of 29 circumscriptions is a small sample of huge units. Moreover, Yates underlines the fact that the reliability of the results is not assessable with the purposive selection method. Neyman recalls that to obtain a reliable sample, "we must rely on probability theory and work with great numbers" (Neyman, 1952, p.107). Whereas the total number of people in the balanced sample of Gini and Galvani is very large, it remains a small sample of 29 units from a sampling point of view. Of course, Neyman advocates that the circumscriptions should have been considered as strata, instead of sampling units.

### 4.4.3 Balanced sampling towards randomness

Although representativeness and purposive selection are often associated, Yates (1960, p.39) has stressed that their is no contradiction between the balanced procedures and random sampling. Furthemore, according to Yates, a balanced sample is only satisfactory if it is random. He proposes a method for selecting a random balanced sample. The randomization is done by first drawing a preliminary random sample and then by selecting a further unit. The latter is compared to the first unit selected in the original sample. If the new unit improves the balance, it is kept in the sample in place of the original unit. If not, it is rejected. Then, the second unit in the original sample is compared to another new unit, and so on. This procedure is repeated until the balance is considered satisfactory. An appealing remark from Yates is also that purposive selection balanced sampling lead to the selection of units for which the balanced variables take a value close to the population mean, resulting with a problematic smaller variability in the sample than in the population.

While Galvani (1951) discussed three different kinds of sampling (see Section 4.3.4), a similar classification can be found in an article by Royall & Herson (1973) which compares three kinds of balanced samples. The first category is purposive selection (like in Gini and Galvani's experiment), the second category is random sampling and the third is what the authors call restricted randomization. For the authors, (simple) random

sampling provides a balanced sample "on the average" (Royall & Herson, 1973, p.887), but despite the fact that adjustments can be done with post-stratification, the method can produce severely unbalanced samples. This reasoning on balancing and simple random sampling is very close to Galvani's statement on representativeness. We think however, that the argument of simple random sampling being balanced in average is pointless. It seems to be simply another way of saying that the estimators are unbiased under the sampling design. Furthermore, whereas this is true for total or functions of totals, it is not true for other type of statistics such as ratios, inequality indicators or quantiles.

Restricted randomization, on the other hand, is defined here as a selection process which provides an approximately balanced sample without renouncing to randomization. The selection process is as proposed by Yates (1960) and discussed above. Royall & Herson (1973) point out that none of the methods can however guarantee that a balanced sample is also balanced on variables that are not included in the selection process. Indeed, the balanced sample can mimic the population only to a certain extent, which depends greatly on the availability of the auxiliary information. For the authors, the experiment by Gini and Galvani is a notorious example of an approximately balanced sample which was found out to be unbalanced for numerous other external characteristics.

This discussion on balanced sampling requires some remarks on the variance of the estimator. Indeed, it has been shown (Deville & Tillé, 2005; Fuller, 2009) that the variance under a balanced sampling design can be expressed as the variance of regression residuals. It is thus obvious that the more the auxiliary variables are correlated with the character of interest, the more the variance is reduced. Therefore, even if the sample is not balanced on a particular variable, the variance of the sample mean and total of that variable is nevertheless reduced if a correlation exists between the latter variable and the auxiliary information.

In the last two decades, a lot of research has been conducted on the idea that balanced sampling is not in contradiction with the random sampling framework. Modern methods allow indeed for the selection of balanced samples and at the same time stay consistent with the notion of randomness (Deville, 1988, 1992; Deville et al., 1988; Deville & Tillé, 2004; Hedayat & Majumdar, 1995; Nedyalkova & Tillé, 2008; Tillé & Favre, 2004, 2005). It is therefore clear that the history of balanced sampling has somewhat blurred the initial separation between purposive selection and random sampling.

## 4.5 The Gini index

### 4.5.1 An inequality measure

Although we have, in this paper, introduced Corrado Gini through his work in balanced sampling, his most famous contribution is undoubtedly the Gini coefficient or Gini index (Gini, 1912, 1914, 1921), an inequality index based on the Lorenz curve (Lorenz, 1905). An impressive amount of literature has been written on the Gini index (for a survey paper see Xu, 2004), which is still nowadays the most commonly used inequality measure. Finite population inference for the Gini index has been the center of countless discussions and papers. Indeed, many finite population expressions of the index co-exist. Moreover, variance estimation is not straightforward, and robustness issues are frequent, especially when working with income data which is known to be generally heavily skewed.

The contribution of Corrado Gini's index to inequality measure is immense. It has been the most widely used inequality measure for now nearly a century. Its graphical interpretation, its synthetical comprehension (often written as a percentage, where 0% is full equality and 100% perfect inequality) and the fact that it satisfies many properties of the axiomatic approach to inequality (Cowell, 1988; Dalton, 1920; Pigou, 1912; Shorrocks, 1980, 1984) has favored its preeminence inside the field of inequality theory. Although other measures have gain interest more recently, the Theil index for example (Theil, 1967, 1969) because of its subgroup decomposability property, or the Quintile Share Ratio (Eurostat, 2005; Hulliger & Münnich, 2006; Langel & Tillé, 2011d), because of its simpler interpretation for non-specialists, the Gini index is still by far the most applied and studied measure of income inequality.

### 4.5.2 Continuous case

There is at least three ways of apprehending the Gini index. The first one is based on the Lorenz Curve, which graphically represents the share of total income earned altogether by a given share of income earners, ordered from poorest to richest. For example it is possible that the poorer 75% of the population earn only 25% of the total income. This understanding of the Gini index is the most common because it is graphically straightforward and gives an immediate definition for the continuous case. Indeed, the Gini index is the ratio of the area between the Lorenz Curve of the population of interest and the diagonal (the Lorenz Curve under perfect equality) and the area below the diagonal. The latter area is always equal to $1/2$. Therefore, considering:

(1) a continuous and differentiable strictly increasing cumulative distribution function $F(y)$ of income $y$ in $\mathbb{R}^+$, and $f(y)$ its derivative and probability density function,

(2) $Q_\alpha$, the quantile of order $\alpha$, such that $F(Q_\alpha) = \alpha$ and the quantile function $Q(\alpha)$, which can be written as the inverse of the cumulative distribution function: $Q(\alpha) = F(\alpha)^{-1}$,

(3) the Lorenz function (or quantile share) $L(\alpha)$, which is the share of total income earned by all the income earners up to quantile $\alpha$

$$L(\alpha) = \frac{\int_0^{Q_\alpha} uf(u)du}{\int_0^{\infty} uf(u)du}, \tag{4.1}$$

the definition of the Gini index for an infinite population is then:

$$G = 2\left(\frac{1}{2} - \int_0^1 L(\alpha)d\alpha\right). \tag{4.2}$$

### 4.5.3 Discrete case

For the purpose of measuring inequality in a finite population, the partial sum appears to be a central tool. We propose hereafter two definitions of the partial sum, which can in either case be understood as the sum of all incomes smaller or equal to a given quantile. To link the discrete and continuous cases, it can be noted that the partial sum in the continuous case would be expressed by the numerator of the Lorenz function (4.1).

Let $U$ define a finite population of size $N$, and $y_k$ the income (or other characteristic of interest) of unit $k \in U$. The incomes $y_k$ are assumed to be sorted in increasing order such that $k$ also denotes the rank. A natural expression for the partial sum would therefore be:

$$\sum_{k \in U} y_k \mathbb{1}\left(y_k \leq Q_\alpha\right),$$

where $Q_\alpha$ is the quantile of order $\alpha$ and where $\mathbb{1}(A)$ is equal to 1 if $A$ is true and 0 otherwise. Quantile $Q_\alpha$ can be defined in many different ways in the finite population context (see for example Hyndman & Fan, 1996). The other definition below has two good properties: it gets around the above issue of the finite population quantile and it is strictly increasing upon $\alpha$:

$$Y(\alpha) = \sum_{k \in U} y_k H\left[\alpha N - (k-1)\right], \tag{4.3}$$

where

$$H(x) = \begin{cases} 0 & \text{if} \quad x < 0, \\ x & \text{if} \quad 0 \leq x < 1, \\ 1 & \text{if} \quad x \geq 1, \end{cases} \tag{4.4}$$

is the cumulative distribution function of a uniform distribution in $[0;1]$.

A specific case for (4.3) is the population total, simply denoted by $Y$:

$$Y(1) = \sum_{k \in U} y_k = Y.$$

The second property leads to an important result for the Lorenz Curve. The Lorenz Curve, which can be simply written by

$$L(\alpha) = \frac{Y(\alpha)}{Y},$$

is indeed strictly increasing and convex in $(0;1)$. Recalling (4.2), the Gini index for the continuous case, an expression for a finite population is:

$$G = 2\left(\frac{1}{2} - \frac{1}{Y}\int_0^1 Y(\alpha)d\alpha\right). \tag{4.5}$$

With Expression (4.3), the Gini index becomes:

$$G = \frac{2}{YN}\sum_{k \in U} k y_k - \frac{N+1}{N} = \frac{\sum_{k \in U}\sum_{\ell \in U}|y_k - y_\ell|}{2NY}.$$

### 4.5.4 Sampling from finite population

The level of inequality of a finite population, a country for example, is most often estimated by means of a sample. National statistic institutes around the world usually work with complex sampling designs which allow for the use of auxiliary information to enhance accuracy of the statistics of interest. For this reason, estimation and variance estimation of the Gini index under complex sampling designs has been studied profusely.

A sample $s \subset U$ of size $n(s)$ is a subset of the population. A random sample $S$ is selected from $U$ by means of a sampling design $p(s) \geq 0$, for all $s \subset U$ and

$$\sum_{s \subset U} p(s) = 1.$$

The inclusion probability $\pi_k$ is the probability of unit $k$ to be in the sample. Inclusion probabilities are defined by the sampling design and are such that

$$\pi_k = \sum_{s \ni k} p(s), \text{ for all } k \in U.$$

Some sampling designs are of fixed size, e.g. $\text{var}[n(s)] = 0$. For these designs, the sample size is simply denoted as $n$. Also, if the design is of fixed size, then

$$\sum_{k \in U} \pi_k = n.$$

In sampling theory, a classical estimator of total $Y$ is

$$\widehat{Y} = \sum_{k \in S} w_k y_k,$$

where $w_k$ denotes the sampling weight of unit $k$. The weight can simply be the inverse of inclusion probabilities (the estimator of total is then known as the Horvitz-Thompson estimator) but can also account for calibration and non-response adjustments. A natural estimator of $L(\alpha)$ would then be

$$\widehat{L}(\alpha) = \frac{\widehat{Y}(\alpha)}{\widehat{Y}},$$

where $\widehat{Y}(\alpha)$ is the plug-in estimator of (4.3):

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

where $H$ is as defined in (4.4) and $\widehat{N}_k = \sum_{\ell \in S} w_\ell \mathbb{1}(y_\ell \leq y_k)$. An important feature of this estimator is that the cumulative weight $\widehat{N}_k$ is an estimator of the rank of unit $k$ in the population. The estimation of the rank is one of the most sensible issues in the estimation of the Gini index in the sampling framework. If omitted, this issue leads to substantial errors in the estimation of the sampling variance of the index. Note also that $\widehat{N} = \sum_{k \in S} w_k$. An estimator for (4.5) is then defined by:

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

## 4.6 Variance estimation for the Gini index

### 4.6.1 Linearization: the influence function approach

The Gini index is nearly one century old and has been used in countless empirical applications since then. However, within these applications, it is not uncommon to find only point estimators, lacking variance or standard deviation estimations, which are nevertheless necessary for confidence intervals construction. The question of variance estimation for the Gini index is not trivial and has motivated a great amount of research (for example Hoeffding, 1948; Glasser, 1962; Sandström et al., 1985; Deville, 1996; Ogwang, 2000; Dell et al., 2002; Giles, 2004; Leiten, 2005; Berger, 2008).

One of the main methods of variance estimation of complex statistics in sampling from finite population is the linearization technique. Based on Taylor series, the method has been introduced by Woodruff (1971) and has since motivated many publications presenting different approaches.

Estimating equations (Binder & Patak, 1994; Binder, 1996; Kovacevic & Binder, 1997), influence functions (Deville, 1999) and the Demnati-Rao approach (Demnati & Rao, 2004) are all approaches which can be, or have been, applied to variance estimation for the Gini index in complex surveys.

A generalized linearization method based on influence functions has been developed by Deville (1999). The method is derived from the influence function as defined in the field of robust statistics (Hampel et al., 1985). The influence function in Deville (1999) is slightly different from the latter and is defined by

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

where measure $M$ allocates a unit mass to each $x_k$, $T$ is a functional associating a real number or a vector to each measure, and $\delta_x$ the Dirac measure at $x$. In the sampling framework, measure $M$ is estimated by $\widehat{M}$ with mass $w_k$ at each point $x_k$ in the sample. Moreover, the plug-in estimator of functional $T(M)$ is simply $T(\widehat{M})$. The influence function $z_k = IT(M, x_k)$ is a linearized variable of $T(\widehat{M})$ and

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

The variance of $T(\widehat{M})$ can thus be approximated by the variance of the total of the linearized variable

$$\text{var}[T(\widehat{M})] \approx \text{var} \left( \sum_{k \in S} w_k z_k \right).$$

Thus, the variance of a complex statistic can be estimated under any sampling design for which the expression of the variance of a total is available. Generally, however, the computation of the influence function requires information on the whole population, not only on the sample. Therefore, plug-in estimators $\hat{z}_k$ are used instead of $z_k$.

### 4.6.2 Linearization of the Gini index

The Gini index is not an easy expression to linearize. Solutions based on Taylor linearization resulted in a greatly over-estimated variance (Nygård & Sandström, 1985; Sandström et al., 1985, 1988). Some further results use estimating equations (Kovacevic & Binder, 1997) or introduce the influence function approach (Monti, 1991; Deville, 1999). Applying the latter approach, a linearized variable for the Gini index is

$$z_k = \frac{1}{NY} \left[ 2N_k(y_k - \overline{Y}_k) + Y - Ny_k - G(Y + y_k N) \right]. \tag{4.6}$$

As emphasized previously, this expression involves unavailable information at the population level. It can be substituted by the plug-in estimator

$$\widehat{z}_k = \frac{1}{\widehat{N}\widehat{Y}}\left[2\widehat{N}_k(y_k - \widehat{\overline{Y}}_k) + \widehat{Y} - \widehat{N}y_k - \widehat{G}(\widehat{Y} + y_k\widehat{N})\right],$$

with

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

The linearization approach has proved its efficiency for variance estimation in several simulation studies in the literature (Dell et al., 2002; Osier, 2006, 2009; Berger, 2008).

## 4.7 Balanced sampling and the Gini index

### 4.7.1 Linking two of Gini's main contributions

The estimation of the Gini index by means of a sample is of importance to provide reliable information on the level of income inequality in a population. However, the Gini index is known to be very sensitive to high incomes and the stability of the estimator is therefore very dependent upon the presence of extreme values in the sample. Moreover, when the income distribution contains outliers (very high incomes), the sampling distribution of the Gini index becomes skewed, which creates difficulties when constructing confidence intervals, even if the variance of the index is correctly estimated. In this setup, the choice of the sampling design is crucial. Balanced sampling has proved to make good use of auxiliary information when available. However, the method can be only applied to population totals of auxiliary variables. The idea of Lesage (2008) is to propose methods of balanced sampling for non-linear statistics.

### 4.7.2 The Cube method

Deville & Tillé (2004, 2005) have proposed a general procedure to select a balanced sample called the Cube method. The algorithm is non-rejective and selects a sample with respect to the balancing constraints

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} = \sum_{k \in U} x_{kj},$$

for all auxiliary variable $j = 1, ..., p$. As such, balancing is operated by Horvitz-Thompson estimators of totals for the auxiliary variables. Moreover, the statistic of interest is usually also a total (or a function of a total), say $\widehat{Y}$. If the character of interest $y_k$ is well explained by the auxiliary information $x_{kj}$, the variance of $\widehat{Y}$ is supposedly small.

### 4.7.3 Non-linear balancing constraints

In the case of estimating a strongly non-linear function like the Gini index, the use of available auxiliary information is also desired. It can happen that when estimating the Gini index for a population at a time $t$, the incomes of a previous period in time $t - \delta t$ are known in the population. Denoting $x_k$, the income of a previous year for unit $k$, a simple use of this auxiliary information at the design phase would be to balance the sample on the total $X = \sum_{k \in U} x_k$. However, being eventually interested in the estimation of the Gini index, it seems more favorable to balance, not on the total $X$, but on the Gini index of the $x_k$'s. Lesage (2008) uses for balanced sampling a well-known idea of variance estimation for complex statistics. Indeed, the sampling variance of a statistic $\hat{\theta}$ is easily expressed under a variety of complex sampling designs as long as $\hat{\theta}$ is a total or a function of totals. If instead, $\hat{\theta}$ is a non-linear statistic, the popular approach of linearization consists in bringing the problem back to one of a total, by using the linearized variable.

In the balancing procedure, a similar trick can be used: to balance on a non-linear statistic, the statistic of interest can be linearized and the total of the linearized variable used as a balancing constraint. Moreover, this method does not require any additional computing tools as for standard balanced sampling.

## 4.8 Simulation studies

A simulation study has been operated on real household taxable income data from the canton of Neuchâtel, Switzerland to show the relevance of using non-linear balancing constraints. Complete data is available for a population of $N = 82,489$ households for two consecutive years, 2005 and 2006. The goal of this simulation study is to estimate the Gini index in 2006 by means of a sample and evaluate if balancing procedures are able to reduce the variance of the estimator. The character of interest, denoted $y_k$, is the income of year 2006 for household $k$. The auxiliary information is available at the population level and expressed as follows:

- $\pi_k$: the inclusion probability of unit $k$,
- $x_{k1}$: the income of year 2005 for unit $k$,
- $x_{k2}$: the linearized variable for the Gini index, as expressed in (4.6), of year 2005 for unit $k$.

The results for four different sets of simulations are compared. All four simulations consist in drawing 1000 samples of fixed size $n = 5000$. Equal inclusion probabilities $\pi_k = n/N$, for all $k \in U$, are used across all the simulations. The sampling strategies concerning balancing constraints are

detailed in Table 4.1. Balanced samples have been selected using the algorithm of the Cube method (see Section 4.7.2). With the equal inclusion probabilities used here, Simulation 1 is equivalent to a simple random sampling design without replacement.

Table 4.1 – *Simulations: description of the sampling strategies*

| | Inclusion probabilities | Balancing variables |
|---|---|---|
| Simulation 1 | $\pi_k$. | $\pi_k$. |
| Simulation 2 | $\pi_k$. | $\pi_k$, $x_{k1}$. |
| Simulation 3 | $\pi_k$. | $\pi_k$, $x_{k2}$. |
| Simulation 4 | $\pi_k$. | $\pi_k$, $x_{k1}$, $x_{k2}$. |

The results of the simulation study is summarized in Table 4.2. The second column presents the relative bias of the estimated Gini index $\widehat{G}$ computed over all 1000 samples in each simulation set expressed as

$$\mathrm{RB}(\widehat{G}) = \frac{\mathrm{E}_{sim}(\widehat{G}) - G}{G},$$

where $\mathrm{E}_{sim}(\widehat{G})$ is the mean of the estimated Gini index over the 1000 samples and $G$ is the true value of the Gini index in the population. The last column describes the gain in terms of sampling variance of Simulations 2, 3 and 4 relatively to simple random sampling (Simulation 1).

Table 4.2 – *Simulation results*

| | $\mathrm{RB}(\widehat{G})$ | $\mathrm{var}_{sim}(\widehat{G})/\mathrm{var}_{sim1}(\widehat{G})$ |
|---|---|---|
| Simulation 1 | −0.009% | 1.000 |
| Simulation 2 | 0.036% | 0.773 |
| Simulation 3 | 0.000% | 0.472 |
| Simulation 4 | −0.020% | 0.451 |

The four Monte Carlo simulations show that the bias of $\widehat{G}$ is negligible. The sampling variance of the estimator is lowered for all balanced designs in comparison with the simple random sampling design (Simulation 1). While balancing on $x_{k1}$, the income of the previous year, clearly has an effect on the variance (Simulation 2), the most important result here is that balancing on $x_{k2}$, the linearized variable of the Gini index of the previous year yields a far better improvement (Simulation 3). The same initial auxiliary information is available in both Simulations 2 and 3, but the use of the linearized variable presented in Section 4.6 instead of the plain incomes of year 2005 gives a definitely more stable estimator. In our case, balancing on $x_{k2}$ results in a sampling variance which is more than twice

smaller than with simple random sampling. Finally, balancing on both $x_{k1}$ and $x_{k2}$ (Simulation 4) gives the best result but essentially shows that when $x_{k2}$ is used, also adding $x_{k1}$ as a balancing constraint does not bring much gain in terms of variance.

