Statistics and Its Interface Volume 2 (2009) 227–245

L-functions, processes, and statistics in measuring economic inequality and actuarial risks

Francesca Greselin, Madan L. Puri and Ričardas Zitikis*†

To Professor Joseph L. Gastwirth whose creativity and fondness for indices have been matchlessly inspiring

L-statistics play prominent roles in various research areas and applications, including development of robust statistical methods, measuring economic inequality and insurance risks. In many applications the score functions of L-statistics depend on parameters (e.g., distortion parameter in insurance, risk aversion parameter in econometrics), which turn the L-statistics into functions that we call L-functions. A simple example of an L-function is the Lorenz curve. Ratios of L-functions play equally important roles, with the Zenga curve being a prominent example. To illustrate real life uses of these functions/curves, we analyze a data set from the Bank of Italy year 2006 sample survey on household budgets. Naturally, empirical counterparts of the population L-functions need to be employed and, importantly, adjusted and modified in order to meaningfully capture situations well beyond those based on simple random sampling designs. In the processes of our investigations, we also introduce the L-process on which statistical inferential results about the population L-function hinges. Hence, we provide notes and references facilitating ways for deriving asymptotic properties of the L-process.

AMS 2000 subject classifications: Primary 62P05, 62P20, 62P25; secondary 62G05, 62G20, 62G30.
Keywords and phrases: Gini index, Zenga index, Lorenz curve, Zenga curve, L-statistic, L-function, L-process, Vervaat process, economic inequality, risk measure.

1. INTRODUCTION

Linear combinations of order statistics, commonly known as L-statistics, and their numerous variations and generalizations have played notable roles in diverse areas of application such as measuring economic inequality and insurance risks, deriving premium calculation principles.

Guided by econometric and actuarial applications, and also by a mathematical point of view, in this paper we analyze L-statistics and arrive at their extensions and generalizations that satisfy interesting properties and open up venues for further applications and research in the area. To illustrate the richness and beauty of L-statistics, and to also provide a natural background for the introduction of what we call the L-function in Section 2, we first have a closer look at several examples.

Suppose we are interested in the distribution of incomes in a society. We randomly select n individuals from the society and record their incomes, which are non-negative real numbers assuming that we do not take into account their debts. Hence, we are dealing with outcomes of n non-negative random variables (rv's)

$$X_1, X_2, \dots, X_n. \tag{1}$$

Obviously, the arithmetic mean $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ can hardly convey much useful information, particularly in view of the fact that statistical populations in this context are usually skewed. On the other hand, the median or, more generally, percentiles are more informative measures. They can be represented using the order statistics

$$X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n} \tag{2}$$

of rv's (1). From the mathematical point of view, by ordering rv's (1) we have obtained a *monotone* sequence and thus extracted additional information from rv's (1). We may go one step further and extract *convexity* from the rv's by considering the sequence of lower partial sums $n^{-1}\sum_{i=1}^k X_{i:n}$, $1 \le k \le n$. This naturally takes us back to the work of Lorenz [57] from where the so-called Lorenz curve originates. We shall come back to this and other related curves in later sections.

Similarly to the above noted econometric context, actuaries frequently measure insurance risks, calculate premiums and thus, among many other things, deal with loss rv's $X_1, X_2, \dots, X_n$. They are, in many situations, non-negative rv's, just like the earlier noted incomes. However, and contrary to the econometric context, the focus of actuaries is on large losses, and thus on large order statistics of the X's. This point of view naturally leads to the sequence of upper partial sums $n^{-1}\sum_{i=k}^n X_{i:n}$, $1 \le k \le n$. Naturally, there is a duality between the lower and upper partial sums, expressed in the form of the equation

*Corresponding author.
†Research supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

$n^{-1}\sum_{i=1}^{k} X_{i:n} = \bar{X} - n^{-1}\sum_{i=k+1}^{n} X_{i:n}$, and we shall encounter this duality in various forms later in this paper.

The lower and upper partial sums of order statistics (2) are of course special cases of the L-statistic

$$\frac{1}{n}\sum_{i=1}^{n} c_{i,n} X_{i:n}, \tag{3}$$

where the coefficients $c_{i,n}$ are chosen by the researcher, or decision maker, depending on the problem at hand. We refer to Chernoff, Gastwirth and Johns [13] for classical examples and asymptotic results related to L-statistic (3). For recent contributions in the area with heavy-tailed population distributions, which play an important role in modeling actuarial risks, we refer to Necir and Boukhetala [62], Necir, Meraghni and Meddi [63], Necir and Meraghni [64]. Later in the present paper, we shall discuss additional examples and variations of the L-statistic. At this moment we only note that L-statistic (3) accommodates numerous indices of economic inequality as well as risk measures of actuarial science, especially those that are related to the distorted expectation theory or, in other words, Yaari's [93] dual theory of choice under risk (see, e.g., Kaas, Goovaerts, Dhaene and Denuit [53]; Denuit, Dhaene, Goovaerts and Kaas [27]).

When accommodating risk measures under the classical utility theory, we encounter averages $n^{-1}\sum_{i=1}^{n} h(X_{i:n})$, where h is a utility function. This naturally leads to the more general statistic

$$\frac{1}{n}\sum_{i=1}^{n} c_{i,n} h(X_{i:n}), \tag{4}$$

where the coefficients $c_{i,n}$ and the function h are usually chosen by the researcher, or decision maker. L-statistic (4) encompasses risk measures in the classical utility theory and also a number of those related to Yaari's dual theory of choice under risk (see Section 2.6 in Denuit, Dhaene, Goovaerts and Kaas [27] for notes and references on the topic). L-statistic (4) has also frequently and naturally appeared in the theory of statistics as discussed in the books by, e.g., Serfling [84], Shorack [87], Shao [86].

We may naturally wish to add to the class of L-statistics other ones such as the Cramér–von Mises, Anderson-Darling, and Watson statistics. This leads to the following sum of functions of order statistics

$$\frac{1}{n}\sum_{i=1}^{n} h_{i,n}(X_{i:n}), \tag{5}$$

which is called the FL-statistic in Zitikis [96, 97], where differentiability of the cumulative distribution function (cdf) of the statistic is investigated in the context asymptotic expansions (see, e.g., Helmers [44] for a background). As we have already hinted at above, the FL-statistic encompasses a very large class of statistics (see Zitikis, [96, 97]). For a far-reaching analysis of the FL-statistic, we refer to Borisov and Baklanov [7], Baklanov and Borisov [8, 3], where the authors call statistic (5) the generalized L-statistics.