## 4.9 Conclusion

In this paper, we have reviewed two of Corrado Gini's main contributions and how they have participated in the development of their respective field. For a start we have discussed the ambiguous notions of representativeness, randomness and balanced sampling in order to get the article of Gini & Galvani (1929) back in its perspective and show how it has participated in the debates that have resulted in modern survey sampling theory.

Secondly, we have presented the Gini inequality index and its application in the sampling framework. The linearization of the Gini index through the influence function method is also introduced for two distinct goals. The first one is variance estimation, which is briefly discussed. The second one is for balanced sampling, which is of particular interest in this paper. Indeed, we have shown how balanced sampling and the use of a linearized variable as a balancing constraint can improve estimation. With the help of a simulation study on real data we have shown that, if the information is available, balancing on the income of a previous year when estimating the Gini index has a positive impact on the stability of the estimator. However, we have also shown that this auxiliary information can be used in a much better way, that is using the linearized Gini index of a previous year as balancing constraints instead of the plain incomes.

## Acknowledgements

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

# 5 Variance estimation of the Gini index: Revisiting a result several times published

**Abstract**

Since Corrado Gini suggested the index that bears his name as a way of measuring inequality, the computation of variance of the Gini index has been subject to numerous publications. In this paper, we survey a large part of the literature related to the topic and show that the same results, as well as the same errors, have been republished several times, often with a clear lack of reference to previous work. Whereas existing literature on the subject is very fragmented, we regroup papers from various fields and attempt to bring a wider view of the problem. Moreover, we try to explain how this situation occurred and the main issues involved when trying to perform inference on the Gini index, especially under complex sampling designs. The interest of several linearization methods is discussed and the contribution of recent papers is evaluated. Also, a general result to linearize a quadratic form is given, allowing the approximation of variance to be computed in only a few lines of calculation. Finally, the relevance of the regression-based approach is evaluated and an empirical comparison is proposed. [1]

**Keywords:** Gini, inequality, variance estimation, linearization, influence function, sampling

## 5.1 Introduction

Despite the fact that the Gini index is the most widely used indicator of income inequality in a population, it is not a trivial measure to handle.

[1] This chapter is a reprint of: Langel, M. and Tillé, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society: Series A,* In press.

To start with, its construction is not easy to understand for an uninitiated audience. Also, because the Gini index of a population is commonly estimated by means of a sample, its estimation should be completed by some knowledge of the accuracy of the point estimate, namely by reporting a variance estimator or standard error, allowing for the computation of a confidence interval. Since the Gini index is a non-linear function of interest, variance estimation is not straightforward, especially when data is collected by means of a complex sampling strategy. Thus, the computation of the sampling variance of the Gini index has prompted a great amount of research in statistics and economics.

Although the problem is now mostly solved, the result has been republished many times in recent years. The original motivation of this paper is to understand why. For that purpose, we survey a large portion of the literature on the subject in a historical perspective. A thorough analysis of the evolution of the literature has stressed two important features: the main methodological issue in the computation of variance of the Gini index has sometimes been overlooked, and a serious lack of references to previous works is often witnessed.

One of the main contributions of this paper is to present and compare many different approaches to variance estimation of the Gini index. Additionally we give a general result on the linearization of a quadratic form which allows the initially intricate problem to become computationally quite simple. We also show why the regression-based methods, which have attracted a lot of attention recently, should be avoided. Finally, while publications on the topic have arisen from different fields, we try to propose a global clarification of the present state of the art on the subject.

In the next section, the Gini index and estimators are presented. A discussion on the evolution and issues regarding variance estimation of the index is proposed in Section 5.3. Section 5.4 is dedicated to linearization techniques, which encompass a variety of approaches for deriving a variance estimator. Many approaches are presented and a new result allowing for a fast linearization of the Gini index is suggested. The relevance of recent publications is also discussed. In Section 5.5, we examine the so-called regression approach which has prompted a lot of recent research studies. Finally, a comparative numerical application is performed in Section 5.6. In order to show the shortcomings of the regression approach, an empirical study was conducted using the same data as in two recent papers (Giles, 2004; Davidson, 2009). The paper ends with some concluding remarks.

## 5.2 The Gini index

### 5.2.1 Definition and estimation in an infinite population

Like several other inequality measures, the Gini (1912, 1914, 1921) index is based on the cumulative income of a proportion of the poorer units. Let $f(y)$ be a probability density function of a positive continuous random variable $Y$ that represents the income and $F(y)$ its cumulative distribution function. First define the Lorenz (1905) curve given by

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

where $F^{-1}(.)$ is the inverse function of $F(.)$ and

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

The Gini index can be defined in several ways (see, for example Xu, 2004):

$$G = 2\int_0^1 [\alpha - L(\alpha)]\, d\alpha = 1 - 2\int_0^1 L(\alpha)d\alpha. = \frac{2}{\mu}\int_0^{\infty} yF(y)f(y)dy - 1. \tag{5.1}$$

If $y_i, i = 1, \ldots, n$ is a sequence of positive random variables with the same probability density function $f(.)$, the Gini index $G$ can be estimated by

$$\widehat{G} = \frac{2\sum_{i=1}^n iy_{(i)}}{n\sum_{i=1}^n y_{(i)}} - \frac{n+1}{n}. \tag{5.2}$$

where the $y_{(i)}$ are the $y_i$ ordered in increasing order. The above expression can be found for example in Sen (1973) or Fei et al. (1978). A complete review of all the expressions of the Gini index proposed originally by Corrado Gini is found in Ceriani & Verme (2011). Note also that a controversy exists on the use of $\widehat{\widetilde{G}} = n\widehat{G}/(n-1)$ instead of $\widehat{G}$. Indeed, $\widehat{\widetilde{G}}$ is less biased than $\widehat{G}$, but the latter will be used hereafter because the construction of $\widehat{\widetilde{G}}$ becomes difficult when the observations are weighted.

### 5.2.2 Definition and estimation in a finite population

The Gini index is generally estimated by means of a sample survey. In survey sampling, we are interested in a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size $N$ from which a random sample $S$ of size $n$ is selected by means of a sampling design $p(s) = \Pr(S = s)$, for all $s \in U$. Let also $\pi_k = \Pr(k \in S)$ denote the inclusion probability of unit $k \in U$ and $d_k = 1/\pi_k$ the Horvitz & Thompson (1952) weights. In survey methodology, the observations are usually weighted. The sampling weight $w_k$ associated with observation $k$ can be equal to $d_k$ or can be improved by a

calibration technique (Deville & Särndal, 1992) or a non-response adjustment. Let $y_1, \ldots, y_k, \ldots, y_N$ denote the incomes of the units in the population. Let also $N_k$ and $n_k$ denote the rank of unit $k$ in population $U$ and in sample $S$ respectively, with tied observations treated using increasing integer ranks such that, for example, the series of observations $4, 6, 6, 8$ would be assigned ranks $1, 2, 3, 4$. In order to estimate totals

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

one can use weighted estimators

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

The total income of the $\alpha N$ poorest units is defined by

$$\widetilde{Y}(\alpha) = \sum_{k \in U} y_k \mathbb{1}(y_k \leq Q_\alpha),$$

where $Q_\alpha$ is the $\alpha$-quantile and $\mathbb{1}(A)$ is an indicator function equal to 1 if $A$ is true and 0 otherwise. However, this definition is not very accurate because the quantiles can be defined in several different ways when the cumulative distribution function is a step function (for a review, see Hyndman & Fan, 1996). We thus prefer to use the less ambiguous definition of the total income of the $\alpha N$ poorest units proposed in Langel & Tillé (2011d) and given by

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

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

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

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

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

where $\widehat{N}_k$ are the cumulative weights of the sampled units, i.e.

$$\widehat{N}_k = \sum_{\ell \in S} w_\ell \mathbb{1}(n_\ell \leq n_k), \tag{5.3}$$

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

$$L(\alpha) = \frac{Y(\alpha)}{Y}, \tag{5.4}$$

and can be estimated by

$$\widehat{L}(\alpha) = \frac{\widehat{Y}(\alpha)}{\widehat{Y}}. \tag{5.5}$$

Accordingly, functions $L(\alpha)$ and $\widehat{L}(\alpha)$ are also strictly increasing in $(0,1)$. If we use the definition of the Lorenz curve given in (5.4) and the Gini index as given in (5.1), we obtain, after some algebra, the following expressions of the Gini index for a finite population:

$$G = \frac{2}{NY}\sum_{k\in U} N_k y_k - \frac{N+1}{N} = \frac{\sum_{k\in U}\sum_{\ell\in U}|y_k - y_\ell|}{2NY}.$$

Finally, if we use the estimator of the Lorenz curve given in (5.5), we obtain an estimator of the Gini index for weighted observations:

$$\begin{aligned}
\widehat{G} &= \frac{2}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k \widehat{N}_k y_k - \left(1 + \frac{1}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k^2 y_k\right) \\
&= \frac{2\sum_{k\in S} w_k \widehat{N}_k y_k - \sum_{k\in S} w_k^2 y_k}{\widehat{N}\widehat{Y}} - 1 && (5.6)\\
&= \frac{\sum_{k\in S} w_k y_k (2\widehat{N}_k - \widehat{N} - w_k)}{\widehat{N}\widehat{Y}} && (5.7)\\
&= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell |y_k - y_\ell|}{2\widehat{N}\widehat{Y}}. && (5.8)
\end{aligned}$$

Note that $\widehat{L}(\alpha)$ and $\widehat{G}$ are generally biased. If instead of (5.5), the Lorenz curve is estimated through an empirical distribution function (step function), the resulting expression of $\widehat{G}$ may differ from that proposed above, depending upon the type of step function used (see also Davidson, 2009). Our approach avoids this ambiguity.

## 5.3 Variance estimation: A result several times published

### 5.3.1 First results and evolution

The evolution of literature on the variance of the Gini index may be summarized as follows. Until the 1980s, the number of papers devoted to the subject was very limited. This period of time is briefly discussed below. Research was then split into three main directions: survey sampling, robust statistics and economics.

The first calculation of variance relating to the Gini index was proba-

bly realized by Nair (1936) who computed the exact variance of the Gini Mean Difference (GMD), that is the numerator of Expression (5.8) of the Gini index. The proposed expression of variance is nevertheless particularly cumbersome, even in the simplest case of unweighted observations presented by Nair. Note that in a survey context, it is also always possible to compute the variance of a double sum like the GMD, but the computation requires inclusion probabilities up to the fourth order (see for example Ardilly & Tillé, 2006). Lomnicki (1952) and Glasser (1962) have approximated this expression and give simpler variance estimators. The chronology between these three papers is clear: Lomnicki (1952) refers to Nair (1936), while Glasser (1962) cites both articles. Finally, Sillitto (1969) also computed an expression for the variance of the Gini Mean Difference by using L-moments to approximate the quantile function by polynomials (see also Hosking, 1990).

Meanwhile, one of the first results for the variance of the full Gini index is attributed to Hoeffding (1948). After defining the notion of U-statistics based on earlier work by Halmos (1946) and expressing the variance of a U-statistic, Hoeffding shows that the Gini index is a function of two U-statistics and gives a result for the variance of the Gini index. In addition to giving a general form and an unbiased estimator for the variance of the GMD using some results on U-statistics, Glasser (1962) proposes an approximation of the variance of the Gini index for simple random sampling without replacement from a finite population.

Expressions for an asymptotic variance of the Gini index has also been given by Sendler (1979) based on results of Shorack (1972) on functions of order statistics, by Cowell (1989) who generalizes the U-statistic approach to weighted observations, as well as by Schechtman (1991) who combined the work of Hoeffding (1948) and the idea of infinitesimal jackknife (Jaeckel, 1972). In related topics, Gastwirth (1972) computed lower and upper bounds for the estimated Gini index, while Beach & Davidson (1983) proposed inference on the Lorenz curve.

### 5.3.2 Fragmentation of the literature

Because the measure is widely used in practice (in official statistics or policy making for example), The question of sampling variance of the Gini index has interested scholars from partially unrelated domains. When analyzing the whole corpus of literature, a clear separation between publications from the field of statistics and publications in economic journals is witnessed. As papers from one field are seldom cited in the other, it seems evident that researchers from these two fields do not necessarily read each other's work. Moreover, inside statistical literature, some contributions come from survey sampling while others come from robust statistics.

In the past 20 years, the gap between economic and statistical literature on the Gini index seems to have become wider. The survey paper by Xu (2004) is a perfect illustration of this fragmentation. In this paper, which reviews literature on the Gini index across the 20th century, no references are made to survey sampling studies on the topic. Moreover, papers from robust statistics are not mentioned either. One could argue that this paper does not focus on inference *per se*, but in another paper, Xu (2007) proposes an introductory overview of inference for the Gini index, in which the literature from robust statistics and survey sampling of the last two decades is also absent.

The variety of existing methods to tackle the problem is another obstacle to a unified understanding of the latter. Indeed, at least three general categories of approaches have been used: linearization (or asymptotic theory-based) methods, resampling methods and regression-based methods. Moreover, each of these categories encompass many different techniques.

In Section 5.4, a number of linearization techniques are presented. Linearization-based computation of variance of the Gini index has generated a large amount of literature and thus, embodies the main focus of this paper. Furthermore, even after the problem of linearizing the Gini index was solved, the result has been republished several times. The main reason why such a situation happened is probably the plurality of approaches that can be defined as linearization methods: direct computation of variance, use of estimating equations, computation of influence function, Delta method or Demnati & Rao (2004) method. These methods lead to the same result but are done by using different developments. Moreover, while some of these approaches have indeed been proposed directly for variance approximation, some were developed for other purposes such as robustness analysis.

The regression-based methods are described in Section 5.5. A jackknife approach is also discussed inside the regression section as well as in the empirical illustrations. Bootstrap methods are not treated in this paper but have also been applied to propose Gini index standard errors (Mills & Zandvakili, 1997; Kuan, 2000; Giorgi et al., 2006). Finally, note that the problem has recently been addressed using the empirical likelihood method (Qin et al., 2010; Peng, 2011).

### 5.3.3 The pitfall of the computation of the variance

There is certainly another, more methodological reason why variance estimation for the Gini index has resulted in a lot of publications. From today's point of view the problem seems quite simple, thus we show in Section 5.4 a result which gives an expression for the variance in only a few lines of calculation, but the complexity of the statistic has been a chal-

lenge for variance estimation in the past. In particular, some authors who proposed straightforward or simplified solutions were in fact falling into a methodological pitfall. This pitfall can be described as follows: when one examines Expression (5.6) of the estimated Gini index, one could believe that the numerator is composed of two simple sums and that an expression of the variance can be directly derived. Unfortunately, this reasoning is mistaken. Indeed, the quantity $\widehat{N}_k$ defined in (5.3) is an estimator of $N_k$, the rank of unit $k$ in the population, and it is random because its value depends on the selected sample. Thus, part of the variability of the estimated Gini index is due to $\widehat{N}_k$, and this aspect must be taken into account. As shown by Sandström et al. (1985, 1988), the variance of the index is in fact considerably overestimated when this randomness is not taken care of.

One could imagine some situations where the rank is known in the population, for example when the sample is selected from a register. This question has been discussed by Deville (1996). If so, computing variance of the Gini index amounts to expressing the variance of a ratio. The estimator of the Gini index nevertheless has a smaller variance when an estimator of the rank is used rather than the true population rank. Although this overestimation may at first seem surprising, it can be easily illustrated. Suppose an unexpected number of high incomes are selected in the sample. The true population rank $N_k$ of selected unit $k$ will then have a tendency to be underestimated by $\widehat{N}_k$. If on the other hand, very few high incomes are selected, $\widehat{N}_k$ will have a tendency to overestimate the rank of unit $k$. The sampling distribution of $\sum_{k\in S} w_k \widehat{N}_k y_k$ becomes less scattered than that of the respective expression computed with the true population rank, namely $\sum_{k\in S} w_k N_k y_k$, resulting in a smaller variance for the Gini index estimated using the former sum.

## 5.4 Linearization techniques

### 5.4.1 The rationale behind linearization

Linearization combines a range of techniques used to calculate the approximated variance of a non-linear statistic. It consists in approximating a non-linear or complex statistic (here, the Gini index) by a sum of terms, i.e. finding a linearized variable $z_k$ such that

$$\widehat{G} - G \approx \sum_{k\in S} w_k z_k - \sum_{k\in U} z_k.$$

Next, the variance of $\widehat{G}$ is simply approximated by the variance of the estimated total

$$\widehat{Z} = \sum_{k\in S} w_k z_k.$$

Nevertheless, the $z_k$'s often depend on population parameters that must be estimated. By estimating these parameters, one can obtain $\widehat{z}_k$ an estimator of $z_k$ and thus construct an estimator of variance by plugging $\widehat{z}_k$ into the expression of the variance of a total corresponding to the given sampling design. Thus, the method is applicable to any sampling design, provided that an expression for the variance of the total is known. For instance, if the sampling design is simple without replacement, with fixed sample size $n$, we have the estimator

$$\widehat{\text{var}}_{lin}(\widehat{G}) = \frac{N-n}{Nn(n-1)} \sum_{k \in S} (\widehat{z}_k - \bar{\widehat{z}})^2, \tag{5.9}$$

where

$$\bar{\widehat{z}} = \frac{1}{n} \sum_{k \in S} \widehat{z}_k.$$

For details on the asymptotic framework validating linearization, one can relate to Isaki & Fuller (1982), Deville & Särndal (1992) and Deville (1999). Without using this terminology, Glasser (1962) already pointed out that the linearized variable given by

$$u_k = 2 \sum_{\ell \in U} |y_k - y_\ell| \tag{5.10}$$

can be used to approximate the variance of the Gini Mean Difference by plugging it into the estimator of variance of a total in place of the interest variable. For instance, if the sampling design is simple without replacement with fixed sample size $n$, we have the approximation of variance

$$\text{AVar}(\widehat{A}) = N^2 \frac{N-n}{Nn(n-1)} \sum_{k \in U} (u_k - \bar{u})^2,$$

where

$$\widehat{A} = \sum_{k \in S} \sum_{\ell \in S} |y_k - y_\ell|$$

and

$$\bar{u} = \frac{1}{N} \sum_{k \in U} u_k.$$

### 5.4.2 Taylor series expansion

In practice, there are several ways to calculate this linearized variable. For smooth functions of totals, one can linearize by performing a Taylor series expansion with respect to these totals (Woodruff, 1971). In the survey sampling literature, Nygård & Sandström (1981, 1985, 1989) and especially Sandström et al. (1985, 1988) are usually considered as seminal works on the variance of the Gini index. The latter discuss four different variance estimators for the Gini index, first in simple random sampling, then in

unequal probability sampling. Furthermore, Sandström et al. (1988) shows links between their work and those of Glasser (1962) or Sendler (1979). Two of the estimators presented in Sandström et al. (1985, 1988) are based on first order Taylor approximation, and are thus probably the first explicit tentative of linearization of the Gini index. However, the results have been shown to be only partially satisfactory because the Gini index cannot be expressed as a smooth function of totals.

For the first estimator (*ratio* estimator) presented by these authors, the Gini index is simply considered to be a ratio of totals. Thus, the classical first order Taylor approximation of a ratio is used to linearize the index. The authors note that this estimator does not take into account the fact that the ranks depend upon the other units in the sample (namely, that $\widehat{N}_k$ is random) and that these ranks should be estimated in some way. These conclusions are in line with the pitfall discussed in Section 5.3.3. Indeed, the *ratio* estimator is unsatisfactory because it can only be constructed if the Gini index is mistakenly interpreted as a function of totals.

The second estimator (hereafter denoted *Taylor* estimator) proposed by Sandström et al. (1985, 1988) is also based on Taylor series but this time, it takes the randomness of the ranks into account. However, the expression is cumbersome in the simple random sampling case and becomes virtually inapplicable in the probability sampling framework because it requires joint inclusion probabilities up to order four.

In opposition to the first two design-based estimators, the third proposed estimator is what the authors call a *model-based* estimator, denoting that the observations are realizations of *iid* random variables which form a fixed, given, sample. This estimator is related to the U-statistics approach, and, under simple random sampling, reduces to the variance estimator proposed by Sendler (1979). The last estimator discussed is a straightforward jackknife estimator.

In Sandström et al. (1985), the simulation studies under simple random sampling show that the first *ratio* estimator, which is the only estimator to treat the ranks as constants, greatly overestimates the variance. This should act as an indication that handling the random ranks issue is crucial. Other estimators show good results. In the probability sampling case (Sandström et al., 1988), the *ratio* estimator is unsurprisingly also ineffective and the *Taylor* estimator is, as noted previously, not applicable. Thus, in that situation, only the jackknife and the *model-based* estimators are considered satisfactory by the authors. Sandström et al. (1988) also note that the *Taylor* estimator is similar to that of Glasser (1962) and that when both sample and finite population sizes become large, *Taylor* and *model-based* estimators are equal.

### 5.4.3 Influence functions

When the function of interest is, like the Gini index, not a function of totals, a more radical way of computing a linearized variable involves computing the influence function initially proposed by Hampel (1974) and Hampel et al. (1985). The influence function was first proposed to study the robustness of an estimator but can also be used to approximate the variance. The influence function of the Gini index seems to have been computed for the first time by Monti (1991) and next by Cowell & Victoria-Feser (1996b, 2003) but was used to robustify the Gini estimator rather than to estimate the variance. The influence function of the Gini given by Monti (1991) can be rewritten by:

$$z_k = \frac{1}{NY}\left[2N_k(y_k - \overline{Y}_k) + Y - Ny_k - G(Y + y_k N)\right], \tag{5.11}$$

where

$$\overline{Y}_k = \frac{\sum_{\ell \in U} y_\ell \mathbb{1}(N_\ell \leq N_k)}{N_k}.$$

The result obtained by Monti (1991) has been overlooked by survey statisticians, probably because the role of the influence function as a tool for variance estimation was not well established.

We will show in the next sections that this linearized variable can be computed in only a few lines of calculation by means of different methods. This linearized variable can be estimated by:

$$\widehat{z}_k = \frac{1}{\widehat{N}\widehat{Y}}\left[2\widehat{N}_k(y_k - \widehat{\overline{Y}}_k) + \widehat{Y} - \widehat{N}y_k - \widehat{G}(\widehat{Y} + y_k\widehat{N})\right]. \tag{5.12}$$

where

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

### 5.4.4 Estimating equations

Another method used to compute a linearized variable for the Gini index is to express the latter as the solution of an estimating equation (Binder & Kovacevic, 1995; Kovacevic & Binder, 1997). Using the estimating equation methodology (for details on this approach, see for example Binder & Patak, 1994), a linearized variable for the Gini index is derived. The result is equal to that of Expression (5.11) and can be estimated by (5.12).

### 5.4.5 Deville approach

Deville (1996) uses papers by Sandström et al. (1985, 1988) as a starting point for the linearization of the Gini index. He identifies clearly the source of the overestimation obtained by these authors when the randomness of the ranks were neglected. Later, Deville (1999) proposed a mod-

ified version of the influence function in order to compute a linearized variable for sampling from a finite population. Unfortunately, in the paper of Deville (1999), a term is missing in the final expression of the linearized variable of the Gini index.

In order to define the influence function, Deville uses a measure $M$ with unit mass for each point of the population. According to Deville's definition, the measure $M$ is positive, discrete, with a total mass $N$ while the total mass is equal to 1 for the measure used in the influence function proposed by Hampel (1974). A function of interest can be presented as a functional $T(M)$ that associates for each measure a real number or a vector. For instance, a total $Y$ can be written

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

Besides, we also suppose that the considered functionals are homogeneous in the sense that there always exists a real number $\alpha$ such that

$$T(tM) = t^{\alpha} T(M), \text{ for all } t \in \mathbb{R}_+^*.$$

Coefficient $\alpha$ is called the degree of the functional $T(M)$. The measure $M$ is estimated by a measure $\widehat{M}$ that has a mass equal to $w_k$ for each point $x_k$ of sample $S$. The plug-in estimator of a functional $T(M)$ is simply $T(\widehat{M})$. For instance, the estimator of a total is given by

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

Deville's influence function is defined by

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

when this limit exists, where $\delta_x$ is the Dirac measure at point $x$. This influence function is the Gâteaux differential in the direction of the Dirac mass at point $x$. Deville (1999) shows that this influence function $z_k = IT(M, x_k)$ is a linearized variable of $T(\widehat{M})$ in the sense that it allows for the approximation of the interest function:

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

The approximation of the variance of $T(\widehat{M})$ is obtained by computing the variance of the weighted sum of $z_k$ on the sample:

$$\text{Avar}[T(\widehat{M})] = \text{var} \left( \sum_{k \in S} w_k z_k \right).$$

The influence function generally depends on population parameters that are unknown and can simply be estimated by replacing the latter by their plug-in estimators. Thereby, we obtain $\widehat{z}_k$, the estimator of the linearized variable $z_k$. Computation of influence functions follows the rules of differential calculus. Deville (1999) has shown amongst other properties, the following results.

**Result 5.1** *If $T(M) = \sum_{k\in U} y_k = \int y dM(y)$, then the influence function is $IT(M, x_k) = y_k$.*

**Result 5.2** *If $S$ and $T$ are two functionals, $I(S+T) = IS + IT$ and $I(ST) = SIT + TIS$.*

**Result 5.3** *Let $\mathbf{S}$ a functional of $\mathbb{R}^q$, $\mathbf{T}_{\boldsymbol{\lambda}}$ a family of functional that depend of $\boldsymbol{\lambda} \in \mathbb{R}^q$, then*

$$I(\mathbf{T}_{\mathbf{S}}) = I(\mathbf{T}_{\boldsymbol{\lambda}=\mathbf{S}}) + \left.\frac{\partial \mathbf{T}}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\mathbf{S}} I\mathbf{S}.$$

Result 5.3 shows that the computation of the influence function can also be realized by steps. We propose hereafter an additional result that enables us to directly compute the linearized variable of a double sum (e.g. quadratic form) such that

$$S = \sum_{k\in U}\sum_{\ell\in U} \phi(y_k, y_\ell).$$

**Result 5.4** *If*

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

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

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

*Proof.* Let

$$\begin{aligned}
C(t) &= \frac{1}{t}\left[S(M + t\delta_\xi) - S(M)\right] \\
&= \frac{1}{t}\left[\int\int \phi(x,y) d(M + t\delta_\xi)(x) d(M + t\delta_\xi)(y) - \int\int \phi(x,y) dM(x) dM(y)\right] \\
&= \frac{1}{t}\int\left[\int \phi(x,y) d(M + t\delta_\xi)(x) - \int \phi(x,y) dM(x)\right] dM(y) \\
&\quad + \frac{1}{t}\int\int \phi(x,y) d(M + t\delta_\xi)(x) d(t\delta_\xi)(y) \\
&= \frac{1}{t}\left[\int\int \phi(x,y) dM(y) d(M + t\delta_\xi)(x) - \int\int \phi(x,y) dM(y) dM(x)\right] \\
&\quad + \int \phi(x,\xi) d(M + t\delta_\xi)(x).
\end{aligned}$$

We obtain

$$IS(M,\xi) = \lim_{t\to 0} C(t) = \int \phi(\xi,y)dM(y) + \int \phi(x,\xi)dM(x).$$

□

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

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

Result 5.4 allows for a fast computation of the linearized variable of the Gini index. The latter can be written

$$G = \frac{A}{2NY},$$

where $A$ is a double sum

$$A = \sum_{k\in U}\sum_{\ell\in U} |y_k - y_\ell|.$$

By using Result 5.4, we get a linearized variable of $A$ immediately:

$$IA(M,x_k) = 2\sum_{\ell\in U} |y_k - y_\ell| = 2\left[2N_k(y_k - \overline{Y}_k) + Y - Ny_k\right],$$

which is the result presented in Expression (5.10) obtained by Glasser (1962) by using a different method. Now, if we apply the technique of linearization by steps as well as Result 5.1 for totals $Y$ and $N$, we obtain a linearized variable of the Gini index

$$\begin{aligned} z_k &= IG(M,x_k) = \frac{IA(M,x_k)}{2NY} - \frac{A}{2N^2Y}IN(M,x_k) - \frac{A}{2NY^2}IY(M,x_k) \\ &= \frac{1}{NY}\left[\frac{IA(M,x_k)}{2} - G(Y + y_kN)\right] \\ &= \frac{1}{NY}\left[2N_k(y_k - \overline{Y}_k) + Y - Ny_k - G(Y + y_kN)\right], \end{aligned}$$

which is the result proposed in Monti (1991) and can be estimated by (5.12).

### 5.4.6 Demnati and Rao approach

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

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

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

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

The computation of a simple derivative with respect to the weights is advocated by Demnati & Rao (2004) in order to compute the linearized variable of a function of totals. This method also enables us to compute a linearized variable for any function of interest whose observations are weighted by $w_k$. By computing the derivative of Expression (5.6) with respect to $w_k$, we obtain, in a couple of lines, the estimator of the linearized variable given in Expression (5.12). The result is exactly the same as the one obtained by Deville's method.

### 5.4.7 Graf approach

In a recent paper, Graf (2011) proposes another way of computing the linearized variable by applying a Taylor expansion with respect to the indicator variables $I_k$, where for all $k \in U$

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

determines the presence of unit $k$ in the sample. The Graf method is coherent because the expansion is done with respect to the only source of randomness in the estimator. In sampling from a finite population, the estimator of the Gini index $\widehat{G}$ can be written as a functions of these indicator variables: $\widehat{G} = Q(I_1, \ldots, I_k, \ldots, I_N)$. By using a Taylor expansion, we can then write:

$$Q(I_1, \ldots, I_k, \ldots, I_N) \approx Q(\pi_1, \ldots, \pi_k, \ldots, \pi_N) + \sum_{k \in U} (I_k - \pi_k) Q'_k, \quad (5.13)$$

where $Q'_k(.)$ is the partial derivative of $Q(I_1, \ldots, I_k, \ldots, I_N)$ with respect to $I_k$. Since

$$Q(\pi_k, \ldots, \pi_k, \ldots, \pi_N) = G,$$

we can then compute a linearized variable as

$$z_k = I_k Q'_k.$$

From Expression (5.13), we obtain

$$\widehat{G} - G \approx \sum_{k \in S} \frac{z_k}{\pi_k} - \sum_{k \in U} z_k.$$

If we use the Horvitz-Thompson weights $d_k = 1/\pi_k$, Expression (5.6) can be written:

$$\widehat{G} = \frac{\sum_{k\in U}\sum_{\ell\in U} d_k d_\ell |y_k - y_\ell| I_k I_\ell}{2\sum_{k\in U} d_k I_k \sum_{k\in U} d_k I_k y_k}.$$

The computation of $I_k Q_k'$ directly gives Expression (5.12). Note that in all the proposed linearization methods, if the weights are not fixed but depend upon the $I_k$'s (for example if they result from a calibration procedure) they must be differentiated as well. The advantage of the Graf approach however, is that the effect of the calibration procedure on the estimator is directly accounted for when the weights are differentiated with respect to $I_k$.

### 5.4.8 Other recent publications

In survey sampling, Dell et al. (2002) and Osier (2006, 2009) also show results based on the influence function in the sense of Deville. Both papers derive a linearized variable for various inequality and poverty measures including the Gini index as well as application to survey data. Although the whole technique and computation rules are attributed to Deville (1999) in both articles, the result for the Gini index is not presented as a problem already solved.

In the work of Cowell & Victoria-Feser (2003), the influence function result by Monti (1991) is presented and reference to Deville (1999) is made to show that influence functions can also be used for variance estimation, not only to study robustness. Finally, the computation of the influence function and the approximation of the variance of the Gini index by linearization are presented as well-known results in the works of Berger (2008) and Lesage (2009), the former referring to Kovacevic & Binder (1997), the latter citing Deville (1999).

Outside of survey statistics, the same result using different methodologies related to influence functions and linearization has also been republished at least three times in recent years (Bhattacharya, 2007; Davidson, 2009; Barrett & Donald, 2009). The common point of these articles is an obvious lack of references to prior papers from robust and survey statistics, nor to seminal papers like Hoeffding (1948) or Glasser (1962). The fact is that these papers all propose a short "historical" introduction on the state of the art of the problem, in which the whole literature presented in Sections 5.3 and 5.4 is absent. These papers are briefly discussed below.

Bhattacharya (2007) has proposed asymptotic inference for the Gini index using an asymptotic framework based on a generalized method of moments and empirical process theory. An influence function for the Gini index is derived and the method is extended to clustered and stratified sampling. Although the outcome is similar to previously published papers, the development and theoretical basis is different. The paper makes

no reference to previous research studies on the influence function of the Gini index. Instead Cowell (1989), Zheng (2001) or Bishop et al. (1997) and an earlier draft of Barrett & Donald (2009) are cited. Moreover, the author notes that existing literature on inference for inequality measures has almost always assumed simple random sampling, while we have shown that many papers in the survey sampling literature have tackled the issue of complex sampling designs.

In a paper by Davidson (2009), the delta method is used to produce an asymptotically valid expression for the variance of the Gini index. The paper starts with a brief history of the subject. Reference is made to recent papers (Bishop et al., 1997; Xu, 2007) on the U-statistic approach instead of older papers like Hoeffding (1948), as well as references from the regression based approach. From survey sampling literature, only Sandström et al. (1988) is cited. The proposed variance estimator for the Gini index is constructed via the delta method and the result obtained is equivalent to earlier suggestions in Monti (1991), Kovacevic & Binder (1997), Deville (1999) or Cowell & Victoria-Feser (1996b). The numerical illustration applied in this paper is of interest and is discussed in Section 5.6.

Barrett & Donald (2009) work upon a similar corpus of literature as Davidson (2009). Their work is said to take grounds in results by Cowell (1989), Bishop et al. (1997), Xu (2007), Zitikis & Gastwirth (2002), Zitikis (2003). The authors use influence functions to derive asymptotic variance expressions for generalized Gini indices (the E-Gini and S-Gini). Cowell & Victoria-Feser (1996b) is referred to as one of the first application of the influence function in econometrics. The main contribution of the paper is that the influence function is derived for a whole class of Gini indices and not only for the classical Gini index. However, references to previous literature on the influence function of the latter and its use for variance estimation is again nearly completely omitted.

## 5.5 Regression-based variance estimation

### 5.5.1 Ogwang's Jackknife

A different way of expressing the Gini index has prompted a new wave of publications. In fact, it has been advocated already since the 1980s that the Gini index can be expressed by means of a covariance (Anand, 1983; Lerman & Yitzhaki, 1984; Shalit, 1985). With $F(y)$ denoting the cumulative distribution of incomes $y$, and $\mu(y)$ the mean income, the Gini index can thus be written as follows:

$$G = \frac{2\text{cov}\left[y, F(y)\right]}{\mu(y)}. \tag{5.14}$$

This idea has been exploited by Ogwang (2000) in order to derive a fast algorithm for the computation of jackknife estimates of the variance of the Gini index. Ogwang (2000) shows that using (5.14) and sorting incomes in non-decreasing order, the Gini index estimated by means of a sample of size $n$ is a function of a regression coefficient such that

$$\widehat{G} = \frac{n^2-1}{6n}\frac{\widehat{\beta}}{\bar{y}},$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, $\widehat{\beta}$ is the ordinary least square (OLS) estimator of $\beta$ from the regression model $y_i = \alpha + \beta i + u_i$, for $i = 1, ..., n$ and error terms $u_i$ are assumed to be homoscedastic with mean 0 and variance $\sigma^2$. Equivalently, the Gini index can be estimated by

$$\widehat{G} = \frac{2}{n}\widehat{\theta} - 1 - \frac{1}{n}, \tag{5.15}$$

where $\widehat{\theta} = \sum_{i=1}^{n} i y_i / \sum_{i=1}^{n} y_i$ is the weighted least square (WLS) estimator of $\theta$ in the regression model $i = \theta + u_i$ with heteroscedastic error $u_i$ of variance $\sigma^2/y_i$. Note that (5.15) is equivalent to (5.2). The algorithm proposed by Ogwang is fast because it avoids re-ranking incomes at each step, i.e. every time an observation is dropped from the sample. The Ogwang-jackknife variance estimator is

$$\widehat{\text{var}}_{Oj}(\widehat{G}) = \frac{n-1}{n}\sum_{k=1}^{n}\left(\widehat{G^{(k)}} - \frac{1}{n}\sum_{\ell=1}^{n}\widehat{G^{(\ell)}}\right)^2, \tag{5.16}$$

where $\widehat{G^{(k)}}$ is the estimated Gini index of the remaining $(n-1)$ observations after deletion of unit $k$:

$$\begin{aligned}\widehat{G^{(k)}} &= \widehat{G} + \frac{2}{\sum_{i=1}^{n} y_i - y_k}\left[\frac{y_k\widehat{\theta}}{n} + \frac{\sum_{i=1}^{n} i y_i}{n(n-1)} - \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{k} y_i + k y_k}{n-1}\right] \\ &\quad - \frac{1}{n(n-1)},\end{aligned} \tag{5.17}$$

and can be computed as an $n$-component vector in only one pass through the data. It is crucial to note that $\widehat{G^{(k)}}$ takes into account the fact that the rank of $y_i$ changes when observation $k$ is dropped from the sample and $y_i \geq y_k$. Computing $\widehat{G^{(k)}}$ is thus equivalent to applying (5.15) successively to each sample of size $(n-1)$. The standard error $\text{SE}_{Oj}(\widehat{G})$ is obtained simply by

$$\text{SE}_{Oj}(\widehat{G}) = \sqrt{\widehat{\text{var}}_{Oj}(\widehat{G})}. \tag{5.18}$$

Note that a jackknife estimator had already been applied with satisfactory results by Sandström et al. (1985, 1988) in their simulation study, as well as by other authors (for a review see Berger, 2008). In their approach, the

Gini index is recomputed on the $(n-1)$ remaining observations for each jackknife sub-sample using Expression (5.7). The result is no different than that of Ogwang (2000), but the contribution of the latter is that by using (5.17), computation becomes far less intensive because all $\widehat{G^{(k)}}$'s can be computed at once while still taking the rank changes into account. This is an important feature when the sample size becomes large. A similar simplification is also proposed by Karagiannis & Kovacevic (2000).

### 5.5.2 Direct regression

In a later paper, Giles (2004) suggests a direct analytical variance estimator for the Gini index based on regression theory. Indeed, as Expression (5.15) shows, $\widehat{G}$ is a function of a regression coefficient and an estimator of the variance of the Gini index can thus be derived directly from the variance of the regression coefficient by

$$\widehat{\text{var}}_{reg}(\widehat{G}) = \frac{4}{n^2}\widehat{\text{var}}(\widehat{\theta}), \tag{5.19}$$

and, accordingly, the standard error by

$$\text{SE}_{reg}(\widehat{G}) = \frac{2}{n}\text{SE}(\widehat{\theta}),$$

where $\text{SE}(\widehat{\theta}) = \sqrt{\widehat{\text{var}}(\widehat{\theta})}$ can be obtained from any basic statistical package handling OLS or WLS. Thereby, Giles advocates that using the jackknife method to derive a variance estimator is unnecessary. We argue that this idea is a deep methodological error and that the variance estimator in (5.19) must not be used. Giles (2004)'s paper has been followed by a discussion concerning the violation of the model assumptions leading to a substantial overestimation of the variance (Ogwang, 2004, 2006; Giles, 2006; Modarres & Gastwirth, 2006). Above all, Modarres & Gastwirth (2006) argue that because the $y_i$'s are ordered, they are dependent, and therefore the independence assumption for the error terms in the regression model does not hold. This dependency is totally ignored in Giles's proposal. In a classical regression, the independent variable is supposed to be non-random, which absolutely does not correspond to the problem of the variance of the Gini index. The solution of Giles (2004) is thus wrong since he committed the same error as in the first *ratio* estimator of Sandström et al. (1985, 1988). In other words, he fell into the pitfall described in Section 5.3.3. The discussion that follows the paper (see Ogwang, 2006; Modarres & Gastwirth, 2006; Giles, 2006) is a little bit vain since this problem was already identified by Deville (1996, 1999). Indeed, as discussed earlier in this paper, the major issue in leading reliable inference for the Gini index is that the population rank $N_k$ of unit $k$ has to be estimated by means of the sample. Also, the proposed model does not correspond to

the way the data has been produced or could be modelled, because it assumes that the independent variable is not random. Moreover, as pointed out by Berger (2008), the regression approach makes no account of the sampling design.