The generalized L-statistic in the above noted papers by Baklanov and Borisov should not, however, be confused with the generalized L-statistic, which is also known as the GL-statistic, that has been extensively studied by Serfling [83], Helmers, Janssen and Serfling [46, 47], Helmers and Ruymgaart [45], Gilat and Helmers [35], Putt and Chinchilli [74], Serfling [85]; see also references therein. The GL-statistic is of the form

$$\frac{1}{m(n)}\sum_{i=1}^{m(n)} c_{i,m(n)} W_{i:m(n)}, \tag{6}$$

where the order statistics $W_{1:m(n)}, W_{2:m(n)}, \ldots, W_{m(n):m(n)}$ are not those of the original sample but of rv's $W_1, W_2, \ldots, W_{m(n)}$ that are generated from the original sample $X_1, X_2, \ldots, X_n$ using a specially designed mechanism. We refer to Serfling [83] for the definition of this mechanism, as well as for examples illustrating the encompassing nature of the GL-statistic, which includes L-, U- and many other statistics.

Actuarial and econometric problems related to portfolio theory have suggested other generalizations of L-statistics. One of such generalizations is the nested L-statistic, called also the NL-statistic, introduced by Brazauskas, Jones, Puri and Zitikis [10]. In a way, the NL-statistic is related to the GL-statistic, since it is of a similar (6) form:

$$\frac{1}{m}\sum_{i=1}^{m} c_{i,m} R_{i:m}, \tag{7}$$

where $R_{1:m} \le R_{2:m} \le \cdots \le R_{m:m}$ are the order statistics of R_i, $1 \le i \le m$, which are L-statistics (3) based on $\{X_1(i), X_2(i), \ldots, X_{n_i}(i)\}$, $1 \le i \le m$.

Another generalization of L-statistic (4) concerns with the situation when the rv's $X_1, X_2, \ldots, X_n$ are ordered according not to their own values, which would lead to the usual order statistics $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$, but according to the ordered values of other rv's, say $Y_1, Y_2, \ldots, Y_n$. Namely, suppose that we have n paired rv's $(X_1, Y_1), \ldots, (X_n, Y_n)$. We order the pairs so that the resulting n pairs have non-decreasing second coordinates $Y_{1:n} \le Y_{2:n} \le \cdots \le Y_{n:n}$. Denote the corresponding first coordinates by

$$X_{(1:n)}, X_{(2:n)}, \ldots, X_{(n:n)}, \tag{8}$$

which are called induced order statistics (see Bhattacharya [6]) or concomitants of order statistics (see, David [16] and [17], David and Nagaraja [18], references therein). Certainly, when $Y_i = X_i$ for all $1 \le i \le n$, then the induced order statistics $X_{(i:n)}$, $1 \le i \le n$, coincide with the corresponding usual order statistic $X_{i:n}$, $1 \le i \le n$. Having thus defined

the induced order statistics, we next generalize, for example, L-statistic (3) as follows:

$$\frac{1}{n}\sum_{i=1}^{n} c_{i,n} X_{(i:n)}. \tag{9}$$

Statistic (9) is frequently called the induced L-statistic. Large sample asymptotic properties of statistic (9) have been investigated by, for example, Bhattacharya [6], Seoh and Puri [82], Davydov and Egorov [21, 22], Rao and Zhao [77, 78]; see also references therein. For applications of the induced order statistics in econometrics and particularly in portfolio management, we refer to Schechtman, Shelef, Yitzhaki and Zitikis [80] as well as to the references therein.

We have by now seen a rich mosaic of L-statistics, but more applications, examples, and extensions will follow. We have organized the rest of the paper as follows. In Section 2 we introduce and justify the introduction of what we call the L-function, which encompasses numerous L-statistics and curves that have been used for measuring economic inequality and insurance risks. Section 3 devoted to a substantial application of L-functions (e.g., Lorenz and Zenga curves) in the analysis of a data set from the Bank of Italy year 2006 sample survey on household budgets. There we clearly see the importance of extending statistical inferential results from simple random sapling designs (i.e., the 'classical' i.i.d. assumption) to much more complex, and thus relevant in practice, sampling designs. In Section 4 we discuss indices of economic inequality whose definitions are based on the Lorenz curve, which is yet another example of the L-function. Since the L-function depends on the population cdf, which is unknown, in Section 5 we define an empirical L-function and, based on it, discuss methods for deriving statistical inferential results about the population L-function. Section 6 concludes the paper with a brief summary of main contributions.

2. L-FUNCTION

We can depict the lower and upper partial sums of order statistics (2) using the stepwise functions $n^{-1}\sum_{i=1}^{[tn]} X_{i:n}$ and $n^{-1}\sum_{i=[tn]+1}^{n} X_{i:n}$, respectively, defined for all $0 \le t \le 1$. The latter two sums can be written as the integrals $\int_0^{[tn]/n} F_n^{-1}(s)ds$ and $\int_{[tn]/n}^1 F_n^{-1}(s)ds$, respectively, where $F_n^{-1}(s)$ denotes the empirical quantile function $\inf\{x : F_n(x) \ge s\}$ corresponding to the empirical cdf $F_n(x) = n^{-1}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}$, where $\mathbf{1}$ is the indicator function. Replacing the limit of integration $[tn]/n$ in the two integrals by t, which modifies the two integrals by asymptotically (when $n \to \infty$) negligible terms, we turn the integrals into continuous functions

$$ALC_n(t) = \int_0^t F_n^{-1}(s)ds$$

and $DALC_n(t) = \int_t^1 F_n^{-1}(s)ds$, that are called, respectively, the empirical absolute Lorenz curve (ALC) and the dual of the ALC, which we simply denote by DALC. The duality between the two curves is expressed by the equation $ALC_n(t) = \bar{X} - DALC_n(t)$.

The corresponding population ALC and DALC are

$$ALC_F(t) = \int_0^t F^{-1}(s)ds$$

and $DALC_F(t) = \int_t^1 F^{-1}(s)ds$, where $F^{-1}(s) = \inf\{x : F(x) \ge s\}$ is the population quantile function. Obviously, $ALC_F(t) = \mu_F - DALC_F(t)$, where $\mu_F = \mathbf{E}[X]$.

In the actuarial literature (see, e.g., Denuit, Dhaene, Goovaerts and Kaas [27]), the quantile $F^{-1}(s)$ is frequently considered a risk measure, called value-at-risk and denoted by $VaR[X,t]$.