## 5.6 Empirical comparisons

Giles (2004) proposes an empirical illustration of the regression-based method using data available from the Penn World Table (Heston et al., 1995) for 133 countries. The variable of interest is, for each country, the real consumption per capita in constant USD (international prices based on year 1985). The Gini index and its standard error are computed for four different time periods (1970, 1975, 1980, 1985). In this numerical application, Giles (2004) compares $\text{SE}_{reg}(\widehat{G})$ to a jackknife variance estimator which the author incorrectly attributes to Ogwang (2000) and shows that both variance estimators behave similarly. However, we will show hereafter that the jackknife estimator computed by Giles is by no means equal to $\text{SE}_{Oj}(\widehat{G})$ of Expression (5.18). Recently, Davidson (2009) used the same data set to compare a linearization variance estimator based on the Delta method to both Giles' and Ogwang's approaches. Note that the linearization estimator proposed in Davidson (2009) is equivalent to Expression (5.9), the only difference being the way the cumulative distribution function is estimated in $\hat{z}_k$. Because we have witnessed only a negligible effect in the empirical applications proposed later in this paper, we simply consider both of them under the term linearization variance estimator and denoted $\widehat{\text{var}}_{lin}(\widehat{G})$. We also define

$$\text{SE}_{lin}(\widehat{G}) = \sqrt{\widehat{\text{var}}_{lin}(\widehat{G})},$$

because standard errors, rather than variances, are reported in both Giles's and Davidson's papers.

The results of these empirical illustrations are confusing and the conclusions unclear. Indeed, Giles (2004), who in addition also conducted a simulation, argues that the jackknife approach produces biased estimates in samples smaller than $5,000$ observations, where the regression approach is more reliable, and that both methods tend to converge as sample size increases. Thus, the use of the regression method is advocated by Giles because it is computationally much simpler and allows for some robustness testing.

In his paper, Davidson (2009) argues on the other hand that both methods might well be unreliable. When reproducing the empirical application of Giles, he obtains the same numerical results as the latter and both approaches are shown to yield much larger variance estimators than the

linearization method. When analyzing the results however, the author does not clearly discard the jackknife and regression estimators. No clear comment is made on the quality of the three estimators and on the discrepancies between them.

Hereafter, we show that the regression approach is conceptually wrong and that, when used correctly as suggested by Ogwang (2000), the jackknife provides reliable estimates in addition to being computationally straightforward. Both Giles (2004) and Davidson (2009) use the same method for computing the jackknife variance estimator and obviously, the estimator is not the one proposed in (5.16). The jackknife approach used in these two papers is hereafter denoted the Davidson-Giles jackknife (DGj) and the resulting variance estimator is :

$$\widehat{\text{var}}_{DGj}(\widehat{G}) = \frac{n-1}{n} \sum_{k=1}^{n} \left( \widehat{\widetilde{G^{(k)}}} - \frac{1}{n} \sum_{\ell=1}^{n} \widehat{\widetilde{G^{(k)}}} \right)^2, \tag{5.20}$$

where $\widehat{\widetilde{G^{(k)}}}$ is the Gini index computed using Expression (5.15) directly on the $(n-1)$ values without taking into account the effect of dropping unit $k$ on the ranks of the other sampled units (i.e. without recomputing $\hat{\theta}$ each time a unit is dropped). Likewise, the respective standard error is

$$\text{SE}_{DGj}(\widehat{G}) = \sqrt{\widehat{\text{var}}_{DGj}(\widehat{G})}.$$

Within the procedure of Davidson and Giles, the rank of observation $y_i$ is kept constant among all the $n$ jackknife sub-samples, while in fact as discussed in Section 5.3.3, the ranks depend upon the sample and should thus be recalculated within each jackknife sub-sample. This issue is taken into account in (5.16) but not in (5.20). In our empirical study below, we show that (5.20) leads to a heavy overestimation of the variance, as was the case with the direct regression approach. Indeed, using this version of the jackknife is nothing other than falling again into the pitfall described in Section 5.3.3. In addition to the fact that the jackknife estimator is computed incorrectly, these empirical illustrations have another major issue: because the setup is not that of a repeated random sampling simulation method, the obtained estimates cannot be compared to a Monte Carlo variance which would mimic the true value. Thus, it is not possible to assess the genuine reliability of the proposed variance estimators.

In order to take these issues into account, we first repeat the exact empirical setup proposed by Giles (2004) and Davidson (2009) and add a fourth estimator, the Ogwang jackknife estimator $\text{SE}_{Oj}(\widehat{G})$. Subsequently, we use the same data to perform a Monte Carlo simulation study using two different sample sizes. In both numerical illustrations, only data for the year 1970 is used.

Table 5.1 summarizes the results of the first replication. The figures of

Table 5.1 – *Standard errors of $\widehat{G}$ using different approaches (Penn World Table data, year 1970) and comparing results obtained in Giles (2004), Davidson (2009) and in our replication.*

| | Regression approach $\text{SE}_{reg}(\widehat{G})$ | Davidson-Giles jackknife $\text{SE}_{DGj}(\widehat{G})$ | Linearization approach $\text{SE}_{lin}(\widehat{G})$ | Ogwang jackknife $\text{SE}_{Oj}(\widehat{G})$ |
|---|---|---|---|---|
| Giles (2004) | 0.0417 | 0.0481 | | |
| Davidson (2009) | 0.0418 | 0.0478 | 0.0173 | |
| Our replication | 0.0418 | 0.0478 | 0.0173 | 0.0176 |

the previous studies are reported for comparisons. Note that the results we produced are exactly the same as in Davidson (2009) and that they differ very slightly from those of Giles (2004), probably because data was processed differently in order to create the desired real consumption per capita variable. These results show that the Ogwang jackknife algorithm gives a variance estimator that is very close to the linearization estimator and very different from the other two estimators, showing that taking the changes of the ranks into account in the jackknife procedure has a massive effect on variance estimation.

We then proceed to a Monte Carlo simulation setup on the same data. Ten thousand simple random samples of size $n = 30$ are drawn without replacement from the full data of $N = 133$ countries. On each sample, the Gini index $\widehat{G}$ as well as the four competing variance estimators are computed. The Monte-Carlo standard error of the Gini index, denoted $\text{SE}_{sim}(\widehat{G})$, is then computed on the $10,000$ obtained estimators of $G$. The value of $\text{SE}_{sim}(\widehat{G})$ is the benchmark to which the four estimators can be compared. The exact same simulation setup was also conducted with a larger sample size of $n = 100$. In Table 5.2, the Monte Carlo expected value $\text{E}_{sim}[\text{SE}(\widehat{G})]$ for each method is reported. The right-hand column contains the Monte Carlo which approximates the true value for each sample size. Results show that the linearization technique as well as the jackknife provide reliable estimates for the variance of the Gini index. Indeed, both methods give estimators that are close to the Monte Carlo estimator. Unlike what has been reported in previous papers, the jackknife procedure proposed by Ogwang (2000) is valid if applied correctly. On the contrary, the two methods that do not account for the rank issue yield unsatisfactory results. As suggested earlier in the paper, variance is greatly overestimated when these approaches are used. In our simulations for example, the variance estimated by the regression approach is overestimated by a factor of more than 20 with a sample size of $n = 100$. We thus point out that the main problem with the regression-based variance estimation is

Table 5.2 – *Simulation results: standard errors of $\widehat{G}$ using different approaches (Penn World Table data, year 1970).* 10,000 *replicates, simple random sampling without replacement.*

| | | $n = 30$ | $n = 100$ |
|---|---|---|---|
| Regression approach | $\mathrm{E}_{sim}[\mathrm{SE}_{reg}(\widehat{G})]$ | 0.0896 | 0.0483 |
| Davidson-Giles jackknife | $\mathrm{E}_{sim}[\mathrm{SE}_{DGj}(\widehat{G})]$ | 0.1066 | 0.0554 |
| Linearization approach | $\mathrm{E}_{sim}[\mathrm{SE}_{lin}(\widehat{G})]$ | 0.0333 | 0.0100 |
| Ogwang jackknife | $\mathrm{E}_{sim}[\mathrm{SE}_{Oj}(\widehat{G})]$ | 0.0355 | 0.0102 |
| Monte Carlo estimator | $\mathrm{SE}_{sim}(\widehat{G})$ | 0.0337 | 0.0101 |

that the rank is not identified as a major source of randomness, leading to over-estimation. The method should therefore be avoided.

## 5.7 Conclusion

As this paper tries to illustrate, the computation of variance of the Gini index has been subject to a vast number of publications. For numerous reasons, the same results have been republished several times. The segmentation of the different research fields and the multiplicity of methods that can be applied to obtain an approximation of variance are probably the main causes of this phenomenon. Indeed, an examination of the referenced papers in the publications clearly shows the lack of communication between statisticians, economists and survey statisticians. As a consequence, the same "mistakes" have been reproduced when tackling the problem, an example being the recent regression-based variance estimator suggested by Giles (2004).

In addition to giving a small contribution (the linearization of a double sum) which tends to simplify the problem, we try to propose a global picture of the state of the art regarding inference for the Gini index. We hope that the paper clarifies the situation to a much larger extent than previous survey papers on the Gini index have. We also hope that it will provide researchers from different fields with an update on what is done elsewhere as well as a comprehensive survey of the literature. Finally, we hope it will enlighten the reader on the interesting issues regarding inference on a non-linear statistic and on the different possible approaches leading to the result.

## Acknowledgements

The authors would like to thank the Joint Editor and three anonymous referees for their insightful comments and suggestions. This research is supported by grant no. 200021-121604 of the Swiss National Science Foundation.

# 6

# Inference by linearization for Zenga's new inequality index: A comparison with the Gini index

**Abstract**

Zenga's new inequality curve and index are two recent tools for measuring inequality. Proposed in 2007, they should thus not be mistaken for anterior measures suggested by the same author. This paper focuses on the new measures only, which are hereafter referred to simply as the Zenga curve and Zenga index. The Zenga curve $Z(\alpha)$ involves the ratio of the mean income of the $100\alpha\%$ poorest to that of the $100(1-\alpha)\%$ richest. The Zenga index can also be expressed by means of the Lorenz Curve and some of its properties make it an interesting alternative to the Gini index. Like most other inequality measures, inference on the Zenga index is not straightforward. Some research on its properties and on estimation has already been conducted but inference in the sampling framework is still needed. In this paper, we propose an estimator and variance estimator for the Zenga index when estimated from a complex sampling design. The proposed variance estimator is based on linearization techniques and more specifically on the direct approach presented by Demnati and Rao. The quality of the resulting estimators are evaluated in Monte Carlo simulation studies on real sets of income data. Finally, the advantages of the Zenga index relative to the Gini index are discussed. [1]

**Keywords:** Gini, inequality, influence function, sampling, variance estimation

[1]This chapter is a reprint of: Langel, M. and Tillé, Y. (2011). Inference by linearization for Zenga's new inequality index: A comparison with the Gini index. *Metrika*, doi:10.1007/s00184-011-0369-1, 1-18.
(The original publication is available at www.springerlink.com)

## 6.1 Introduction

Research on inequality measures can be conducted in different directions. One direction consists in improving methodology on broadly used statistics such as the Gini index or entropy measures, while another direction focuses on proposing new inequality measures and places emphasis on the corresponding advantages. It is a fact that the level of income inequality in a population is often accounted for by using the Gini index. Many discussions concerning the latter measure have arisen in statistical and economic literature and a lot of competing inequality measures have been proposed. Some, like the Atkinson index, the Theil index or the Quintile Share Ratio, have been known and used for decades. This paper focuses on finite population inference for a very recent measure, Zenga's new inequality index (Zenga, 2007) which is seen as a potential alternative to the Gini index. This new inequality index should not be mistaken for anterior measures proposed some years ago by the same author (Zenga, 1984) which are also often referred to as Zenga indices in the literature. For the sake of simplicity, Zenga's new inequality curve and index are hereafter denoted by the terms Zenga curve and Zenga index and respectively expressed by $Z(\alpha)$ and $Z$.

Like the Gini index, the Zenga index can be expressed by means of the Lorenz curve. However, it is also associated with a new inequality curve, the Zenga curve which provides interesting and direct interpretations on inequality. The paper is structured as follows. In the next section, the index and the curve are defined and an estimator allowing for complex sampling designs is derived. Section 6.3 is dedicated to variance estimation. Linearization techniques are used to propose a variance estimator for the Zenga index. Some simulations on real data sets are then run in Section 6.4 to apply our theoretical results, while Section 6.5 focuses on comparisons with the Gini index and on the advantages of the Zenga index. The paper ends with some concluding remarks.

## 6.2 The Zenga index and Zenga curve

### 6.2.1 Definition and notation

Some inequality indices are synthetic values based on an underlying curve or function. The most obvious example is the Gini index and the underlying Lorenz curve. Although the Gini index is the main inequality measure, it does not have unanimous support from statisticians and practitioners and thus has prompted research on other curves and synthetic indices. Zenga (1984) had already proposed two curves as well as the inequality measures $\lambda$ and $\xi$. The $\xi$ index and its underlying curve have drawn particular attention (for a review see Zenga, 2007), but according

to the author, it has not been widely used because it requires estimation of the quantile function as well as of the inverse of the incomplete first moment.

Zenga (2007) has presented a new alternative to the Gini index and other existing inequality measures and curves. Although literature on the Zenga index is not as plentiful as on the Gini index, it has drawn increasing attention in the scientific community. The literature includes some publications on the properties of the index (Maffenini & Polisicchio, 2010), inference and applications (Polisicchio, 2008; Greselin et al., 2009, 2010) as well as subgroup decomposition (Radaelli, 2008, 2010). The literature on the Zenga index and curve also focuses greatly on its advantages compared to the Gini index. Some of these features are described below.

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

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

where $0 \leq \alpha \leq 1$ and $L(\alpha)$ is the quantile share or Lorenz curve (Lorenz, 1905; Gastwirth, 1972; Cowell, 1977; Kovacevic & Binder, 1997; Langel & Tillé, 2011d), which is a central tool of inequality theory and is defined by

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

The Zenga index, which can be written

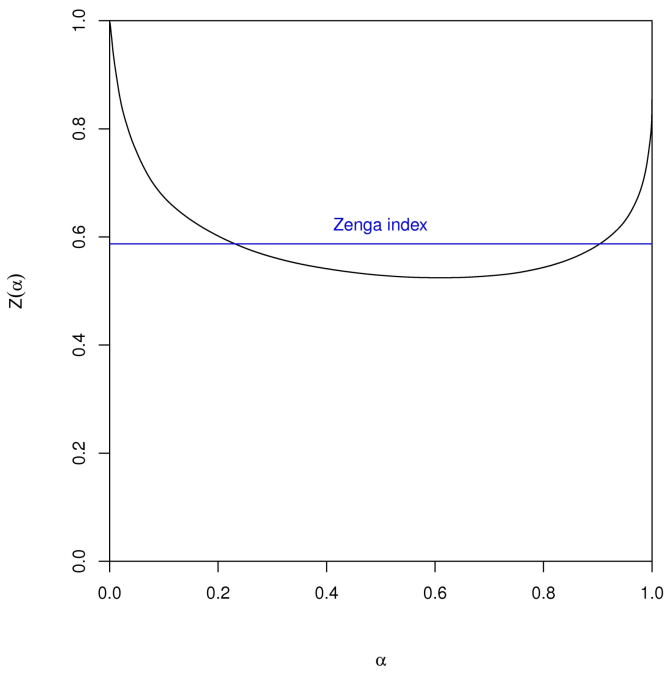
$$Z = \int_0^1 Z(\alpha) d\alpha,$$

can thus be defined, like the Gini index, in terms of the Lorenz curve. Figure 6.1 shows how the index can be plotted together with the Zenga curve and interpreted as a mean level of inequality.

### 6.2.2 The Zenga index in finite population

Let $U$ denote a finite population of $N$ identifiable units $u_1,...,u_k,...,u_N$. For the sake of simplicity, we will hereafter denote unit $u_k$ by its identifier $k$. Associated with each unit $k$ is the value $y_k$ of some characteristic of interest, for example income. To lighten the notation, we will assume with

Figure 6.1 – *Example of a Zenga curve and index for a synthetic data set generated from real Austrian EU-SILC survey data (see Section 6.4.1 for more details).*




no loss of generality that all $y_k$'s are distinct and sorted. Let us define

$$Y = \sum_{\ell \in U} y_\ell, \tag{6.1}$$

$$Y_k = \sum_{\ell \in U} y_\ell \mathbb{1}(\ell \leq k), \tag{6.2}$$

where $\mathbb{1}(A) = 1$ if $A$ is true and 0 otherwise. As suggested in Langel & Tillé (2011d), let us also denote partial sum $Y(\alpha)$, the sum of incomes up to quantile $\alpha$ by

$$Y(\alpha) = Y_{k-1} + y_k[\alpha N - (k-1)], \tag{6.3}$$

where the value of $k$ is such that $\alpha N < k \leq \alpha N + 1$. With Expression (6.3), the finite population quantile share can be defined by

$$L(\alpha) = \frac{Y(\alpha)}{Y}.$$

The Zenga index for a population of size $N$ is then

$$Z = \sum_{k \in U} Z_k, \tag{6.4}$$

where

$$\begin{aligned} Z_k &= \int_{(k-1)/N}^{k/N} 1 - \frac{Y(\alpha)}{\alpha} \cdot \frac{1-\alpha}{Y - Y(\alpha)} d\alpha \\ &= \frac{1}{N} - \int_{(k-1)/N}^{k/N} \frac{Y(\alpha)}{\alpha} \cdot \frac{1-\alpha}{Y - Y(\alpha)} d\alpha \\ &= \frac{(k-1)y_k - Y_{k-1}}{Y + (k-1)y_k - Y_{k-1}} \log\left(\frac{k}{k-1}\right) \\ &+ \left[\frac{Y}{Ny_k} - \frac{Y}{Y + (k-1)y_k - Y_{k-1}}\right] \log\left(\frac{Y - Y_{k-1}}{Y - Y_k}\right). \end{aligned}$$

We now assume $Y_0 = 0$ and define $A_k = (k-1)y_k - Y_{k-1}$ for all $k \in U$. The above can thus be rewritten

$$Z_k = \begin{cases} \left(\frac{Y}{Ny_1} - 1\right) \log\left(\frac{Y}{Y - Y_1}\right), & \text{if } k = 1, \\ \frac{A_k}{Y + A_k} \log\left(\frac{k}{k-1}\right) & \\ \quad + \left[\frac{Y}{Ny_k} - \frac{Y}{Y + A_k}\right] \log\left(\frac{Y - Y_{k-1}}{Y - Y_k}\right), & \text{if } k = 2, ..., N-1, \\ \left(1 - \frac{Y}{Ny_N}\right) \log\left(\frac{N}{N-1}\right), & \text{if } k = N. \end{cases} \tag{6.5}$$

### 6.2.3 An estimator of the Zenga index

A random sample $S$ of size $n$ is drawn from a finite population $U$ of size $N$ from a random sampling design, such that $p(s) = \Pr(S = s)$ is the probability of selecting sample $s \subset U$. The probability for unit $k \in U$ to be included in the sample is written $\pi_k = \Pr(k \in S)$ and $d_k$ denotes the design weight of $k$ such that $d_k = 1/\pi_k$. Note that the design weights $d_k$ are used here for simplicity, and that the following is still valid if the set of weights result from a calibration procedure. Let us also denote

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

$$D_k = \sum_{\ell \in S} d_\ell \mathbb{1}(\ell \leq k)$$

and

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

Expressions (6.1), (6.2) and (6.3) can be respectively estimated from a sample by

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

$$\widehat{Y}_k = \sum_{\ell \in S} d_\ell y_\ell \mathbb{1}(\ell \leq k),$$

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

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

$$\widehat{L}(\alpha) = \frac{\widehat{Y}(\alpha)}{\widehat{Y}}.$$

A natural estimator for the Zenga index is then:

$$\widehat{Z} = \sum_{k \in S} \widehat{Z}_k, \tag{6.6}$$

where

$$\begin{aligned}
\widehat{Z}_k &= \int_{\alpha_{k-1}}^{\alpha_k} 1 - \frac{\widehat{Y}(\alpha)}{\alpha} \cdot \frac{1-\alpha}{\widehat{Y} - \widehat{Y}(\alpha)} d\alpha \\
&= \frac{d_k}{D} - \int_{\alpha_{k-1}}^{\alpha_k} \frac{\widehat{Y}(\alpha)}{\alpha} \cdot \frac{1-\alpha}{\widehat{Y} - \widehat{Y}(\alpha)} d\alpha \\
&= \frac{D_{k-1} y_k - \widehat{Y}_{k-1}}{\widehat{Y} + D_{k-1} y_k - \widehat{Y}_{k-1}} \log\left(\frac{D_k}{D_{k-1}}\right) \\
&+ \left[\frac{\widehat{Y}}{D y_k} - \frac{\widehat{Y}}{\widehat{Y} + D_{k-1} y_k - \widehat{Y}_{k-1}}\right] \log\left(\frac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right).
\end{aligned}$$

Assuming $\widehat{Y}_0 = 0$, $D_0 = 0$ and defining $\widehat{A}_k = D_{k-1} y_k - \widehat{Y}_{k-1}$ for all $k \in S$, we have:

$$\widehat{Z}_k = \begin{cases}
\left(\dfrac{\widehat{Y}}{D y_1} - 1\right) \log\left(\dfrac{\widehat{Y}}{\widehat{Y} - \widehat{Y}_1}\right), & \text{if } k = 1, \\
\dfrac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k} \log\left(\dfrac{D_k}{D_{k-1}}\right) & \\
\qquad + \left[\dfrac{\widehat{Y}}{D y_k} - \dfrac{\widehat{Y}}{\widehat{Y} + \widehat{A}_k}\right] \log\left(\dfrac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right), & \text{if } k = 2, \ldots, n-1, \\
\left(1 - \dfrac{\widehat{Y}}{D y_n}\right) \log\left(\dfrac{D_n}{D_{n-1}}\right), & \text{if } k = n.
\end{cases} \tag{6.7}$$

The particular case of inference (point and variance estimator) with non-

weighted observations is fully discussed and applied in Greselin et al. (2010).