The curves $ALC_F(t)$ and $DALC_F(t)$ can be unified into one function $L_F : [0,1] \to [0,\infty]$ defined by

$$L_F(t) = \int_0^1 F^{-1}(s)K(s,t)ds,$$

which we call the L-function. The kernel $K : [0,1]\times[0,1] \to [0,\infty]$ is specified by the researcher depending on the problem at hand. To illustrate the L-function, note that when $K(s,t) = \mathbf{1}\{s \le t\}$, then $L_F(t) = ALC_F(t)$, and when $K(s,t) = \mathbf{1}\{s > t\}$, then $L_F(t) = DALC_F(t)$. Other examples of the L-function are conditional versions of the ordinary ALC and its dual counterpart defined, respectively, by the equations

$$ABC_F(t) = \frac{1}{t}\int_0^t F^{-1}(s)ds$$

and $DABC_F(t) = (1-t)^{-1}\int_t^1 F^{-1}(s)ds$, and called the absolute Bonferroni curve (ABC) and the dual of the ABC (i.e., DABC). These curves are L-functions with the kernels, respectively, $K(s,t) = t^{-1}\mathbf{1}\{s \le t\}$ and $K(s,t) = (1-t)^{-1}\mathbf{1}\{s > t\}$. Another example of the L-function is the proportional hazards transform (PHT), which is $L_F(t)$ with the kernel $K(s,t) = t/(1-s)^{1-t}$.

Note 1. Using mathematical terminology, the curve $ABC_F(t)$ is the Hardy transform of the quantile function $F^{-1}(t)$. In the actuarial literature (see, e.g., Denuit, Dhaene, Goovaerts and Kaas [27]), $ABC_F(t)$ is usually called the tail value-at-risk (TVaR) risk measure and denoted by $TVaR[X,t]$. The risk measure is closely related to the conditional tail expectation (CTE) risk measure $CTE[X,t] = \mathbf{E}[X|X > F^{-1}(t)]$. Indeed, when the cdf $F(x)$ is continuous, then the TVaR and CTE risk measures coincide.

Note 2. There are numerous examples when the kernel $K(s,t)$ does not depend on the parameter t. In such cases we

simply denote the kernel by $J(s)$. This turns the L-function $L_F(t)$ into the parameter

$$R_F = \int_0^1 F^{-1}(s)J(s)ds,$$

which is linked to L-statistic (3) in the following way. Namely, when the sample size n grows indefinitely, L-statistic (3) with the coefficients

$$c_{i,n} = n\int_{(i-1)/n}^{i/n} J(s)ds$$

converges (see, e.g., Serfling [84] and references therein) to the limit R_F. The function $J(s)$ is usually called the score, or weight, function. Several examples of the function $J(s)$ follow.

Example 1. When $J(s) \equiv 1$, then R_F is the population mean μ_F. The area beneath $ALC_F(t)$ but above the x-axis is R_F with $J(s) = 1 - s$. The area beneath $ABC_F(t)$ is R_F with $J(s) = \log(1/s)$. The absolute Gini index AG_F is twice the area between the absolute Lorenz curve $ALC_F(t)$ and the 'egalitarian' line $I_F(t) = t\mu_F$. Hence, AG_F is R_F with $J(s) = 2s - 1$; thus the equation

$$AG_F = \int_0^1 F^{-1}(s)(2s-1)ds. \tag{10}$$

Note 3. Viewing $L_F(t)$ and R_F as risk measures in an actuarial context, Jones and Zitikis [52] suggest, and justify, finding a 'distortion' parameter t such that the equation $L_F(t) = R_F$ holds. To solve the problem, Jones and Zitikis [52] specify conditions under which a unique solution, say t_0, to the equation exists, and then construct an empirical estimator t_n for t_0 as a solution to the equation $L_n(t) = R_n$, which is an empirical counterpart to the equation $L_F(t) = R_F$. Jones and Zitikis [52] also investigate the asymptotic distribution of t_n when the sample size n increases indefinitely.

For deeper understanding of the L-function $L_F(t)$, and to also get another point of view about its role in actuarial and econometric applications, we rewrite the function as follows:

$$L_F(t) = \int_0^\infty \kappa(1 - F(x), t)dx, \tag{11}$$

where the kernel $\kappa : [0,1] \times [0,1] \to [0,\infty]$ is defined by the equation

$$\kappa(v,t) = \int_0^v K(1-u,t)du.$$

(To prove equation (11), apply the definition of $\kappa(v,t)$ on the right-hand side of equation (11) and use the Fubini theorem.) Depending on the problem, the role of the kernel $\kappa(v,t)$ is to emphasize or, de-emphasize, the tail of the survival function $S(x) = 1 - F(x)$, thus making $L_F(t)$ larger or smaller than the mean μ_F.

Example 2. In the context of indices of economic inequality, the function $\kappa(v,t)$ frequently satisfies the property $\kappa(1,t) = 0$. To illustrate, we take the kernel $K(s,t) = 2s-1$, which leads to the absolute Gini index AG_F (see equation (10)). We have $\kappa(v,t) = v(1-v)$ and thus $\kappa(1,t) = 0$.

Example 3. In the context of actuarial risk measures, the function $\kappa(v,t)$ frequently satisfies the property $\kappa(1,t) = 1$. For example, consider the TVaR kernel $K(s,t) = (1-t)^{-1}\mathbf{1}\{s > t\}$, in which case we have that

$$\kappa(v,t) = \begin{cases} \dfrac{v}{1-t}, & \text{when } v < 1-t, \\ 1, & \text{when } v \geq 1-t. \end{cases}$$

For the PHT kernel $K(s,t) = t/(1-s)^{1-t}$, we have that $\kappa(v,t) = v^t$. In the two cases we obviously have $\kappa(1,t) = 1$.

Note 4. Since $K(s,t)$ is non-negative, the function $v \mapsto \kappa(v,t)$ is non-decreasing. This property together with $\kappa(0,t) = 0$ and, when it holds, $\kappa(1,t) = 1$ imply that, for every 'distortion' parameter $t \in (0,1)$, the function $v \mapsto \kappa(t,v)$ is a 'distortion' function.

Note 5. When $\kappa(v,t) \geq v$, which is satisfied by the TCE and PHT kernels (see Example 3), then we have the loading property (see, e.g., Denuit, Dhaene, Goovaerts and Kaas [27]) for the 'risk measure' $L_F(t)$. The property means that the bound $L_F(t) \geq \mu_F$ holds.

In addition to equation (11), it is also instructive to express the L-function $L_F(t)$ by the equation $L_F(t) = \mathbf{E}[F^{-1}(U)K(t,U)]$, where U denotes a uniform rv on $[0,1]$. Consequently, we have the equation

$$L_F(t) = \kappa(1,t)\mu_F + \mathbf{Cov}[F^{-1}(U), K(t,U)]. \tag{12}$$

The next two illustrate the use of the above representation in econometric and actuarial contexts.