## 6.3 Approximation of the variance by linearization

### 6.3.1 Linearization by the Demnati-Rao approach

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

$$\widehat{\theta} - \theta \approx \sum_{\ell \in S} d_\ell v_\ell - \sum_{\ell \in U} v_\ell.$$

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

$$\text{var}\left(\sum_{\ell \in S} d_\ell v_\ell\right) \approx \text{var}(\widehat{\theta}).$$

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

The linearization method was introduced by Woodruff (1971) using Taylor series. Deville (1999) presented a more general method based on influence functions (Hampel, 1974; Hampel et al., 1985). In both methods, the linearized variable is computed on the function of interest and is then estimated on the sample. Binder (1996) proposed a direct approach in which the linearized variable is directly computed on the estimator. However, like in Woodruff (1971), it is only adapted for smoothed functions of totals. Demnati & Rao (2004) have proposed yet another direct approach which is of broad application. In the Demnati-Rao approach, we consider weights $a_\ell = d_\ell I_\ell$, for all $\ell \in U$, where

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

An estimator $\widehat{\theta}$ can be written as a function of the weights $a_\ell$: $\widehat{\theta} = f(a_1, a_2, ..., a_N)$. The population parameter is obtained by replacing the $a_\ell$'s

by 1's: $\theta = f(1, 1, ..., 1)$. By using Taylor series expansion, we can write

$$\widehat{\theta} \approx \theta + \sum_{\ell \in U} (a_\ell - 1) \frac{\partial \widehat{\theta}}{\partial a_\ell}.$$

Thus,

$$\widehat{\theta} - \theta \approx \sum_{\ell \in S} d_\ell \hat{v}_\ell - \sum_{\ell \in U} \hat{v}_\ell,$$

where

$$\hat{v}_\ell = \frac{\partial \widehat{\theta}}{\partial a_\ell} = \frac{\partial \widehat{\theta}}{\partial d_\ell}.$$

In this paper, we use the Demnati & Rao (2004) approach to derive an estimated linearized variable of the Zenga index, and consequently a variance estimator. The estimated linearized variable $\hat{v}_\ell$ is computed directly on the sample and obtained by calculating the partial derivative of the estimator with respect to the weight $d_\ell$. Once $\hat{v}_\ell$ is computed, variance estimation is done in the standard framework and usual asymptotic conditions of linearization techniques (Woodruff, 1971; Isaki & Fuller, 1982; Deville & Särndal, 1992; Binder, 1996; Kovacevic & Binder, 1997; Deville, 1999). Note that the design weights $d_\ell$ are used, but the method holds for calibration weights as well (Demnati & Rao, 2004).

### 6.3.2 Linearization of the Zenga index

Using the Demnati-Rao approach, the estimated linearized variable $\hat{v}_\ell$ of the Zenga Index at $y_\ell$ can be computed by

$$\hat{v}_\ell = \frac{\partial \widehat{Z}}{\partial d_\ell} = \sum_{k \in S} \frac{\partial \widehat{Z}_k}{\partial d_\ell}. \tag{6.8}$$

Thus, for each sample element $\ell$, the partial derivative with respect to $d_\ell$ of $\widehat{Z}_k$ for all $k \in S$ is computed. Similarly as for point estimation, three cases are derived. We present hereafter the final expressions for $\partial \widehat{Z}_k / \partial d_\ell$ and have included the complete derivation of Expression (6.9) in Appendix A.1.

$$\frac{\partial \widehat{Z}_k}{\partial d_\ell} = \begin{cases} \frac{Dy_\ell - \widehat{Y}}{D^2 y_1} \log\left(\frac{\widehat{Y}}{\widehat{Y} - \widehat{Y}_1}\right) + y_\ell \left(\frac{\widehat{Y}}{Dy_1} - 1\right) \left[\frac{1}{\widehat{Y}} - \frac{\mathbb{1}(\ell > 1)}{\widehat{Y} - \widehat{Y}_1}\right], & \text{if } k = 1, \\ \frac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell < k) - \widehat{A}_k y_\ell}{(\widehat{Y} + \widehat{A}_k)^2} \log\left[\frac{D_k(\widehat{Y} - \widehat{Y}_{k-1})}{D_{k-1}(\widehat{Y} - \widehat{Y}_k)}\right] \\ \quad + \frac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k} \left[\frac{\mathbb{1}(\ell \le k)}{D_k} - \frac{\mathbb{1}(\ell < k)}{D_{k-1}}\right] + \frac{Dy_\ell - \widehat{Y}}{D^2 y_k} \log\left(\frac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right) \\ \quad + \frac{\widehat{Y} y_\ell}{\widehat{Y} - \widehat{Y}_{k-1}} \left[\mathbb{1}(\ell = k) - \frac{y_k d_k}{\widehat{Y} - \widehat{Y}_k} \mathbb{1}(\ell > k)\right] \left(\frac{1}{Dy_k} - \frac{1}{\widehat{Y} + \widehat{A}_k}\right), & \text{if } k = 2, \ldots, n-1, \\ \frac{\widehat{Y} - Dy_\ell}{D^2 y_n} \log\left(\frac{D}{D_{n-1}}\right) + \left(1 - \frac{\widehat{Y}}{Dy_n}\right) \left[\frac{1}{D} - \frac{\mathbb{1}(\ell < n)}{D_{n-1}}\right], & \text{if } k = n. \end{cases} \tag{6.9}$$

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

$$\widehat{\text{var}}(\widehat{Z}) = \frac{N(N-n)}{n(n-1)} \sum_{\ell \in S} (\hat{v}_\ell - \bar{v})^2, \tag{6.10}$$

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

## 6.4 Simulation studies

### 6.4.1 Synthetic Austrian EU-SILC data

At first, a simulation study is run in the R environment (R Development Core Team, 2010) on a synthetic data set generated from original Austrian EU-SILC data. The data is available from the *laeken* R package (Alfons et al., 2010) and incorporates 14,824 non-null individual observations from 6,000 households. The simulation design is kept simple: data at the individual level is considered to be the finite population from which random samples of size $n = 3,000$ are selected with a simple random sampling design without replacement. One thousand replications are made. In each sample, the Zenga index (Expression 6.6) and its linearization variance (Expression 6.10) are estimated. Results are summarized in Table 6.1. The relative bias for point and variance estimation are defined respectively by

$$\text{RB}(\widehat{Z}) = \frac{\text{E}(\widehat{Z}) - Z}{Z},$$

and

$$\text{RB}[\widehat{\text{var}}_{lin}(\widehat{Z})] = \frac{\text{E}[\widehat{\text{var}}_{lin}(\widehat{Z})] - \widehat{\text{var}}_{sim}(\widehat{Z})}{\widehat{\text{var}}_{sim}(\widehat{Z})},$$

where $\widehat{\text{var}}_{lin}(\widehat{Z})$ stands for the estimated variance obtained with the linearization technique and $\widehat{\text{var}}_{sim}(\widehat{Z})$ denotes the Monte-Carlo variance com-

puted on the 1000 replications. Results show that point estimation is very successful and that the linearization technique only very slightly underestimates the variance with a relative bias of $-1.65\%$. The coverage rate for a 95% confidence interval is close to the desired level.

Table 6.1 – *Simulation results (Austrian EU-SILC data, 1000 replications).*

| Point estimation | $Z$ | $\mathrm{E}(\widehat{Z})$ | $\mathrm{RB}(\widehat{Z})$ |
|---|---|---|---|
| | 0.5872 | 0.5870 | $-0.04\%$ |
| Variance Estimation | $\widehat{\mathrm{var}}_{sim}(\widehat{Z})$ | $\mathrm{E}[\widehat{\mathrm{var}}_{lin}(\widehat{Z})]$ | $\mathrm{RB}(\widehat{\mathrm{var}}_{lin})$ |
| | $3.0310 \cdot 10^{-5}$ | $2.9811 \cdot 10^{-5}$ | $-1.65\%$ |
| Coverage Rate for $Z$ (95% confidence interval) | 95.9 | | |

### 6.4.2 Taxable incomes of Canton of Neuchâtel, Switzerland

In the previous example, extreme observations do not have a large effect on the accuracy of estimation. Our second simulation study is run on real taxable income data in the Canton of Neuchâtel, Switzerland for year 2006. It is composed of all 82,489 non-null taxpayers of the canton and includes some extreme observations. The same strategy, design and sample size are used as for the first simulation study in order to allow for comparisons. To account for the extreme observations issue, two sets of simulations are performed: one on the full data set and one on truncated data. In the truncated data, all observations lying above $Q_{0.999}$ are deleted, involving the 83 richest income earners. Truncation of the data reduces the ratio between the largest income and the median income by a factor of 13.2. The results are summarized in Table 6.2. Note that estimates and true values of the Zenga index and its sampling variance for the Neuchâtel data are not displayed in Table 6.2 because this data set has been made available to us exclusively for academic purposes. Thus, the quality of estimation for this data is merely summarized by relative biases and coverage rates. For similar reasons, income values in Figure 6.2 as well as Zenga and Gini index estimates in Figure 6.3 have been masked.

Table 6.2 – *Simulation results (Neuchâtel data, 1000 replications).*

| | Point estimation $\mathrm{RB}(\widehat{Z})$ | Variance estimation $\mathrm{RB}\,(\widehat{\mathrm{var}}_{lin})$ | Coverage Rate CR (95%) |
|---|---|---|---|
| Full data | $-0.08\%$ | $-6.79\%$ | 93.5% |
| Truncated data | $-0.06\%$ | 0.22% | 94.7% |

Although point estimation is accounted for in a satisfactory manner

in both situations, we can see that the variance is not as well estimated when the most severely extreme observations are part of the population. However, it can be advocated that even in the presence of extreme values, which we believe to be frequent when working on income data, quality of inference for the Zenga index remains reasonable with a relative bias for the variance of $-6.79\%$ and a coverage rate of 93.5% for a 95% confidence interval.

## 6.5 Comparison with the Gini index

### 6.5.1 Definition and properties of the Gini index

Working on a new synthetic inequality index like the Zenga index raises one key question: Do we need yet another inequality measure? The question is not without merit considering the vast collection of already existing inequality measures and the amount of research dedicated to enhancing the quality of inference for these measures. However, by comparing the Zenga index to the leading inequality measure, the Gini index, we try to point out why the Zenga index is a serious and interesting alternative to existing indices. Let us first define the Gini index $G$ by

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

an estimator of $G$ by

$$\widehat{G} = \frac{2}{D\widehat{Y}}\sum_{k\in S} d_k D_k y_k - \left(1 + \frac{1}{D\widehat{Y}}\sum_{k\in S} d_k^2 y_k\right) = \frac{\sum_{k\in S}\sum_{\ell\in S} d_k d_\ell \,|y_k - y_\ell|}{2D\widehat{Y}},$$

and a linearized variable estimated on the sample (Monti, 1991; Langel & Tillé, 2011a) by

$$\widehat{u}_k = \frac{1}{D\widehat{Y}}\left[2D_k(y_k - \widehat{\widehat{Y}}_k) + \widehat{Y} - Dy_k - \widehat{G}(\widehat{Y} + y_k D)\right],$$

with

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

Both indices have thus in common that they can be defined by means of the Lorenz curve $L(\alpha)$. Also, both measures fulfill the most common properties of the axiomatic approach to inequality theory (Cowell & Kuga, 1981b) such as anonymity, scale invariance, population principle or principle of transfers (Zenga, 2007). Moreover, Radaelli (2010) proposed a subgroup decomposition of the Zenga index which is closely related to the decomposition of the Gini index (Dagum, 1997).

### 6.5.2 Advantages of the Zenga index

The two measures differ however in many ways. One argument in favor of the Zenga index is described by Greselin et al. (2010, p.3):

> [...] the Gini index underestimates comparisons between the very poor and the whole population and emphasizes comparisons which involve almost identical population subgroups [...] the Zenga index detects, with the same sensibility, all deviations from equality in any part of the distribution.

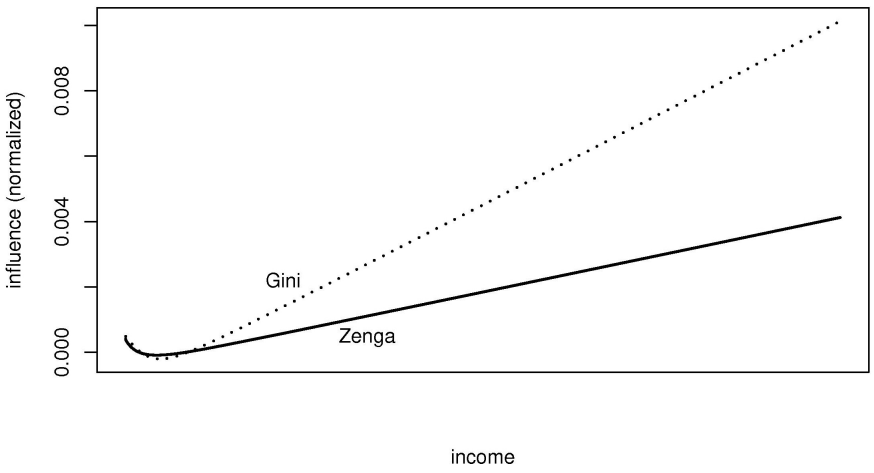
A comparative simulation study regrouping 17 different inequality indices (Langel & Tillé, 2011b) seems to confirm this idea by showing that the Zenga index is one of the most appropriate measures to detect changes at any level of the income distribution and in many different situations. Another argument in favor of the Zenga index concerns interpretation. A lot of intuitive information can indeed be obtained from analyzing the curve itself. For instance, any point measure $Z(\alpha)$ on the curve indicates that the mean income of the $100\alpha\%$ poorest is equal to $[1 - Z(\alpha)]$ times the mean income of the richest $100(1-\alpha)\%$. Moreover, the Zenga index can be plotted alongside the curve and thus, clearly displays the intervals of $\alpha$ where inequality is lower or higher than the mean level of inequality, which is represented by the index itself. Finally, Maffenini & Polisicchio (2010) have shown that when adding an identical positive income to all observations, the effect on the Zenga curve is more intuitive than on the Lorenz curve. Indeed, the Zenga curve shows that, after translation, the level of inequality decreases more heavily for small incomes than for larger ones, whereas the latter intuition is not captured by the Lorenz curve.

### 6.5.3 Influence functions and sampling distributions

In statistics, influence functions (Hampel, 1974) are mainly used as a tool to study robustness. However, Deville (1999) showed that the linearized variable is an influence function, only very slightly modified from the definition of (Hampel, 1974) so that it could be used within a finite population framework. Thus, it is possible to study the sensitivity to extreme observations of the statistic of interest simply by analyzing its linearized variable, or influence curve. Unsurprisingly, the influence curve of the Zenga index shows that the statistic is highly sensitive to extreme observations. As a result, inference can be heavily affected by the presence of very large incomes in the sample. Similar results are found in robust statistics regarding the influence function of the Gini index (Monti, 1991; Cowell & Victoria-Feser, 1996a, 2003) and the Quintile Share Ratio (Hulliger & Schoch, 2009), which are both unbounded from above.

However, a comparative study with the Gini index reveals an interesting result. Figure 6.2 displays the influence function of the Gini in-

Figure 6.2 – *Normalized influence curves of the Zenga index and Gini index estimated on one sample of size* $n = 3000$ *drawn from the Neuchâtel data set.*

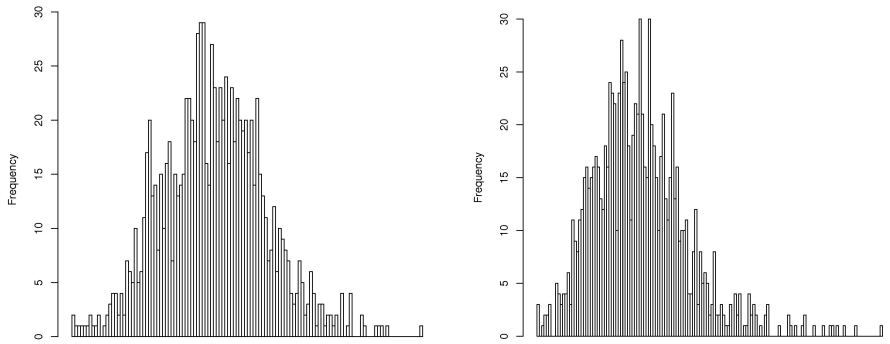



dex alongside that of the Zenga index computed on one sample of size $n = 3000$ from the Neuchâtel simulation study. To allow for comparisons, both influence functions are normalized following the notion of *relative influence function* proposed by Cowell & Flachaire (2007). The estimated linearized variable of the Gini and Zenga indices are thus divided by the value of the respective index estimated on the sample. Figure 6.2 shows that the Zenga index is significatively less affected by extreme observations than the Gini index. This is a very important advantage of the Zenga index because inference from income data is often confronted with extreme values.

The outcome of this feature is that the sampling distribution of the Zenga index is by far closer to the Normal distribution than that of the Gini index, allowing for the construction of more reliable confidence intervals. This intuition is confirmed by a small simulation study performed to estimate the skewness and excess kurtosis of the sampling distribution of both indices $\widehat{G}$ and $\widehat{Z}$. We have simulated 1000 samples (simple random sampling design without replacement) of size $n = 100$ from the Neuchâtel income data and estimated both indices in each sample. The histograms in Figure 6.3 displays the respective sampling distributions of $\widehat{G}$ and $\widehat{Z}$ for this set of simulations. They show that the sampling distribution of the Zenga index is clearly more symmetric than that of the Gini index.

The skewness and excess kurtosis for each index are then estimated on the 1000 samples. The results, displayed in Table 6.3, show that unlike what is observed with the Zenga index, the sampling distribution of the Gini index is a serious obstacle in the construction of good confidence intervals around $\widehat{G}$. Indeed, the skewness and excess kurtosis of $\widehat{G}$ are far from the desired level (0 for both statistics).

Figure 6.3 – *Histograms of the distributions of* $\widehat{Z}$ *and* $\widehat{G}$ *computed on* 1000 *samples of size* $n = 100$.




(a) Zenga index $\widehat{Z}$ (b) Gini index $\widehat{G}$

Table 6.3 – *Skewness and kurtosis: Simulation results.*

| | Zenga index $\widehat{Z}$ | Gini index $\widehat{G}$ |
|---|---|---|
| skewness | 0.22 | 1.15 |
| excess kurtosis | 0.24 | 2.22 |

## 6.6 Conclusion

To effectively bring new insights in the study of income inequality a recent measure like the Zenga index needs a general and valid framework for inference in finite populations. In this paper, we have firstly proposed an estimator of the Zenga index which takes the sampling design into account. Secondly, a variance estimator has been presented. The Demnati and Rao linearization technique has been used to derive an estimator that can be applied to samples selected from a complex sampling design. The theoretical results have then been tested successfully in simulation studies. Finally the relevance of the Zenga index has been emphasized by comparing it to the Gini index. It it shown that in addition to having similar properties as the Gini index, the characteristics of the Zenga index facilitate reliable inference. Indeed, in the presence of extreme observations, the sampling distribution of $\widehat{Z}$ is both markedly less skewed and less heavy-tailed than that of the Gini index. Moreover, the Zenga index and its underlying curve display interesting graphical interpretations. We hope that these features can motivate the use of the Zenga index in future research studies and applications.

## Acknowledgments

The authors are grateful to the Office Cantonal de la Statistique (Canton de Neuchâtel, Switzerland) and especially to Gérard Geiser for the dataset. The authors would also like to thank two anonymous referees for insightful comments and suggestions. This research is supported by grant no. 200021-121604 of the Swiss National Science Foundation.