Note 6. Assuming $\kappa(1,t) = 0$, we see that equation (12) becomes $L_F(t) = \mathbf{Cov}[F^{-1}(U), K(t,U)]$, which is a covariance representation for the L-function. We refer to Schechtman and Zitikis [79] and also to references therein for details on covariance representations for the S-Gini index.

Note 7. In view of equation (12), when $\kappa(1,t) = 1$, then we have the loading property for $L_F(t)$ if and only if $\mathbf{Cov}[F^{-1}(U), K(t,U)] \geq 0$. Using a general result by Lehmann [55], the latter condition is satisfied when the function $s \mapsto K(s,t)$ is non-decreasing.

Inspired by the research of Zenga [94], who introduces and justifies the use of the ratio $ABC_F(t)/DABC_F(t)$, we next introduce a ratio L-function, that we simply call RL-function, defined by the equation

$$RL_F(t) = \frac{\int_0^1 F^{-1}(s)K_1(s,t)ds}{\int_0^1 F^{-1}(s)K_2(s,t)ds}.$$

The RL-function encompasses a number of curves. For example, when $K_1(s,t) = \mathbf{1}\{s \leq t\}$ and $K_2(s,t) = 1$, then

$RL_F(t)$ becomes the Lorenz curve

$$LC_F(t) = \frac{1}{\mu_F}\int_0^t F^{-1}(s)ds \tag{13}$$

(see Note 8 for details). When $K_1(s,t) = t^{-1}\mathbf{1}\{s \le t\}$ and $K_2(s,t) = (1-t)^{-1}\mathbf{1}\{s > t\}$, then $1 - RL_F(t)$ is Zenga's [94] curve (see Note 9 for details).

Note 8. The notion of Lorenz curve appeared in Lorenz [57] and was later formalized in the form of equation (13) by Pietra [68] (see Giorgi [39] for historical notes). Gastwirth's [29] research initiated a revival of the Lorenz curve and, in turn, of many other curves and indices. Beyond econometric applications, the Lorenz curve has been used in many other areas of research and application, including actuarial science, geography, health sciences and law.

Note 9. As we have noted above, Zenga's [94] curve $Z_F(t)$ is equal to $1 - RL_F(t)$ when $K_1(s,t) = t^{-1}\mathbf{1}\{s \le t\}$ and $K_2(s,t) = (1-t)^{-1}\mathbf{1}\{s > t\}$. Hence, we have the equation

$$ZC_F(t) = 1 - \frac{ABC_F(t)}{DABC_F(t)}$$

with the earlier defined absolute Bonferroni curve $ABC_F(t)$ and the dual of it $DABC_F(t)$. The ratio of these two Bonferroni curves takes on values in $[0,1]$ for every $t \in (0,1)$, and thus Zenga's curve $ZC_F(t)$ also takes on values in $[0,1]$. When the random variable X is equal to a constant almost surely, then the quantile $F^{-1}(s)$ is also equal to the constant for every $s \in (0,1)$ and thus $ZC_F(t) = 0$ for every $t \in (0,1)$, meaning 'zero inequality' or 'egalitarian society'. The other extreme scenario is when, loosely speaking, there is only one member of the society who possesses the entire wealth of the society, and in this case $ZC_F(t) = 1$ for every $t \in (0,1)$. To make the latter statement precise, consider a sample of size n with $n-1$ values being equal to 0 and only one, the largest, equal to $x_{n:n} > 0$. In this case $F_n(x) = 0$ for all $x < 0$, $F_n(x) = (n-1)/n$ for all $0 \le x < x_{n:n}$, and $F_n(x) = 1$ for all $x \ge x_{n:n}$. Zenga's curve corresponding to the cdf $F = F_n$ is then equal to 1 for all $0 < t < (n-1)/n$ and $(n-1)/(nt)$ for all $(n-1)/n \le t < 1$. Obviously now, when the sample size n tends to infinity (meaning a continuous model of the underlying population), Zenga's curve $ZC_F(t)$ approaches 1 at every $t \in (0,1)$, which corresponds to the case that is often called 'absolute inequality' in the research area of Economic Inequality. Given these interpretations of Zenga's curve, it is now natural to define an index of economic inequality by calculating the area beneath Zenga's curve:

$$Z_F = \int_0^1 ZC_F(t)dt.$$

This is the index that Zenga [94] introduced, and which since then has been called Zenga's (new) index of economic inequality. Statistical inferential results in the area have been initiated and developed by Greselin and Pasquazzi [42].

3. ZENGA'S CURVE AND INDEX IN ACTION: A BANK OF ITALY SURVEY

In this section we illustrate how Zenga's curve and inequality index act on real data, obtained from the Bank of Italy (see [4]). We note in passing that other interesting applications of the approach adopted in this section concern Kenyan annual earnings [1] and the 1975–1976 UK pre-tax and post-tax individual incomes (see [95]).

Specifically, in the present paper we work with a data set from the Bank of Italy year 2006 sample survey on household budgets [4]. The information collected in the survey includes demographic characteristics, housing, health, education, employment, incomes, payment instruments, forms of saving, non-durable and durable consumption, forms of insurance.

We have organized this section as follows. Firstly, we discuss main features of the aforementioned data set, present how the data has been collected, and then discuss how one could deal with income data that refers to different family sizes. Then we evaluate Zenga's inequality curve and compare it with the classical Lorenz curve based on the same data set. Following Radaelli's [76] developed methodology, we present a decomposition of a uniformity index into 'within' and 'between' terms. The decomposition preserves the structure of the index itself and leads naturally to an analogous decomposition for the inequality index. Finally, we apply the decomposition in our analysis of the Bank of Italy data set [4].

3.1 The Bank of Italy year 2006 survey

The interviews for the Bank of Italy sample survey of Italian household incomes and wealth were conducted between March and October 2006. The sampling scheme was the same as used in previous surveys: the sample was drawn in two stages, with the primary sampling units being municipalities, and the secondary ones being households.

Before selecting primary units, they were stratified by region and size. Within each stratum, the municipalities were selected so that all those with more than 40,000 inhabitants were automatically included (i.e., self-representing municipalities) whereas smaller towns were selected with probabilities proportional to their size. The individual households to be interviewed were then selected randomly. Until 1987, surveys were conducted with time-independent samples of households (cross sections), but after 1989 samples also included some households interviewed previously (panel households) in order to facilitate an analysis of changes over time.

The 2006 survey covered 7,768 households of which 3,957 (51%) were panel households. The response rate was 42% and, as usual, was higher for panel households (67%) than for non-panel ones (30%). To reduce effects of non-participation, a post-stratification of the sample was done at the end of the survey by reweighing various segments of

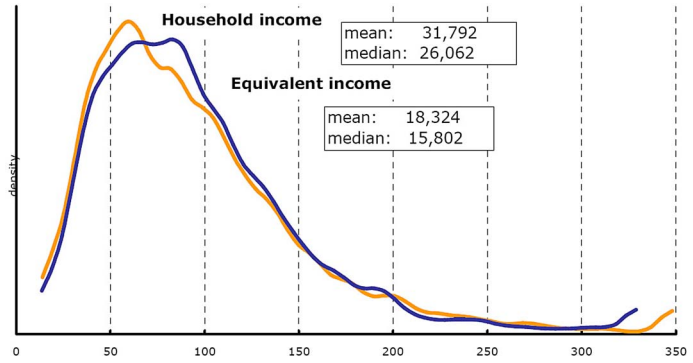

Figure 1. Distribution of incomes in Italy (extracted from [4]).

the population in order to align the characteristics of the final sample to those of the population according to gender, age group, geographical area, size of the municipality of residence (see [4, p. 36] for details). From the items in the questionnaire, main economic aggregates such as net disposable income and net wealth were calculated.

In our present analysis we deal with the net household income, which is the sum of payroll income (net wages and salaries, fringe benefits), pensions and net transfers (pensions, arrears, financial assistance, scholarships, alimony and gifts); we refer to [4, p. 41] for aggregation details. In all the computations that follow we consider the weights $w_i > 0$, $i = 1, \ldots, 7{,}768$, with the sum $\sum w_i = 7{,}768$. The weights are supplied by the central Bank of Italy for each household, and they take into account probabilities of inclusion in the sample and non-participation.

It should be noted that the net household income X could be negative since paid alimony and gifts were subtracted when forming the income. In the 2006 survey this happened for two households, whose overall relative weight was 1.227851 or, in other words, 0.002%. These households were then discarded in order to deal with only non-negative values. For the remaining 7,766 households, the re-scaled weights w_i', $i = 1, \ldots, 7{,}766$, were calculated, with a negligible correction incorporated to get the sum $\sum w_i'$ equal to the number of households, that is, to 7,766.

The distribution of the household incomes (see Fig. 1; extracted from [4]) shows the usual asymmetric form with a relatively low frequency of very low incomes, a bulge around medium–low incomes, and a progressively lower frequency for higher incomes. For completeness of the discussion, we note that the non-parametric estimate of the distribution density in Fig. 1 was obtained using the standard normal kernel function and the bandwidth selected according to the criterion minimizing the asymptotic value of the mean squared error. To obtain more robust results, the values below the 1st and above the 99th percentiles were set equal to the respective percentiles (windsorized estimates). A description of this technique can be found in, for example, Silverman [89].

Note that about 20% of the households had annual incomes less than € 15,334, and a half of the households had incomes less than € 26,062. Approximately 10% of the most affluent households had incomes over € 55,712. The likelihood of being in the right tail of the distribution increased significantly for households whose head had a university degree, was between 51 and 65 of age, self-employed, and lived in the North or Center regions.

To deal correctly with the data, we need to keep in mind that household income is a measure that does not take into account the number of household members, and per capita income does not reflect consumption among members of the same family. To correct these shortcomings, the degree of inequality and poverty can be measured by adjusting the total household income according to an equivalence scale. The result, called *equivalent income*, is obtained as the ratio between the total household income and a scale coefficient. The equivalent income is therefore the income that each individual of a household would need if she/he had lived alone maintaining the same standard of life that she/he enjoys as a member of the household.

Following almost all Bank of Italy studies so far, in the present study we use the modified OECD (Organisation for Economic Co-operation and Development) equivalence scale, which assigns 1 to the household head, 0.5 to the other adult members of the household, and 0.3 to the members under 14 years of age (see Brandolini and Cipollone [9], and references therein). An alternative approach is due to Cutler and Katz [15], who proposed using the scale coefficient

$$CK = [(\#\text{adults}) + (\#\text{children})^{\alpha}]^{\beta} \tag{14}$$

with some parameters $\alpha, \beta \in (0, 1]$. To compare the modified OECD and Cutler-Katz scales, see Table 1. In general, different equivalence scales can affect the inequalities between groups with the same number of household components, but they do not modify inequality within those groups. An interesting application would be to consider sub-populations composed of families of same size and then compare the inequality among the groups. However, for the sake of clarity and simplicity, in the present study we restrict ourselves to a decomposition into only three groups. In any case, it is useful to keep in mind that different equivalence scales produce only slight changes in the coefficients, as we see in Table 1.

3.2 Zenga's curve for the Bank of Italy year 2006 survey equivalent incomes

We begin our study of the Bank of Italy data set by evaluating the so-called uniformity and inequality curves, whose definitions are given below. As a consequence, the Zenga inequality curve and index will be calculated, and then compared with the Lorenz curve and the Gini index. But to this end, we first need to introduce notation, starting with a general set-up of data given in Table 2, where n_{ij} is the frequency of the value x_i in subgroup j. Note that

Table 1. The modified OECD and Cutler-Katz scale coefficients

			Scale coefficients		
Family size	# of adults	# of children	Modified OECD	CK $\alpha=1.0$ $\beta=0.7$	CK $\alpha=0.5$ $\beta=0.8$
1	1		1.0	1.00	1.00
2	2	0	1.5	1.62	1.74
	1	1	1.3	1.62	1.38
3	3	0	2.0	2.16	2.41
	2	1	1.8	2.16	2.08
	1	2	1.6	2.16	1.74
4	4	0	2.5	2.64	3.03
	3	1	2.3	2.64	2.72
	2	2	2.1	2.64	2.41
	1	3	1.9	2.64	2.08
5	5	0	2.5	2.64	3.03
	4	1	2.3	2.64	2.72
	3	2	2.1	2.64	2.41
	2	3	1.9	2.64	2.08
	1	4	1.9	2.64	2.08
6	6	0	2.5	2.64	3.03
	5	1	2.3	2.64	2.72
	4	2	2.1	2.64	2.41
	3	3	1.9	2.64	2.08
	2	4	1.9	2.64	2.08
	1	5	1.9	2.64	2.08

Table 2. The distribution of X in the case of c subgroups

X	1	...	j	...	c	Row total
x_1	n_{11}	...	n_{1j}	...	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	...	n_{ij}	...	n_{ic}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r\cdot}$
Column total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot c}$	N

$n_{ij}=0$ if and only if the variable X does not take the value x_i in the j^{th} subgroup.