# General conclusion

As this work points out on different occasions, research on inequality measures is both profuse and fragmented. In the context of survey methodology, emphasis is often placed on specific issues and measures rather than on setting up a general framework. The economic literature, in contrast, focuses on a large panel of inequality indices and on their properties, paying only little attention to inferential issues. Though leaving out a lot of open questions, this document tries to take advantage of both the recent developments in survey methodology and the abundant literature on income inequality to propose a unified setting where reliable inference can be achieved in a large number of situations. As discussed throughout the document, the complexity of both the statistic of interest and the sampling design are obstacles that this methodology tries to overcome. This is operated by two main achievements: clarification of the finite population setting and notation on the one hand, the use of linearization techniques for variance estimation on the other.

Firstly, the use of a coherent notation allows us to propose finite population expressions of most inequality measures as well as estimators that are able to handle survey weights. In particular, the finite population Lorenz curve (or quantile share), which appears in numerous measures, is expressed and estimated in a straightforward and convenient way. Secondly, linearization is a very general approach to variance estimation. Indeed, once the asymptotic framework is assumed to hold, the approach simply requires the statistic of interest to be linearizable and the sampling design to allow for unbiased variance estimation of a linear estimator. While the latter condition is usually met, the former has benefited from recent developments by Deville (1999) and Demnati & Rao (2004) allowing for a larger class of statistics to admit a linearized variable. In addition, our result for the linearization of the partial sum (the numerator of the finite population quantile share) in Chapter 3 shows the uselessness of applying kernel smoothing on the density functions when linearizing a quantile share.

The diversity of inequality indicators that are discussed and proposed in the literature can thus be estimated under typical complex sampling designs used in official statistics. Variance estimators of most inequality measures can be directly derived from the linearized variables proposed across the previous chapters and the Appendix. Although linearization

may seem somewhat intricate in some cases (e.g. the Zenga index), it is not a difficult, nor a computer intensive, method for practitioners to use once the expression of the linearized variable is implemented into a software package.

Due to his contribution to both inequality measures and survey sampling, a large portion of the document is dedicated to the work of Corrado Gini. Unlike the QSR or the Zenga index, the Gini index has already been at the center of a lot of research. For that reason, Gini's work has been discussed here from a different, more historical angle. His work on balanced sampling is first used as a pretext for a historical analysis of survey sampling. Chapter 4 shows how balanced sampling has reshaped the distinction between purposive selection and random sampling methods. Besides, the use of the linearized variable of the Gini index as balancing constraints has shown to yield interesting improvements in terms of accuracy of estimation.

While the lack of literature on finite population inference for the QSR or the Zenga index is obvious, the situation for the Gini index is opposite. Rather than innovations, what is needed regarding inference for the latter is a clarification and a comprehensive understanding of the state of the art on the subject. This was the objective of Chapter 5, where variance estimation for the Gini index is studied both vertically (from a historical perspective) and horizontally (comparison of different approaches). An important result of this part of the research reveals the misconception leading to the attractive yet incorrect regression approach to estimate the variance of the Gini index.

The Gini index appears again in Chapter 6 where comparisons with the Zenga index show that there is room for competing inequality measures. Indeed, the popularity of the Gini index seems to be based on habit rather than on some intrinsic qualities. While no existing measure appears to be unilaterally better than others, the Gini index is outdistanced by other measures in most main concerns such as interpretability, robustness or decomposability.

Studying income inequality in the finite population context also paves the way for possible further lines of research. In particular, focusing on a fully non-parametric approach points out some issues relating to the structure of the income distribution itself. Though not always the case, the allocation of income in a finite population generally results in a very positively skewed income distribution with a frequent occurrence of extreme values. Usual inequality measures such as the QSR, the Theil, Gini or Atkinson indices are very sensitive to extreme observations. This fragility is increased by the fact that reliable data is known to be more difficult to collect at the tails of the distribution. Taking this issue into account is essential and further investigation focusing on this aspect can be operated in at least four directions.

As a start, one can place emphasis on the use of auxiliary information, for exemple by overrepresenting units at the tails of the distribution. This can be done at both the design and estimation stage. As shown in Chapter 4, linearized variables of non-linear auxiliary information can be favorably used as balancing constraints. Lesage (2009) also suggests using linearized variables in calibration procedures.

Another direction for further research is to study the survey non-response mechanism associated with income. In recent years, new developments related to the treatment of non-ignorable unit non-response have been proposed (Deville, 2002; Särndal & Lundström, 2005). A non-response pattern is said to be non-ignorable when the response mechanism depends on the character of interest, which is a probable situation when the latter is income. Thus, taking this mechanism into account may have a meaningful effect on the estimation of income inequality measures and improve the quality of inference.

Some recent research in the parametric approach to income inequality also deserves further attention. A four-parameter distribution, the Generalized Beta Distribution of the Second Kind (GB2) proposed initially by McDonald (1984) and studied in the survey context by Graf et al. (2011) has shown good results for the parametric estimation of inequality measures.

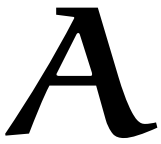
Lastly, while different robust estimators of highly sensitive measures like the QSR or Gini index are suggested in Hulliger et al. (2011), robust inequality measures like plain ratios of quantiles have not been much explored in the literature. Although often considered as too insensitive, the inter-decile ratio (IDR) is nevertheless widely used by social science practitioners. A detailed analysis of this class of measure in the field of survey sampling would be of interest.

Whereas the above lines of research are centered on improving the quality of estimation inside the proposed framework, another topic of interest is to broaden this framework from the one-sample case to longitudinal surveys. While Chapter 2 discusses the question of sensitivity to change of various inequality measures, the estimation of the measure of change is not addressed in this work. In a finite population setting, estimating the sampling variance of change in the level of inequality between two samples drawn at two different occasions is intricate. Indeed, the two samples are generally not independent and the covariance between the inequality measure estimated at both occasions has to be taken into account. Goga et al. (2009) have recently generalized the influence function approach of Deville (1999) to the two-sample case, allowing for the variance estimation of change of inequality measures like the Gini index (see also Münnich & Zins, 2011). This approach could be applied to other inequality measures and be used by practitioners to decide whether an

observed variation in the inequality level through time is significant or instead only due to random sampling errors.

A final and general remark on income inequality is that the choice of the measure(s) computed by the practitioners is generally not solely determined by statistical concerns. For instance, Laeken indicators have been chosen for a variety of reasons such as interpretability and popularity. Therefore, although statistical science should provide insights into which measure should be used when and why, it should also be able to propose a reliable methodology for a whole set of measures. Even if some particular inequality indices are studied more thoroughly than others, it is one of the goals that this research has tried to achieve, notably by proposing a linearized variable for many common indices in the Appendix section. With the hope that inequality measure estimates will be more systematically associated with a variance estimate, confidence interval or standard error in the future.

# Appendix



## A.1 Linearization of the Zenga index [1]

The three cases of Expression (6.7) are linearized separately in order to obtain an estimated linearized variable $\hat{v}_\ell$ such that

$$\hat{v}_\ell = \frac{\partial \widehat{Z}}{\partial d_\ell} = \sum_{k \in S} \frac{\partial \widehat{Z}_k}{\partial d_\ell}. \tag{A.1}$$

### A.1.1 Linearization of $\widehat{Z}_k$ for $k = 2, ..., n-1$

First, $\widehat{Z}_k$ for $k = 2, ..., n$ is rewritten as the sum of two terms, $P_1$ and $P_2$:

$$\begin{aligned} \widehat{Z}_k &= \underbrace{\frac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k} \log\left(\frac{D_k}{D_{k-1}}\right)}_{P_1} + \underbrace{\left[\frac{\widehat{Y}}{Dy_k} - \frac{\widehat{Y}}{\widehat{Y} + \widehat{A}_k}\right] \log\left(\frac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right)}_{P_2} \\ &= P_1 + P_2. \end{aligned}$$

Thus,

$$\frac{\partial \widehat{Z}_k}{\partial d_\ell} = \frac{\partial P_1}{\partial d_\ell} + \frac{\partial P_2}{\partial d_\ell}, \tag{A.2}$$

and the derivation can be split into two separate steps, the linearization of terms $P_1$ and $P_2$. Linearization of term $P_1$ can be done by computing the partial derivative with respect to $d_\ell$. Using differentiation rules, we obtain

$$\begin{aligned} \frac{\partial P_1}{\partial d_\ell} &= \left[(\widehat{Y} + \widehat{A}_k) \log\left(\frac{D_k}{D_{k-1}}\right) \frac{\partial \widehat{A}_k}{\partial d_\ell} - \widehat{A}_k \log\left(\frac{D_k}{D_{k-1}}\right) \frac{\partial(\widehat{Y} + \widehat{A}_k)}{\partial d_\ell}\right] \\ &\quad \times \frac{1}{(\widehat{Y} + \widehat{A}_k)^2} + \frac{D_{k-1}}{D_k} \frac{\partial\left(\frac{D_k}{D_{k-1}}\right)}{\partial d_\ell} \frac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k}. \end{aligned} \tag{A.3}$$

[1] Appendix A.1 is a reprint of the Appendix section in Langel, M. and Tillé, Y. (2011). Inference by linearization for Zenga's new inequality index: A comparison with the Gini index. *Metrika*, doi:10.1007/s00184-011-0369-1, 1-18.
(The original publication is available at www.springerlink.com)

We now compute the derivatives that are needed in Equation (A.3):

$$\frac{\partial \widehat{Y}}{\partial d_\ell} = y_\ell, \tag{A.4}$$

$$\frac{\partial \widehat{A}_k}{\partial d_\ell} = (y_k - y_\ell)\,\mathbb{1}(\ell < k), \tag{A.5}$$

$$\frac{\partial \left(\frac{D_k}{D_{k-1}}\right)}{\partial d_\ell} = \frac{D_{k-1}\mathbb{1}(\ell \leq k) - D_k\mathbb{1}(\ell < k)}{D_{k-1}^2}, \tag{A.6}$$

and replace them in Expression (A.3) to obtain

$$\begin{aligned}\frac{\partial P_1}{\partial d_\ell} &= \frac{\widehat{A}_k}{\widehat{Y}+\widehat{A}_k}\left[\frac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell < k) - \widehat{A}_k y_\ell}{\widehat{A}_k(\widehat{Y}+\widehat{A}_k)}\log\left(\frac{D_k}{D_{k-1}}\right)\right.\\ &\quad\left. - \frac{\mathbb{1}(\ell < k)}{D_{k-1}} + \frac{\mathbb{1}(\ell \leq k)}{D_k}\right].\end{aligned} \tag{A.7}$$

Similarly, for term $P_2$:

$$\begin{aligned}\frac{\partial P_2}{\partial d_\ell} &= \left[\frac{\widehat{Y}\frac{\partial(\widehat{Y}+\widehat{A}_k)}{\partial d_\ell}}{(\widehat{Y}+\widehat{A}_k)^2} - \frac{\frac{\partial \widehat{Y}}{\partial d_\ell}}{\widehat{Y}+\widehat{A}_k} + \frac{\frac{\partial \widehat{Y}}{\partial d_\ell}}{Dy_k} - \frac{\widehat{Y}\frac{\partial(Dy_k)}{\partial d_\ell}}{(Dy_k)^2}\right]\log\left(\frac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right)\\ &\quad + \frac{\widehat{Y}-\widehat{Y}_k}{\widehat{Y}-\widehat{Y}_{k-1}}\frac{\partial\left(\frac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right)}{\partial d_\ell}\left(\frac{\widehat{Y}}{Dy_k} - \frac{\widehat{Y}}{\widehat{Y}+\widehat{A}_k}\right).\end{aligned} \tag{A.8}$$

In addition to Result (A.4), the following derivatives are needed :

$$\frac{\partial(\widehat{Y}+\widehat{A}_k)}{\partial d_\ell} = y_\ell - (y_\ell - y_k)\,\mathbb{1}(\ell < k), \tag{A.9}$$

$$\frac{\partial Dy_k}{\partial d_\ell} = y_k, \tag{A.10}$$

$$\frac{\partial\left(\frac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right)}{\partial d_\ell} = \frac{y_\ell}{\widehat{Y}-\widehat{Y}_k}\left[\mathbb{1}(\ell = k) - \frac{y_k d_k}{\widehat{Y}-\widehat{Y}_k}\mathbb{1}(\ell > k)\right]. \tag{A.11}$$

Results (A.4), (A.9), (A.10) and (A.11) are substituted into Equation (A.8):

$$\begin{aligned}\frac{\partial P_2}{\partial d_\ell} &= \left[\frac{y_\ell\widehat{A}_k - \widehat{Y}\,(y_\ell - y_k)\,\mathbb{1}(\ell < k)}{(\widehat{Y}+\widehat{A}_k)^2} + \frac{Dy_\ell - \widehat{Y}}{D^2 y_k}\right]\log\left(\frac{\widehat{Y}-\widehat{Y}_{k-1}}{\widehat{Y}-\widehat{Y}_k}\right)\\ &\quad + \frac{\widehat{Y}y_\ell}{\widehat{Y}-\widehat{Y}_{k-1}}\left[\mathbb{1}(\ell = k) - \frac{y_k d_k}{\widehat{Y}-\widehat{Y}_k}\mathbb{1}(\ell > k)\right]\\ &\quad \times\left(\frac{1}{Dy_k} - \frac{1}{\widehat{Y}+\widehat{A}_k}\right).\end{aligned} \tag{A.12}$$

The final expression for the linearization of $\widehat{Z}_k$ for $k = 2, ..., n-1$ is obtained by replacing (A.7) and (A.12) in (A.2):

$$\begin{aligned}
\frac{\partial \widehat{Z}_k}{\partial d_\ell} &= \frac{\partial P_1}{\partial d_\ell} + \frac{\partial P_2}{\partial d_\ell}, \\
&= \frac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k} \left[ \frac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell < k) - \widehat{A}_k y_\ell}{\widehat{A}_k(\widehat{Y} + \widehat{A}_k)} \log\left(\frac{D_k}{D_{k-1}}\right) \right. \\
&\quad \left. - \frac{\mathbb{1}(\ell < k)}{D_{k-1}} + \frac{\mathbb{1}(\ell \leq k)}{D_k} \right] \\
&\quad + \left[ \frac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell < k) - \widehat{A}_k y_\ell}{(\widehat{Y} + \widehat{A}_k)^2} + \frac{D y_\ell - \widehat{Y}}{D^2 y_k} \right] \log\left(\frac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right) \\
&\quad + \frac{\widehat{Y} y_\ell}{\widehat{Y} - \widehat{Y}_{k-1}} \left[ \mathbb{1}(\ell = k) - \frac{y_k d_k}{\widehat{Y} - \widehat{Y}_k} \mathbb{1}(\ell > k) \right] \left( \frac{1}{D y_k} - \frac{1}{\widehat{Y} + \widehat{A}_k} \right).
\end{aligned}$$

The latter can be rewritten by

$$\begin{aligned}
\frac{\partial \widehat{Z}_k}{\partial d_\ell} &= \frac{\widehat{Y}\,(y_k - y_\ell)\,\mathbb{1}(\ell < k) - \widehat{A}_k y_\ell}{(\widehat{Y} + \widehat{A}_k)^2} \log\left[ \frac{D_k(\widehat{Y} - \widehat{Y}_{k-1})}{D_{k-1}(\widehat{Y} - \widehat{Y}_k)} \right] \\
&\quad + \frac{\widehat{A}_k}{\widehat{Y} + \widehat{A}_k} \left[ \frac{\mathbb{1}(\ell \leq k)}{D_k} - \frac{\mathbb{1}(\ell < k)}{D_{k-1}} \right] + \frac{D y_\ell - \widehat{Y}}{D^2 y_k} \log\left(\frac{\widehat{Y} - \widehat{Y}_{k-1}}{\widehat{Y} - \widehat{Y}_k}\right) \\
&\quad + \frac{\widehat{Y} y_\ell}{\widehat{Y} - \widehat{Y}_{k-1}} \left[ \mathbb{1}(\ell = k) - \frac{y_k d_k}{\widehat{Y} - \widehat{Y}_k} \mathbb{1}(\ell > k) \right] \\
&\quad \times \left( \frac{1}{D y_k} - \frac{1}{\widehat{Y} + \widehat{A}_k} \right).
\end{aligned} \tag{A.13}$$

### A.1.2 Linearization of $\widehat{Z}_1$

The case for $k = 1$ can be derived:

$$\begin{aligned}
\frac{\partial \widehat{Z}_1}{\partial d_\ell} &= \log\left(\frac{\widehat{Y}}{\widehat{Y} - \widehat{Y}_1}\right) \left[ \frac{D y_1 \frac{\partial \widehat{Y}}{\partial d_\ell} - \widehat{Y} \frac{\partial (D y_1)}{\partial d_\ell}}{(D y_1)^2} \right] + \left( \frac{\widehat{Y}}{D y_1} - 1 \right) \\
&\quad \times \frac{\widehat{Y} - \widehat{Y}_1}{\widehat{Y}} \frac{\partial \left( \frac{\widehat{Y}}{\widehat{Y} - \widehat{Y}_1} \right)}{\partial d_\ell}, \\
&= \frac{D y_\ell - \widehat{Y}}{D^2 y_1} \log\left(\frac{\widehat{Y}}{\widehat{Y} - \widehat{Y}_1}\right) + y_\ell \left( \frac{\widehat{Y}}{D y_1} - 1 \right) \\
&\quad \times \left[ \frac{1}{\widehat{Y}} - \frac{\mathbb{1}(\ell > 1)}{\widehat{Y} - \widehat{Y}_1} \right].
\end{aligned} \tag{A.14}$$

### A.1.3 Linearization of $\widehat{Z}_n$

Finally the $k = n$ case is also linearized:

$$\begin{aligned}\frac{\partial \widehat{Z}_n}{\partial d_\ell} &= \log\left(\frac{D}{D_{n-1}}\right)\left[\frac{\widehat{Y}\frac{\partial(Dy_n)}{\partial d_\ell} - Dy_n\frac{\partial \widehat{Y}}{\partial d_\ell}}{(Dy_n)^2}\right] \\ &\quad + \left(1 - \frac{\widehat{Y}}{Dy_n}\right)\frac{D_{n-1}}{D}\frac{\partial\left(\frac{D}{D_{n-1}}\right)}{\partial d_\ell}, \\ &= \frac{\widehat{Y} - Dy_\ell}{D^2 y_n}\log\left(\frac{D}{D_{n-1}}\right) + \left(1 - \frac{\widehat{Y}}{Dy_n}\right)\left[\frac{1}{D} - \frac{\mathbb{1}(\ell < n)}{D_{n-1}}\right]. \quad \text{(A.15)}\end{aligned}$$

### A.1.4 Linearization of the Zenga index

Finally, by recalling Expression (A.1) and combining Results (A.13), (A.14) and (A.15) into

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

an estimated linearized variable $\hat{v}_\ell$ can now be obtained for all $\ell \in S$, and thus a variance estimator for $\widehat{Z}$, the Zenga index estimated from a sample.

## A.2 Linearization of the Atkinson index

### A.2.1 Atkinson index with $\epsilon \in [0,1)$

The Atkinson index in Expression 2.2.3 can be rewritten for $\epsilon \in [0,1)$ as a function of totals by

$$\begin{aligned} \mathrm{A}_\epsilon &= 1 - \left[\frac{1}{N}\sum_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^{1-\epsilon}\right]^{1/(1-\epsilon)} \\ &= 1 - \left(N^{-\epsilon}Y^{\epsilon-1}\sum_{k\in U} y_k^{1-\epsilon}\right)^{1/(1-\epsilon)} \\ &= 1 - N^{\epsilon/(\epsilon-1)}Y^{-1}B^{1/(1-\epsilon)}, \end{aligned}$$

where $b_k = y_k^{1-\epsilon}$ and $B = \sum_{k\in U} b_k$. A linearized variable $v_\ell^A$ for $\mathrm{A}_\epsilon$ is obtained by applying the Deville (1999) influence function approach, thus by computing partial derivatives with respect to these totals:

$$\begin{aligned} v_\ell^A &= y_\ell \frac{\partial \mathrm{A}_\epsilon}{\partial Y} + b_\ell \frac{\partial \mathrm{A}_\epsilon}{\partial B} + 1\frac{\partial \mathrm{A}_\epsilon}{\partial N} \\ &= y_\ell N^{\epsilon/(\epsilon-1)}Y^{-2}B^{1/(1-\epsilon)} - b_\ell \frac{N^{\epsilon/(\epsilon-1)}Y^{-1}B^{\epsilon/(1-\epsilon)}}{1-\epsilon} \\ &\quad - \frac{\epsilon N^{1/(\epsilon-1)}Y^{-1}B^{1/(1-\epsilon)}}{\epsilon-1} \\ &= [N^{\epsilon/(\epsilon-1)}Y^{-1}B^{1/(1-\epsilon)}]\left(\frac{y_\ell}{Y} - \frac{1}{1-\epsilon}\frac{b_\ell}{B} - \frac{\epsilon}{\epsilon-1}\frac{1}{N}\right) \\ &= (1-\mathrm{A}_\epsilon)\left[\frac{y_\ell}{Y} - \frac{b_\ell}{(1-\epsilon)B} - \frac{\epsilon}{(\epsilon-1)N}\right]. \end{aligned}$$

### A.2.2 Atkinson index with $\epsilon = 1$

The Atkinson index for $\epsilon = 1$ can be rewritten as

$$\begin{aligned} \mathrm{A}_1 &= 1 - \prod_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^{\frac{1}{N}} \\ &= 1 - \frac{N}{Y}\exp\frac{1}{N}\sum_{k\in U}\ln y_k \\ &= 1 - \frac{N}{Y}\exp\frac{B}{N}, \end{aligned}$$

where $b_k = \ln y_k$ and $B = \sum_{k\in U} b_k$. A linearized variable $v_\ell^{A1}$ for $\mathrm{A}_1$ is then

$$
\begin{aligned}
v_\ell^{A1} &= y_\ell \frac{\partial \mathrm{A}_1}{\partial Y} + b_\ell \frac{\partial \mathrm{A}_1}{\partial B} + 1 \frac{\partial \mathrm{A}_1}{\partial N} \\
&= \frac{y_\ell N \exp(B/N)}{Y^2} - \frac{b_\ell \exp(B/N)}{Y} - \frac{\exp(B/N)}{Y} + \frac{B \exp(B/N)}{NY} \\
&= \frac{1-\mathrm{A}_1}{N}\left(\frac{N y_\ell}{Y} - b_\ell - 1 + \frac{B}{N}\right).
\end{aligned}
$$

A linearized variable of the Atkinson index is also derived in Dell et al. (2002).