We next split the overall frequency distribution $\{(x_1,n_{1\cdot}),\ldots,(x_r,n_{r\cdot})\}$ with non-negative and non-decreasing x_t into two disjoint groups: 1) the lower group $\{(x_1,n_{1\cdot}),\ldots,(x_i,n_{i\cdot})\}$, which includes the first $N_i=\sum_{t=1}^{i}n_{t\cdot}$ observations, and 2) the upper group $\{(x_{i+1},n_{i+1\cdot}),\ldots,(x_r,n_{r\cdot}),(x^*_{r+1},0)\}$, which consists of the dual $N-N_i$ observations; we had to include a virtual observation x^*_{r+1} with null frequency to obtain an upper group in the case $i=r$.

To measure the uniformity between the lower and upper groups, Zenga [94] has suggested the point uniformity index

$$U_i=\frac{\bar{M}_i}{\overset{+}{M}_i},\quad i=1,\ldots,r,$$

where $\bar{M}_i$ and $\overset{+}{M}_i$ are the arithmetic means

$$\bar{M}_i=\frac{1}{N_i}\sum_{t=1}^{i}x_t n_{t\cdot}\left(=\frac{Q_i}{N_i}\right),\quad i=1,\ldots,r,$$

$$\overset{+}{M}_i=\begin{cases}\dfrac{T-Q_i}{N-N_i} & \text{for } i=1,\ldots,r-1,\\ x^*_{r+1} & \text{for } i=r.\end{cases}$$

The U_i takes on values in the interval $[0,1]$, with the value 0 in the case of extreme inequality and 1 in the case of perfect equality. In turn, an index of inequality can be defined as

$$Z_i=\frac{\overset{+}{M}_i-\bar{M}_i}{\overset{+}{M}_i}\left(=1-U_i\right) \tag{15}$$

for $i=1,\ldots,r$. Zenga [94] has also proposed an inequality diagram in the unit square, which is defined for all $i=1,\ldots,r$ as the stepwise function with the value Z_i in the interval $(p_{i-1},p_i]$, where $p_i=N_i/N$ and $N_0=0$.

The Zenga inequality curve for the Bank of Italy data set is depicted in Fig. 2 (top panel). For example, the point $(0.250149,0.654701)$ on the curve means that the mean income of the poorest 25,01% of the household members (a lower partial mean) is 34,53% of the mean income of the richest 74,99% individuals (an upper partial mean). The curve is U shaped and resembles the curve obtained for Dagum distributions by Polisicchio and Porro [70], where the authors also derive analytic expressions for the Zenga curve in the case of several widely used income models. The Zenga inequality curve can be compared to the classical Lorenz curve, which we have depicted in Fig. 2 (bottom panel) using the same Bank of Italy data set. For instance, the point $(0.250149,0.103293)$ on the Lorenz inequality curve means that 25,01% of the poorest individuals own 10,33% of the total income.

Note also that in Fig. 2 (top panel), the horizontal line $y=0.658682$ has been drawn to highlight the intervals of values of p_i in the graph for which the inequality measured by $Z(p_i)$ is lower/greater than its mean value. In particular, the inequality is lower than its mean value on the interval $[0.238707,0.891898]$ and reaches its minimal value $Z(p_i)=0.598743$ at $p_i=0.603221$. As we can see, this graph gives a more accurate representation of the inequality in the extreme parts of the distribution.

Note 10. The horizontal line $y=0.658682$ is a Zenga curve, which corresponds to the empirical distribution coming from the truncated Pareto distribution with the parameter $\theta=0.5$ (see Polisicchio [69]) since for this distribution the inequality between the lower and the upper group is always the same irrespectively of the 'cutting' value x_i.

Furthermore, Zenga [94] has derived a synthetic inequality index defined as the sum of the areas beneath the in-

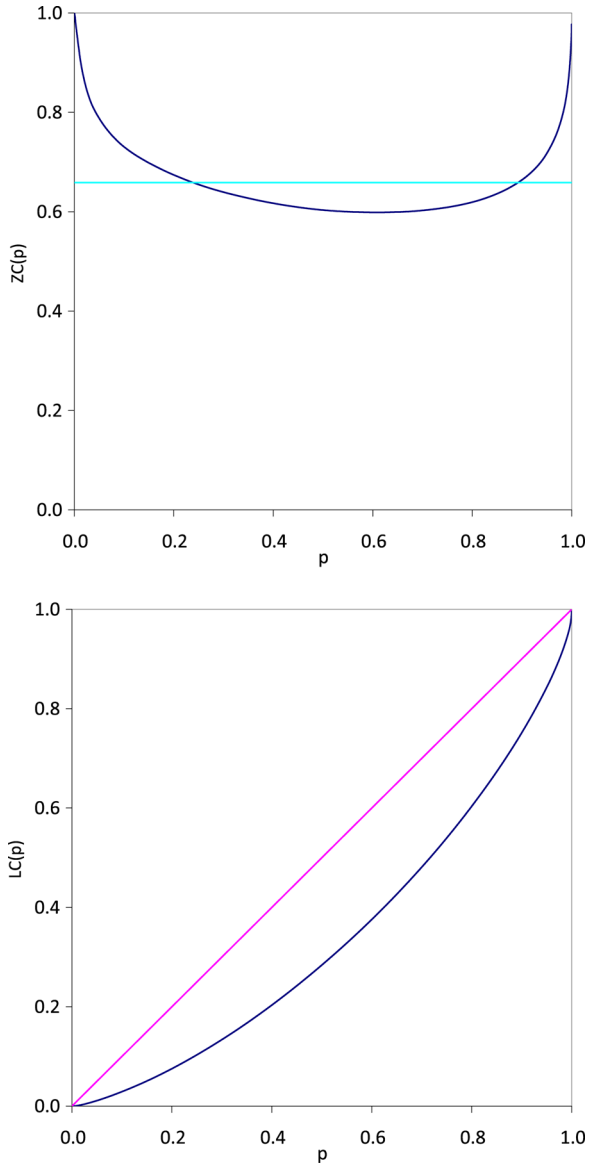

Figure 2. Zenga (top) and Lorenz (bottom) inequality curves for the Bank of Italy year 2006 family income distribution.

equality diagram, which is the weighted arithmetic mean

$$Z = \sum_{i=1}^{r} Z_i \frac{n_{i\cdot}}{N} \tag{16}$$

of the point inequality indices Z_i. Obviously, $Z = 1 - U$, where U is the uniformity index defined by

$$U = \sum_{i=1}^{r} U_i \frac{n_{i\cdot}}{N}. \tag{17}$$

The Zenga inequality index Z in the case of the Bank of Italy data set is equal to 0.658682. To compare, the value of the Gini index is $G = 0.320415$, which means that the

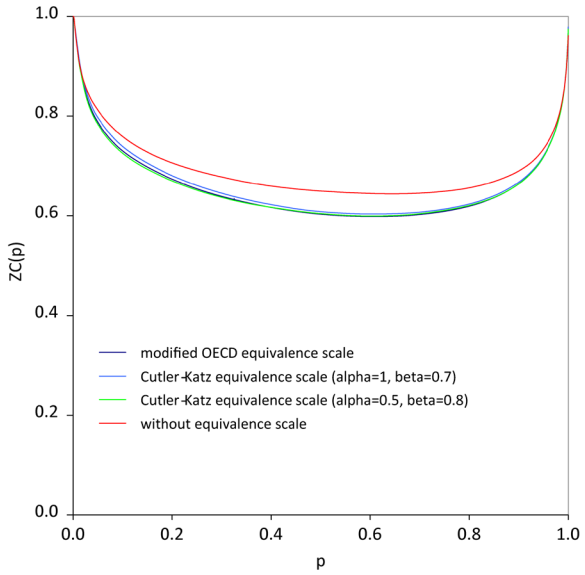

Figure 3. Zenga inequality curves for the Bank of Italy year 2006 family income distribution.

concentration area (i.e., the area between the egalitarian line and the Lorenz curve) is 32,04% of the concentration area that we have in the case of absolute inequality.

We conclude this subsection by noting similarities among the results obtained using different equivalence scales on the Bank of Italy data set. Namely, in Fig. 3 we have depicted the Zenga curves corresponding to equivalent incomes obtained using the modified OECD and Cutler-Katz scales, and also without using any equivalence scale. We see that no matter what scale we use, the inequality curve changes only slightly, whereas a substantial difference is noticeable in the case of the raw data, that is, when no equivalence scale is used. The latter curve would therefore result in a misleading evaluation of the inequality, which confirms the need for pre-treating data when income values correspond to different family compositions.

3.3 Radaelli's decomposition by subgroups of the uniformity and inequality indices

Here we present Radaelli's [76] decomposition of the uniformity and inequality indices and then, in the next subsection, apply it to analyze the Bank of Italy data set.

Let $0 \le x_1 < \cdots < x_r$ denote the distinct values that the variable X takes on all c subgroups (see Table 2 above). For the overall distribution $\{(x_1, n_{1\cdot}), \ldots, (x_r, n_{r\cdot})\}$, in addition to N_i and Q_i introduced above, we define

$$T = \sum_{i=1}^{r} x_i n_{i\cdot} \; (= Q_r) \; > 0 \quad \text{and} \quad M = \frac{T}{N}.$$

Furthermore, let ${}_jN_i$, ${}_jQ_i$, ${}_jT$ and ${}_jM$ denote the analogous quantities for the j^{th} subgroup distribution

$\{(x_i, n_{ij}) : i = 1, \ldots, r\}$. Considering the j^{th} subgroup in Table 2, the j^{th} lower partial mean is defined by (note that the lower group must be non-empty)

$$(18) \qquad {}_j\bar{M}_i = \begin{cases} \dfrac{{}_jQ_i}{{}_jN_i} & \text{when } {}_jN_i > 0, \\ 0 & \text{otherwise,} \end{cases}$$

for all $i = 1, \ldots, r$, and the j^{th} upper mean is

$${}_j\overset{+}{M}_i = \begin{cases} \dfrac{{}_jT - {}_jQ_i}{n_{\cdot j} - {}_jN_i} & \text{for } i = 1, \ldots, r_j - 1, \\ x_r & \text{for } i = r_j, \ldots, r, \end{cases}$$

where r_j denotes the position among $(1, \ldots, r)$ of the subgroup maximal recorded value with non-null frequency, that is, $r_j = \max_{i=1,\ldots,r}\{i : n_{ij} > 0\}$.

The key of Radaelli's [76] decomposition is the point uniformity index

$${}_{j,h}U_i = \frac{{}_j\bar{M}_i}{{}_h\overset{+}{M}_i}$$

for $j, h = 1, \ldots, c$ and $i = 1, \ldots, r$. The index allows comparisons within subgroups and also between two different subgroups. Namely, when $j = h$, then the index ${}_{j,j}U_i$ involves a comparison among means within the same subgroup, whereas for $j \neq h$, the index ${}_{j,h}U_i$ involves a comparison between two different subgroups. In the definition of ${}_{j,h}U_i$ for $j \neq h$, the lower partial mean of the j^{th} subgroup (values $x \leqslant x_i$ in the subgroup j) is compared with the upper mean of the h^{th} subgroup (values $x > x_i$ in the subgroup h). Hence, ${}_{j,h}U_i$ with $j \neq h$ can be interpreted as a point cross uniformity index. The above interpretations therefore suggest splitting the overall uniformity index into 'within' and 'between' components.

In order to examine relationship between the overall i^{th} point uniformity index U_i and the cross uniformity point indices ${}_{j,h}U_i$, we first observe that the overall lower partial mean

$$(19) \qquad \bar{M}_i = \frac{1}{N_i}\sum_j {}_j\bar{M}_i \cdot {}_jN_i \left(= \frac{Q_i}{N_i}\right)$$

is the weighted average of the c group lower partial means ${}_j\bar{M}_i$ with weights given by the cumulative frequencies ${}_jN_i$, that is, the sizes of the groups in which they are evaluated. Second, we observe that, for $i = 1, \ldots, r-1$, the overall i^{th} upper partial mean

$$(20) \qquad \overset{+}{M}_i = (T - Q_i)\left(\sum_h \frac{{}_hT - {}_hQ_i}{{}_h\overset{+}{M}_i}\right)^{-1} \left(= \frac{T - Q_i}{N - N_i}\right)$$

is the weighted harmonic mean of the group upper partial means ${}_j\overset{+}{M}_i$ with weights ${}_hT - {}_hQ_i$. Substituting eqs. (19) and (20) into the definition of U_i, we get

$$(21) \qquad U_i = \sum_j \sum_h {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, \frac{{}_hT - {}_hQ_i}{T - Q_i}$$

for $i = 1, \ldots, r-1$. The overall r^{th} point uniformity index can be analogously decomposed as follows

$$(22) \qquad U_r = \sum_j \sum_h {}_{j,h}U_r \, \frac{n_{\cdot j}}{N} \, \frac{1}{c}.$$

Introducing the weights

$${}_hw_i = \begin{cases} \dfrac{{}_hT - {}_hQ_i}{T - Q_i} & \text{for } i = 1, \ldots, r-1; \\ \frac{1}{c} & \text{for } i = r \end{cases}$$

for $h = 1, \ldots, c$, we combine eqs. (20) and (21) into

$$(23) \qquad U_i = \sum_j \sum_h {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, {}_hw_i$$

for $i = 1, \ldots, r$. Eq. (23) expresses the overall point uniformity index U_i as a weighted average of the uniformities ${}_{j,h}U_i$.