## A.3 Linearization of the Generalized entropy index

### A.3.1 Generalized entropy index with $\varphi \neq 0$, $\varphi \neq 1$

The Generalized entropy index for $\varphi \neq 0$, $\varphi \neq 1$ can be rewritten by

$$
\begin{aligned}
\mathrm{GE}_\varphi &= \frac{1}{\varphi(\varphi-1)}\left[\frac{1}{N}\sum_{k\in U}\left(\frac{y_k}{\overline{Y}}\right)^\varphi - 1\right] \\
&= \frac{N^{\varphi-1}Y^{-\varphi}\sum_{k\in U} y_k^\varphi - 1}{\varphi(\varphi-1)} \\
&= \frac{N^{\varphi-1}Y^{-\varphi}B}{\varphi(\varphi-1)} - \frac{1}{\varphi(\varphi-1)},
\end{aligned}
$$

where $b_k = y_k^\varphi$ and $B = \sum_{k\in U} b_k$. A linearized variable $v_\ell^{GE}$ for $\mathrm{GE}_\varphi$ is

$$
\begin{aligned}
v_\ell^{GE} &= y_\ell \frac{\partial \mathrm{GE}_\varphi}{\partial Y} + b_\ell \frac{\partial \mathrm{GE}_\varphi}{\partial B} + 1 \frac{\partial \mathrm{GE}_\varphi}{\partial N} \\
&= y_\ell \frac{-\varphi Y^{-\varphi-1}N^{\varphi-1}B}{\varphi(\varphi-1)} + b_\ell \frac{N^{\varphi-1}Y^{-\varphi}}{\varphi(\varphi-1)} + \frac{(\varphi-1)Y^{-\varphi}BN^{\varphi-2}}{\varphi(\varphi-1)} \\
&= \frac{Y^{-\varphi}N^{\varphi-1}B}{\varphi(\varphi-1)}\left(\frac{-\varphi y_\ell}{Y} + \frac{b_\ell}{B} + \frac{\varphi-1}{N}\right) \\
&= \left[\mathrm{GE}_\varphi + \frac{1}{\varphi(\varphi-1)}\right]\left(\frac{-\varphi y_\ell}{Y} + \frac{b_\ell}{B} + \frac{\varphi-1}{N}\right).
\end{aligned}
$$

### A.3.2 Theil index (Generalized entropy index with $\varphi = 1$)

The Theil index can first be rewritten as a function of totals:

$$\begin{aligned} \mathrm{T} &= \sum_{k\in U} \frac{y_k}{Y} \log \frac{N y_k}{Y} \\ &= \frac{1}{N} \sum_{k\in U} \frac{y_k}{\overline{Y}} \log \frac{y_k}{\overline{Y}} \\ &= \frac{1}{N} \frac{1}{\overline{Y}} \sum_{k\in U} y_k (\log y_k - \log \overline{Y}) \\ &= \frac{B}{Y} + \log N - \log Y, \end{aligned}$$

where $b_k = y_k \log y_k$ and $B = \sum_{k\in U} b_k$. A linearized variable $v_\ell^T$ for T can then be derived:

$$\begin{aligned} v_\ell^T &= b_\ell \frac{\partial \mathrm{T}}{\partial B} + y_\ell \frac{\partial \mathrm{T}}{\partial Y} + 1 \frac{\partial \mathrm{T}}{\partial N} \\ &= b_\ell \frac{1}{Y} + y_\ell \left( -\frac{B}{Y^2} - \frac{1}{Y} \right) + \frac{1}{N} \\ &= \frac{y_\ell}{Y} \left( \log y_\ell - \frac{B}{Y} - 1 \right) + \frac{1}{N}. \end{aligned}$$

A linearized variable of the Theil index is also given in Dell et al. (2002).

## A.4 Linearization of the Rényi divergence measure

### A.4.1 Rényi divergence with $\varphi \neq 1$

The Rényi divergence with $\varphi \neq 1$ can be rewritten as

$$\begin{aligned} \mathrm{R}_\varphi &= \frac{1}{\varphi - 1} \log \left[ N^{\varphi - 1} \sum_{k\in U} \left( \frac{y_k}{Y} \right)^\varphi \right] \\ &= \frac{\log \left( N^{\varphi-1} Y^{-\varphi} B \right)}{\varphi - 1} \\ &= \frac{(\varphi - 1) \log N - \varphi \log Y + \log B}{\varphi - 1}, \end{aligned}$$

where $b_k = y_k^\varphi$ and $B = \sum_{k\in U} b_k$. A linearized variable $v_\ell^R$ for $\mathrm{R}_\varphi$ is obtained by

$$\begin{aligned} v_\ell^R &= y_\ell \frac{\partial \mathrm{R}_\varphi}{\partial Y} + b_\ell \frac{\partial \mathrm{R}_\varphi}{\partial B} + 1 \frac{\partial \mathrm{R}_\varphi}{\partial N} \\ &= y_\ell \frac{-\varphi}{(\varphi-1)Y} + b_\ell \frac{1}{(\varphi-1)B} + \frac{1}{N} \\ &= \frac{1}{\varphi-1}\left(\frac{\varphi-1}{N} - \frac{\varphi y_\ell}{Y} + \frac{b_\ell}{B}\right). \end{aligned}$$

### A.4.2 Rényi divergence with $\varphi = 1$

The Rényi Divergence with $\varphi = 1$ is the Theil index. Derivations for a linearized variable of the Theil index are presented in Section A.3.2.

## A.5 Linearization of the interquantile ratios

A finite population expression for interquantile ratios is given by

$$\mathrm{QR} = \frac{Q_{1-\alpha}}{Q_\alpha},$$

with $\alpha \leq 0.5$. A linearized variable $v_\ell^{QR}$ for QR is thus simply

$$\begin{aligned} v_\ell^{QR} &= \frac{I(Q_{1-\alpha})_\ell}{Q_\alpha} - \frac{Q_{1-\alpha} I(Q_\alpha)_\ell}{Q_\alpha^2} \\ &= QR\left[\frac{I(Q_{1-\alpha})_\ell}{Q_{1-\alpha}} - \frac{I(Q_\alpha)_\ell}{Q_\alpha}\right], \end{aligned}$$

with $I(Q_\alpha)_\ell$ as defined in (3.14) and proposed initially by Deville (1999).

## A.6 Linearization of the quantile share ratios

The specific case of the Quintile share ratio (QSR) is detailed in Chapter 3. However, any measure of the type

$$\mathrm{SR} = \frac{Y - Y_{1-\alpha}}{Y_\alpha},$$

or

$$\widetilde{\mathrm{SR}} = \frac{Y - \widetilde{Y}_{1-\alpha}}{\widetilde{Y}_\alpha},$$

with $Y_\alpha$ and $\widetilde{Y}_\alpha$ as defined respectively in Expressions (3.1) and (3.5) and $\alpha \leq 0.5$, can be linearized in the same way as the QSR. To obtain a linearized variable of SR, one can indeed generalize to any value of $\alpha \leq 0.5$

the result from Section 3.7:

$$v_\ell^{SR} = \frac{y_\ell - [(1-\alpha)Q_{1-\alpha} - (Q_{1-\alpha} - y_\ell)\mathbb{1}(y_\ell \leq Q_{1-\alpha})]}{Y_\alpha} - \frac{(Y - Y_{1-\alpha})[\alpha Q_\alpha - (Q_\alpha - y_\ell)\mathbb{1}(y_\ell \leq Q_\alpha)]}{Y_\alpha^2}.$$

Similarly, to compute a linearized variable for SR, the result from Section 3.8 is used:

$$v_\ell^{\widetilde{SR}} = \frac{y_\ell - \{y_\ell H[(1-\alpha)N - \ell + 1] + \widetilde{Q}_{1-\alpha}[(1-\alpha) - \mathbb{1}(y_\ell < \widetilde{Q}_{1-\alpha})]\}}{\widetilde{Y}_\alpha} - \frac{(Y - \widetilde{Y}_{1-\alpha})\{y_\ell H(\alpha N - \ell + 1) + \widetilde{Q}_\alpha[\alpha - \mathbb{1}(y_\ell < \widetilde{Q}_\alpha)]\}}{\widetilde{Y}_\alpha^2}.$$

## A.7 Linearization of the ASR

The expression of the ASR in (2.14) when $N$ is even is

$$\text{ASR} = \frac{N}{2}\left[\sum_{\ell=1}^{N/2} \frac{\sum_{k\in U} y_k \mathbb{1}(k \leq \ell)}{\sum_{k\in U} y_k \mathbb{1}(k > N - \ell)}\right]^{-1}.$$

Now, if we define

$$Y_\ell = \sum_{k\in U} y_k \mathbb{1}(k \leq \ell),$$

and

$$R_\ell = \frac{Y_\ell}{Y - Y_{N-\ell}}, \tag{A.16}$$

the ASR can be rewritten by

$$\text{ASR} = \frac{N}{2\sum_{\ell=1}^{N/2} R_\ell}.$$

Using derivation rules and the influence function approach of Deville (1999), a linearized variable for ASR is thus

$$\begin{aligned} v_k^{ASR} &= -\frac{N}{2\left(\sum_{\ell=1}^{N/2} R_\ell\right)^2} \sum_{\ell=1}^{N/2} I(R_\ell)_k \\ &= -\text{ASR}\frac{\sum_{\ell=1}^{N/2} I(R_\ell)_k}{\sum_{\ell=1}^{N/2} R_\ell}, \end{aligned} \tag{A.17}$$

where $I(R_\ell)_k$ denotes the influence function of $R_\ell$ for unit $k$. Recalling (A.16) we can now write

$$I(R_\ell)_k = \frac{I(Y_\ell)_k}{Y - Y_{N-\ell}} - \frac{Y_\ell I(Y - Y_{N-\ell})_k}{(Y - Y_{N-\ell})^2}.$$

Applying Results 5.1 and 5.2 to $I(Y - Y_{N-\ell})_k$, we obtain

$$I(Y - Y_{N-\ell})_k = I(Y)_k - I(Y_{N-\ell})_k = y_k - I(Y_{N-\ell})_k,$$

and

$$I(R_\ell)_k = \frac{I(Y_\ell)_k}{Y - Y_{N-\ell}} - \frac{Y_\ell[y_k - I(Y_{N-\ell})_k]}{(Y - Y_{N-\ell})^2}. \tag{A.18}$$

We can now use Expression (3.19) with $\alpha = \ell/N$ and respectively with $\alpha = (N-\ell)/N$ to derive the influence function of $Y_\ell$ and $Y_{N-\ell}$:

$$I(Y_\ell)_k = \frac{\ell}{N} y_\ell - (y_\ell - y_k)\mathbb{1}(y_k \leq y_\ell), \tag{A.19}$$

$$I(Y_{N-\ell})_k = \frac{N-\ell}{N} y_{N-\ell} - (y_{N-\ell} - y_k)\mathbb{1}(y_k \leq y_{N-\ell}). \tag{A.20}$$

Finally replacing (A.19) and (A.20) into (A.18), we obtain

$$\begin{aligned} I(R_\ell)_k \quad = \quad & \frac{\frac{\ell}{N} y_\ell - (y_\ell - y_k)\mathbb{1}(y_k \leq y_\ell)}{Y - Y_{N-\ell}} \\ & - \frac{Y_\ell \left[ y_k - \frac{N-\ell}{N} y_{N-\ell} + (y_{N-\ell} - y_k)\mathbb{1}(y_k \leq y_{N-\ell}) \right]}{(Y - Y_{N-\ell})^2}, \end{aligned} \tag{A.21}$$

which can be substituted into (A.17) and obtain a linearized variable for ASR. Note that derivations for the case where $N$ is odd follows directly. Expression (A.17) becomes

$$v_k^{ASR} = -\frac{N+1}{2\left(\sum_{\ell=1}^{(N+1)/2} R_\ell\right)^2} \sum_{\ell=1}^{(N+1)/2} I(R_\ell)_k,$$

and $I(R_\ell)_k$ in (A.21) is unchanged.

# Bibliography


Acar, W. & Sankaran, K. (1999). The myth of the unique decomposability: Specializing the Herfindahl and entropy measures? *Strategic Management Journal* **20**, 969–975. (Cited page 37.)

Alfons, A., Holzer, J. & Templ, M. (2010). *laeken: Laeken indicators for measuring social cohesion*. R package version 0.1.3. (Cited page 117.)

Allison, P. D. (1978). Measures of inequality. *American Sociological Review* **43**, 865–880. (Cited page 36.)

Anand, S. (1983). *Inequality and poverty in Malaysia: measurement and decomposition*. World Bank research publication. New York: Oxford University Press. (Cited page 101.)

Ardilly, P. & Tillé, Y. (2006). *Sampling Methods: Exercises and Solutions*. New York: Springer. (Cited page 90.)

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory* **2**, 244–263. (Cited page 36.)

Atkinson, A. B. & Bourguignon, F., eds. (2000). *Handbook of Income Distribution*, vol. 1. Amsterdam: Elsevier. (Cited page 33.)

Atkinson, A. B., Marlier, E. & Nolan, B. (2004). Indicators and targets for social inclusion in the European Union. *Journal of Common Market Studies* **42**, 47–75. (Cited page 20.)

Barrett, G. F. & Donald, S. G. (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *Journal of Business and Economic Statistics* **27**, 1–17. (Cited pages 52, 100, and 101.)

Beach, C. M. & Davidson, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies* **50**, 723–735. (Cited page 90.)

Berger, Y. G. (2008). A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics* **24**, 541–555. (Cited pages 35, 52, 79, 81, 100, 102, and 104.)

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics* **137**, 674–707. (Cited page 100.)

Bhattacharya, N. & Mahalanobis, B. (1967). Regional disparities in household consumption in India. *Journal of the American Statistical Association* **62**, 143–161. (Cited page 43.)

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology* **22**, 17–22. (Cited pages 80, 115, and 116.)

Binder, D. A. & Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equation approach. *Survey Methodology* **21**, 137–145. (Cited page 95.)

Binder, D. A. & Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association* **89**, 1035–1043. (Cited pages 56, 80, and 95.)

Bishop, J. A., Formby, J. & Zheng, B. (1997). Statistical inference and the Sen index of poverty. *International Economic Review* **38**, 381–387. (Cited page 101.)

Blackorby, C., Donaldson, D. & Auersperg, M. (1981). A new procedure for the measurement of inequality within and among population subgroups. *Canadian Journal of Economics* **14**, 665–685. (Cited page 43.)

Bosmans, K. & Cowell, F. A. (2010). The class of absolute decomposable inequality measures. *Economics Letters* **109**, 154–156. (Cited page 42.)

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica* **47**, 901–920. (Cited page 43.)

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. (With discussion). *Journal of the Royal Statistical Society. Series B. Methodological* **26**, 211–252. (Cited page 61.)

Brewer, K. R. W. (1999). Design-based or prediction-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review* **67**, 35–47. (Cited page 26.)

Ceriani, L. & Verme, P. (2011). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality* , 1–2310.1007/s10888-011-9188-x. (Cited page 87.)

Chakravarty, S. R. (1999). Measuring inequality: the axiomatic approach. In *Handbook of Income Inequality Measurement*, J. Silber, ed. Norwell: Kluwer Academic Publishers. (Cited page 42.)

Chotikapanich, D., ed. (2008). *Modeling Income Distributions and Lorenz Curves*. New York: Springer. (Cited page 26.)

Chotikapanich, D. & Griffiths, W. E. (2006). Bayesian assessment of Lorenz and stochastic dominance in income distributions. In *Dynamics of inequality and poverty*, J. Creedy & G. Kalb, eds. Amsterdam: Elsevier. (Cited page 26.)

Coulter, P. B. (1989). *Measuring Inequality: A Methodological Handbook*. Boulder: Westview Press. (Cited pages 33, 34, and 37.)

Cowell, F. A. (1977). *Measuring Inequality*. Oxford: Philip Allan. (Cited pages 33, 53, and 111.)

Cowell, F. A. (1988). Inequality decomposition: Three bad measures. *Bulletin of Economic Research* **40**, 309–311. (Cited pages 40 and 76.)

Cowell, F. A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics* **42**, 27–41. (Cited pages 90 and 101.)

Cowell, F. A. (2000). Measurement of inequality. In *Handbook of Income Distribution*, A. B. Atkinson & F. Bourguignon, eds., vol. 1. Amsterdam: Elsevier. (Cited pages 24, 35, 40, and 41.)

Cowell, F. A., ed. (2003). *The economics of poverty and inequality*. Cheltenham: Edward Elgar. (Cited page 33.)

Cowell, F. A. (2006). Theil, inequality indices and decomposition. *Research on Economic Inequality* **13**, 345–360. (Cited pages 37 and 40.)

Cowell, F. A. & Flachaire, E. (2007). Income distribution and inequality measurement: the problem of extreme values. *Journal of Econometrics* **141**, 1044–1072. (Cited page 121.)

Cowell, F. A. & Kuga, K. (1981a). Additivity and the entropy concept: An axiomatic approach to inequality measurement. *Journal of Economic Theory* **25**, 131–143. (Cited page 43.)

Cowell, F. A. & Kuga, K. (1981b). Inequality measurement: An axiomatic approach. *European Economic Review* **15**, 287–305. (Cited pages 40 and 119.)

Cowell, F. A. & Victoria-Feser, M.-P. (1996a). Poverty measurement with contaminated data: A robust approach. *European Economic Review* **40**, 1761–1771. (Cited page 120.)

Cowell, F. A. & Victoria-Feser, M.-P. (1996b). Robustness properties of inequality measures. *Econometrica* **64**, 77–101. (Cited pages 95 and 101.)

Cowell, F. A. & Victoria-Feser, M.-P. (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* **1**, 191–219. (Cited pages 95, 100, and 120.)

Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio. *Empirical Economics* **22**, 515–531. (Cited pages 43 and 119.)

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal* **30**, 348–361. (Cited pages 40, 41, and 76.)

Das, T. & Parikh, A. (1981). Decompositions of atkinson's measure of inequality. *Australian Economic Papers* **20**, 171–78. (Cited page 43.)

Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics* **150**, 30–40. (Cited pages 86, 89, 100, 101, 104, 105, and 106.)

Dayioglu, M. & Baslevent, C. (2006). Imputed rents and regional income inequality in Turkey: A subgroup decomposition of the Atkinson index. *Regional Studies* **40**, 889–905. (Cited page 43.)

Dell, F., d'Haultfoeuille, X., Février, P. & Massé, E. (2002). Mise en oeuvre du calcul de variance par linéarisation. In *Actes des Journées de Méthodologie Statistique*. Paris: Insee-Méthodes. (Cited pages 58, 59, 79, 81, 100, 134, and 135.)

Demnati, A. & Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology* **30**, 17–34. (Cited pages 56, 80, 91, 99, 115, 116, and 125.)

Deville, J.-C. (1988). Estimation linéaire et redressement sur informations auxiliaires d'enquêtes par sondage. In *Mélanges économiques: Essais en l'honneur de Edmond Malinvaud*, A. Monfort & J.-J. Laffond, eds. Paris: Economica, pp. 915–927. (Cited page 75.)

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*. Örebro: Statistics Sweden. (Cited page 75.)

Deville, J.-C. (1996). Estimation de la variance du coefficient de Gini estimé par sondage. In *Actes des Journées de Méthodologie Statistique*, vol. 69-70-71. Paris: Insee-Méthodes. (Cited pages 52, 79, 92, 95, and 103.)

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204. (Cited pages 52, 56, 57, 58, 59, 80, 93, 95, 96, 97, 100, 101, 103, 115, 116, 120, 125, 127, 133, 136, and 137.)

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In *Actes des Journées de Méthodologie Statistique*. Paris: Insee-Méthodes. (Cited page 127.)

Deville, J.-C., Grosbras, J.-M. & Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceedings in Computational Statistics*. Heidelberg: Physica Verlag. (Cited page 75.)

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382. (Cited pages 54, 88, 93, and 116.)

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893–912. (Cited pages 75 and 81.)

Deville, J.-C. & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* **128**, 569–591. (Cited pages 75 and 81.)

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. (Cited pages 34, 42, 52, and 76.)

Fei, J. C. H., Ranis, G. & Kuo, S. W. Y. (1978). Growth and the family distribution of income by factor components. *The Quarterly Journal of Economics* **92**, 17–53. (Cited page 87.)

Fienberg, S. E. & Tanur, J. M. (1995). Reconsidering Neyman on experimentation and sampling: Controversies and fundamental contributions. *Probability and Mathematical Statistics* **15**, 47–60. (Cited page 68.)

Fienberg, S. E. & Tanur, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* **64**, 237–253. (Cited page 68.)

Foster, J. E. & Shorrocks, A. F. (1991). Subgroup consistent poverty indices. *Econometrica* **59**, 687–709. (Cited page 42.)

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika* **96**, 933–944. (Cited page 75.)

Galvani, L. (1951). Révision critique de certains points de la méthode représentative. *Revue de l'Institut International de Statistique* **19**, 1–12. (Cited pages 71 and 74.)

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* **54**, 306–316. (Cited pages 53, 90, and 111.)

Gibbs, A. L. & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70**, 419–435. (Cited page 37.)

Giles, D. E. A. (2004). Calculating a standard error for the Gini coefficient: some further results. *Oxford Bulletin of Economics and Statistics* **66**, 425–433. (Cited pages 79, 86, 103, 104, 105, 106, and 107.)

Giles, D. E. A. (2006). A cautionary note on estimating the standard error of the Gini index of inequality: comment. *Oxford Bulletin of Economics and Statistics* **68**, 395–396. (Cited page 103.)

Gini, C. (1912). *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppin. (Cited pages 76 and 87.)

Gini, C. (1914). Sulla misura della concertrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Scienze, Lettere e Arti* **73**, 1203–1248. (Cited pages 76 and 87.)

Gini, C. (1921). Measurement of inequality and incomes. *The Economic Journal* **31**, 124–126. (Cited pages 35, 76, and 87.)

Gini, C. & Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione 1. dic. 1921. *Annali di Statistica* **Series 6, 4**, 1–107. (Cited pages 71, 73, 74, and 84.)

Giorgi, G. M., Palmitesta, P. & Provasi, C. (2006). Asymptotic and bootstrap inference for the generalized Gini indices. *Metron* **64**, 107–124. (Cited page 91.)

Glasser, G. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association* **57**, 648–654. (Cited pages 79, 90, 93, 94, 98, and 100.)

Goga, C., Deville, J.-C. & Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika* **96**, 691–709. (Cited page 127.)

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn & A. Buccianti, eds. Chichester: Wiley. (Cited page 99.)

Graf, M., Nedyalkova, D., Münnich, R., Seger, J. & Zins, S. (2011). Parametric estimation of income distributions and indicators of poverty and social exclusion. Research Project Report WP2 – D2.1, FP7-SSH-2007-217322 AMELI. (Cited pages 26 and 127.)

Greselin, F., Pasquazzi, L. & Zitikis, R. (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**. (Cited pages 111, 115, and 120.)

Greselin, F., Puri, M. & Zitikis, R. (2009). L-functions, processes, and statistics in measuring economic inequality and actuarial risks. *Statistics and its Interfaces* **2**, 227–245. (Cited page 111.)

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis pro Pestování Matematiky* **84**, 387–423. (Cited page 70.)

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker. (Cited page 70.)

Halmos, P. R. (1946). The theory of unbiased estimation. *Annals of Mathematical Statistics* **17**, 34–43. (Cited page 90.)

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393. (Cited pages 95, 96, 115, and 120.)

Hampel, F. R., Ronchetti, E., Rousseuw, P. J. & Stahel, W. (1985). *Robust Statistics: The Approach Based on the Influence Function*. New-York: Wiley. (Cited pages 56, 80, 95, and 115.)

Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science* **2**, 180–190. (Cited page 68.)

Hansen, M. H., Dalenius, T. D. & Tepping, B. J. (1985). The development of sample survey in finite population. In *A Celebration of Statistics, The ISI Centenary Volume*, A. B. Atkinson & S. Fienberg, eds. New York: Springer. (Cited page 68.)

Hansen, M. H., Madow, W. G. & Tepping, B. J. (1983). An evaluation of model dependent and probability-sampling inferences in sample surveys (with discussion and rejoinder). *Journal of the American Statistical Association* **78**, 776–807. (Cited page 26.)

Hedayat, A. S. & Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference* **44**, 237–247. (Cited page 75.)

Heston, A., Summers, R. & Aten, B. (1995). Penn World Table. Tech. rep., Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania. (Cited page 104.)

Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review* **54**, 761. (Cited page 37.)

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325. (Cited pages 79, 90, 100, and 101.)

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685. (Cited pages 29 and 87.)

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B. Methodological* **52**, 105–124. (Cited page 90.)

Hulliger, B., Alfons, A., Filzmoser, P., Meraner, A., Schoch, T. & Templ, M. (2011). Robust methodology for Laeken indicators. Research Project Report WP4 – D4.2, FP7-SSH-2007-217322 AMELI. (Cited page 127.)

Hulliger, B. & Münnich, R. (2006). Variance estimation for complex surveys in the presence of outliers. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association. (Cited pages 65 and 76.)

Hulliger, B. & Schoch, T. (2009). Robustification of the Quintile Share Ratio. In *Proceedings of the Colloque sur les méthodes de sondage en l'honneur de Jean-Claude Deville*. Neuchâtel: University of Neuchâtel. (Cited page 120.)

Hyndman, R. J. & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician* **50**, 361–365. (Cited pages 35, 54, 58, 77, and 88.)

Isaki, C. T. & Fuller, W. A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association* **77**, 89–96. (Cited pages 93 and 116.)

Jaeckel, L. (1972). The infinitesimal jackknife. Tech. rep., Bell Laboratories Memorandum, MM 72-1215-11. (Cited page 90.)

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute* **22**, 359–380. (Cited pages 68 and 69.)

Karagiannis, E. & Kovacevic, M. (2000). A method to calculate the jackknife variance estimator for the Gini coefficient. *Oxford Bulletin of Economics and Statistics* **62**, 119–122. (Cited page 103.)

Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique* **9**, 176–183. (Cited pages 68 and 69.)

Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique* **11**, 180–185. (Cited pages 68 and 69.)

Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique* **13**, 66–78. (Cited pages 68 and 69.)

Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique* **14**, 119–134. (Cited pages 68 and 69.)

Kish, L. (1995). *Survey Sampling*. New York: Wiley-Interscience. (Cited page 30.)

Kleiber, C. & Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken: Wiley-Interscience. (Cited page 26.)

Kolm, S.-C. (1976a). Unequal inequalities. I. *Journal of Economic Theory* **12**, 416–442. (Cited page 42.)

Kolm, S.-C. (1976b). Unequal inequalities. II. *Journal of Economic Theory* **13**, 82–111. (Cited page 42.)

Kolm, S. C. (1999). The rational foundations of income inequality measurement. In *Handbook of Income Inequality Measurement*, J. Silber, ed. Norwell: Kluwer Academic Publishers. (Cited page 42.)

Kovacevic, M. S. & Binder, D. A. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics* **13**, 41–58. (Cited pages 53, 56, 80, 95, 100, 101, 111, and 116.)

Kruskal, W. & Mosteller, F. (1979a). Representative sampling, I: Nonscientific literature. *International Statistical Review* **47**, 13–24. (Cited pages 70, 71, and 72.)

Kruskal, W. & Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review* **47**, 111–127. (Cited pages 70, 71, and 72.)

Kruskal, W. & Mosteller, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review* **47**, 245–265. (Cited pages 70, 71, and 72.)

Kruskal, W. & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review* **48**, 169–195. (Cited pages 70, 71, and 72.)

Kuan, X. (2000). Inference for generalized Gini indices using the iterated bootstrap method. *Journal of Business and Economics Statistics* **18**, 223–227. (Cited page 91.)

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The annals of Mathematical Statistics* **22**, 79–86. (Cited pages 37 and 38.)

Lambert, P. J., Millimet, D. L. & Slottje, D. (2003). Inequality aversion and the natural rate of subjective inequality. *Journal of Public Economics* **87**, 1061–1090. (Cited page 36.)

Langel, M. & Tillé, Y. (2011a). Corrado Gini, a pioneer in balanced sampling and inequality theory. *Metron* **69**, 45–65. (Cited pages 21 and 119.)

Langel, M. & Tillé, Y. (2011b). An evaluation of the performance of inequality measures for the detection of changes in an income distribution. Tech. rep., University of Neuchatel. (Cited pages 53 and 120.)

Langel, M. & Tillé, Y. (2011c). Inference by linearization for Zenga's new inequality index: a comparison with the Gini index. *Metrika*, doi:10.1007/s00184-011-0369-1. (Cited page 21.)

Langel, M. & Tillé, Y. (2011d). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference* **141**, 2976–2985. (Cited pages 21, 76, 88, 111, and 112.)

Leiten, E. (2005). On variance estimation of the Gini coefficient. In *Workshop on Survey Sampling Theory and Methodology*. Vilnius: Statistics Lithuania. (Cited page 79.)

Leiten, E. & Traat, I. (2005). Variance of Laeken indicators in complex surveys. Tech. rep., Statistical Office of Estonia. (Cited page 39.)

Lerman, R. I. & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economic Letters* **15**, 363–368. (Cited page 101.)

Lesage, E. (2008). Contraintes d'équilibrage non linéraires. In *Méthodes de sondage : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, P. Guilbert, D. Haziza, A. Ruiz-Gazen & Y. Tillé, eds. Paris: Dunod, pp. 285–289. (Cited pages 81 and 82.)

Lesage, E. (2009). Calage non linéaire. In *Actes des Xème Journées de Méthodologie Statistique*. Paris: Insee. (Cited pages 100 and 127.)

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics* **22**, 1081–1114. (Cited page 37.)

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**, 546–556. (Cited page 26.)

Lomnicki, Z. A. (1952). The standard error of Gini's mean difference. *Annals of Mathematical Statistics* **23**, 635–637. (Cited page 90.)

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* **9**, 209–219. (Cited pages 24, 53, 76, 87, and 111.)

Maffenini, W. & Polisicchio, M. (2010). How potential is the I(p) inequality curve in the analysis of empirical distributions. Tech. rep., Universita degli Studi di Milano-Bicocca. (Cited pages 111 and 120.)

McDonald, J. B. (1984). Some generalised functions for the size distribution of income. *Econometrica* **52**, 647–663. (Cited page 127.)

Merle, P. & Andreyev, Z. (2002). Democratization or increase in educational inequality? changes in the length of studies in France, 1988-1998. *Population* **57**, 631–657. (Cited page 40.)

Mills, J. A. & Zandvakili, S. (1997). Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* **12**, 133–150. (Cited page 91.)

Modarres, R. & Gastwirth, J. L. (2006). A cautionary note on estimating the standard error of the Gini index of inequality. *Oxford Bulletin of Economics and Statistics* **68**, 385–390. (Cited page 103.)

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica* **51**, 561–577. (Cited pages 80, 95, 98, 100, 101, 119, and 120.)

Moore, R. E. (1996). Ranking income distributions using the geometric mean and a related general measure. *Southern Economic Journal* **63**, 69–75. (Cited page 36.)

Münnich, R. & Zins, S. (2011). Variance estimation for indicators of poverty and social exclusion. Research Project Report WP3 – D3.2, FP7-SSH-2007-217322 AMELI. (Cited page 127.)

Mussard, S., Seyte, F. & Terraza, M. (2003). Decomposition of Gini and the generalized entropy inequality measures. *Economics Bulletin* **4**, 1–6. (Cited page 37.)

Nair, U. S. (1936). The standard error of Gini's mean difference. *Biometrika* **28**, 428–436. (Cited page 90.)

Nedyalkova, D. & Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika* **95**, 521–537. (Cited pages 26 and 75.)

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–606. (Cited pages 71 and 72.)

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Washington: Graduate School, U. S. Department of Agriculture. (Cited pages 73 and 74.)

Nygård, F. & Sandström, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International. (Cited page 93.)

Nygård, F. & Sandström, A. (1985). The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics* **1**, 399–412. (Cited pages 80 and 93.)

Nygård, F. & Sandström, A. (1989). Income inequality measures based on sample surveys. *Journal of Econometrics* **42**, 81–95. (Cited page 93.)

Ogwang, T. (2000). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics* **62**, 123–129. (Cited pages 79, 102, 103, 104, 105, and 106.)

Ogwang, T. (2004). Calculating a standard error for the Gini coefficient: some further results: reply. *Oxford Bulletin of Economics and Statistics* **66**, 435–437. (Cited page 103.)

Ogwang, T. (2006). A cautionary note on estimating the standard error of the Gini index of inequality: comment. *Oxford Bulletin of Economics and Statistics* **68**, 391–393. (Cited page 103.)

Osier, G. (2006). Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation. In *Methodological Workshop on EU-SILC*. Helsinki: Eurostat and Statistics Finland. (Cited pages 39, 51, 52, 56, 57, 58, 65, 81, and 100.)

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods* **3**, 167–195. (Cited pages 51, 52, 56, 57, 58, 65, 81, and 100.)

Peng, L. (2011). Empirical likelihood methods for the Gini index. *Australian and New Zealand Journal of Statistics* **53**, 131–139. (Cited page 91.)

Pigou, A. C. (1912). *Wealth and Welfare*. London: Mc Millan. (Cited pages 40, 41, and 76.)

Polisicchio, M. (2008). The continuous random variable with uniform point inequality measure I(p). *Statistica e Applicazioni* **6**, 137–151. (Cited page 111.)

Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficients. *Economic Journal* **86**, 243–55. (Cited page 43.)

Qin, Y., Rao, J. N. K. & Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling* **27**, 1429–1435. (Cited page 91.)

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. (Cited page 117.)

Radaelli, P. (2008). A subgroups decomposition of Zenga's uniformity and inequality indexes. *Statistica e Applicazioni* **6**, 117–136. (Cited page 111.)

Radaelli, P. (2010). On the decomposition by subgroups of the Gini index and Zenga's uniformity and inequality indexes. *International Statistical Review* **78**, 81–101. (Cited pages 43, 111, and 119.)

Rao, J. N. K. (2011). Impact of frequentist and bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science* **26**, 240–256. (Cited page 27.)

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley: University of California Press. (Cited page 38.)

Royall, R. M. (1988). The prediction approach to sampling theory. In *Sampling*, P. R. Krishnaiah & C. R. Rao, eds., vol. 6 of *Handbook of Statistics*. Amsterdam: Elsevier, pp. 399–413. (Cited page 26.)

Royall, R. M. & Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association* **68**, 880–889. (Cited pages 74 and 75.)

Sandström, A., Wretman, J. H. & Walden, B. (1985). Variance estimators of the Gini coefficient: Simple random sampling. *Metron* **43**, 41–70. (Cited pages 79, 80, 92, 93, 94, 95, 102, and 103.)

Sandström, A., Wretman, J. H. & Walden, B. (1988). Variance estimators of the Gini coefficient: Probability sampling. *Journal of Business and Economic Statistics* **6**, 113–120. (Cited pages 80, 92, 93, 94, 95, 101, 102, and 103.)

Särndal, C.-E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: Wiley. (Cited pages 54 and 127.)

Särndal, C.-E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer. (Cited pages 26 and 30.)

Schechtman, E. (1991). On estimating the asymptotic variance of a function of U-statistics. *The American Statistician* **45**, 103–106. (Cited page 90.)

Sen, A. K. (1973). *On Economic Inequality*. Oxford: Clarendon Press. (Cited pages 34, 40, and 87.)

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127. (Cited page 29.)

Sendler, W. (1979). On statistical inference in concentration measurement. *Metrika* **26**, 109–122. (Cited pages 90 and 94.)

Shalit, H. (1985). Calculating the Gini index of inequality for individual data. *Oxford Bulletin of Economics and Statistics* **47**, 185–189. (Cited pages 35 and 101.)

Shorack, G. (1972). Functions of order statistics. *Annals of Mathematical Statistics* **43**, 412–427. (Cited page 90.)

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica* **48**, 613–625. (Cited pages 37, 40, 43, and 76.)

Shorrocks, A. F. (1984). Inequality decomposition by population subgroups. *Econometrica* **52**, 1369–1385. (Cited pages 40 and 76.)

Silber, J., ed. (1999). *Handbook of Income Inequality Measurement*. Norwell: Kluwer Academic Publishers. (Cited pages 33 and 34.)

Sillitto, G. P. (1969). Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika* **56**, 641–650. (Cited page 90.)

Skinner, C. J. (2004). Comment on Demnati and Rao: Linearization variance estimators for survey data. *Survey Methodology* **30**, 17–18. (Cited page 56.)

Stoetzel, J. (1980). Le cours de la vie selon la condition sociale: une étude des revenus selon l'âge dans les diverses professions. *Revue Française de Sociologie* **21**, 155–170. (Cited page 40.)

Swiss Federal Statistical Office (2008). *Scénarios des ménages. Evolution des ménages privés entre 2005 et 2030*. OFS, Neuchâtel. (Cited page 31.)

Theil, H. (1967). *Economics and Information Theory*. Amsterdam: North-Holland. (Cited pages 37 and 76.)

Theil, H. (1969). The desired political entropy. *The American Political Science Review* **63**, 521–525. (Cited page 76.)

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Paris: Dunod. (Cited page 30.)

Tillé, Y. & Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika* **91**, 913–927. (Cited page 75.)

Tillé, Y. & Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters* **74**, 31–37. (Cited page 75.)

Traat, I. (2006). Variance of quantile estimators in household surveys. In *Workshop on Survey Sampling Theory and Methodology*. Ventspils: Central Statistical Bureau of Latvia. (Cited page 52.)

van den Brakel, J. & Bethlehem, J. (2008). Model-based estimation for official statistics. Discussionpaper 08002, Statistics Netherlands. (Cited page 27.)

Wolter, K. M. (2007). *Introduction to Variance Estimation*. New York: Springer, 2nd ed. (Cited pages 28, 29, and 56.)

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411–414. (Cited pages 56, 79, 93, 115, and 116.)

Xu, K. (2004). How has the literature on Gini's index evolved in the past 80 years? Working papers archive, Dalhousie, Department of Economics. (Cited pages 35, 76, 87, and 91.)

Xu, K. (2007). U-statistics and their asymptotic results for some inequality and poverty measures. *Econometric Reviews* **26**, 567–577. (Cited pages 91 and 101.)

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. London: Charles Griffin, 3rd ed. (Cited pages 73, 74, and 75.)

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B. Methodological* **15**, 235–261. (Cited page 29.)

Yitzhaki, S. (1994). Economic distance and overlapping of distribution. *Journal of Economics* **61**, 147–159. (Cited page 43.)

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling Gini. *Research in Economic Inequality* **8**, 13–30. (Cited page 43.)

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. (Cited pages 38 and 110.)

Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica e Applicazioni* **4**, 3–27. (Cited pages 38, 110, 111, and 119.)

Zheng, B. (1994). Can a poverty index be both relative and absolute? *Econometrica* **62**, 1453–1458. (Cited pages 40 and 42.)

Zheng, B. (2001). Statistical inference for poverty measures with relative poverty lines. *Journal of Econometrics* **101**, 337–356. (Cited page 101.)

Zitikis, R. (2003). Asymptotic estimation of the E-Gini index. *Econometric Theory* **19**, 587–601. (Cited page 101.)

Zitikis, R. & Gastwirth, J. L. (2002). The asymptotic distribution of the S-Gini index. *Australian and New Zealand Journal of Statistics* **44**, 439–446. (Cited page 101.)