Splitting the second summation in (23) according to whether $h = j$ or $h \neq j$, we arrive at the within/between groups decomposition

$$(24) \qquad U_i = \sum_j {}_{j,j}U_i \, \frac{{}_jN_i}{N_i} \, {}_jw_i + \sum_j \sum_{h \neq j} {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, {}_hw_i.$$

The first sum on the right-hand side of eq. (24) involves all the uniformity indices obtained within each group, and so the sum can be interpreted as a measure of the 'within subgroups' component of the overall point uniformity index U_i. The second (double) sum on the right-hand side of eq. (24) encompasses the uniformity ratios ${}_{j,h}U_i$ evaluated crosswise ($h \neq j$) by comparing the lower partial mean of the j^{th} subgroup with the upper partial mean of another subgroup, and so it measures the 'between subgroups' contribution. The decomposition of the synthetic uniformity index

$$U = \sum_{i=1}^{r} U_i \, \frac{n_{i\cdot}}{N}$$

is now straightforward:

$$(25) \qquad \begin{aligned} U &= \sum_i \sum_j \sum_h {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, {}_hw_i \, \frac{n_{i\cdot}}{N} \\ &= \sum_i \sum_j {}_{j,j}U_i \, \frac{{}_jN_i}{N_i} \, {}_jw_i \, \frac{n_{i\cdot}}{N} \\ &\quad + \sum_i \sum_j \sum_{h \neq j} {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, {}_hw_i \, \frac{n_{i\cdot}}{N}. \end{aligned}$$

Furthermore, denote

$$A_i = \sum_j \frac{{}_jN_i}{N_i}\ {}_jw_i,$$

which is the sum of the weights associated with the 'within' point uniformity indices ${}_{j,j}U_i$. Likewise, the dual of the above introduced A_i is $1 - A_i$, which is associated with the weights assigned to the uniformity cross indices ${}_{j,h}U_i$, $h \neq j$, evaluated between two different groups. With the above notation, the uniformity index U can now be decomposed as the weighted average (see Radaelli [75] for details)

$$\begin{aligned} U &= \sum_i \left[{}_WU_i\ A_i + {}_BU_i\ (1 - A_i)\right]\ \frac{n_{i\cdot}}{N} \\ &= {}_WU\ A + {}_BU\ (1 - A), \end{aligned} \tag{26}$$

where $A = \sum_i A_i \frac{n_{i\cdot}}{N}$ can be interpreted as the overall weight for the uniformity indices evaluated within the same subgroup. The dual $1 - A$ has an analogous meaning.

Aiming at a decomposition for the inequality indices, we decompose $Z_i = 1 - U_i$ as follows

$$\begin{aligned} Z_i &= 1 - \left[{}_WU_i\ A_i + {}_BU_i\ (1 - A_i)\right] \\ &= (1 - {}_WU_i)\ A_i + (1 - {}_BU_i)\ (1 - A_i) \\ &= {}_WI_i\ A_i + {}_BZ_i\ (1 - A_i)\,. \end{aligned}$$

The corresponding decomposition for the overall inequality index becomes

$$\begin{aligned} Z &= 1 - \left[{}_WU\ A + {}_BU\ (1 - A)\right] \\ &= {}_WI\ A + {}_BI\ (1 - A)\,. \end{aligned} \tag{27}$$

We are now ready to apply the above decompositions of the inequality and uniformity indices to the Bank of Italy income data set. This is the topic of our next subsection.

3.4 The decomposition of indices for the Bank of Italy survey equivalent incomes

We divide the entire sample into three subgroups of households according to geographical area: North, Center, and South (including islands). To synthetically depict these three exhaustive and non-overlapping subgroups, we present some of their aggregate characteristics in Table 3, which for each group reports the minimum, maximum, mean and median equivalent incomes (in €), the sample and income shares, the uniformity ${}_jU$ and inequality ${}_jZ$ indices. We see from Table 3 that North is the biggest group (46.26%) owning more than a half (53.76%) of the overall equivalent income. We also see from the table that the mean and median of the equivalent incomes slightly decrease from North to Center, and they drastically decrease to South. Even if the range of incomes for South is strongly lower with respect to the range recorded for North and Center, the uniformity indices are decreasing from North to South.

Table 3. Aggregate characteristics in each geographical area

Eqv. income	North	Center	South	Italy
Min	63.37	5.88	52.00	5.88
Max	507,556.55	811087.83	205,073.02	811,087.83
Mean	21,595.83	21,414.66	12,897.16	18,324.45
Median	18,859.53	18,282.82	11,000.00	15,802.61
Shares				
Sample	0.462648	0.195347	0.342005	1
Income	0.537585	0.225085	0.237330	1
Indices				
${}_jU$	0.395324	0.360710	0.355784	0.341318
${}_jZ$	0.604676	0.639290	0.644216	0.658682

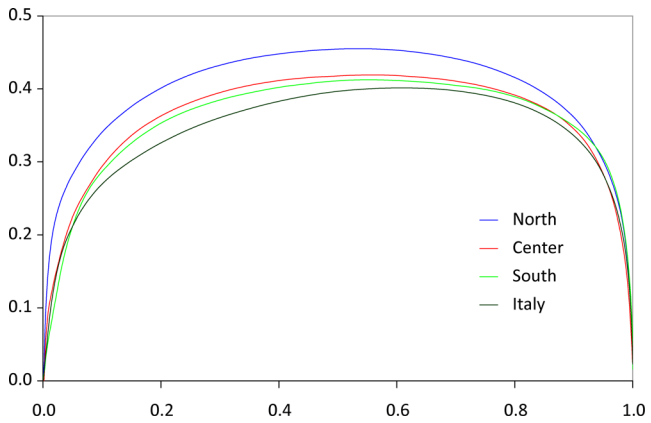

Figure 4. Uniformity curves within each subgroup and for the overall distribution of income.

To proceed with decompositions, we need slightly different notations for 'within' and 'between' quantities, following Radaelli's approach. We start by evaluating the 'within' uniformity curves for the three subgroups. For each of them (i.e., $j = 1, 2, 3$) the stepwise function with the value ${}_{j,j}U_i$ in $(\frac{{}_jN_{i-1}}{n_{\cdot j}}, \frac{{}_jN_i}{n_{\cdot j}}]$ is drawn for $i = 1, \ldots, 7{,}363$, where 7,363 is the number of different equivalent incomes of the 7,766 observed households and ${}_jN_0 = 0$. The three curves, and also the uniformity curve for the overall distribution, are presented in Fig. 4. We glean from the figure that the three subgroup curves keep the same order from the origin of the axes until the 87$^{\text{th}}$ percentile: the North uniformity curve is the highest, followed by the Center curve, with the South curve being the lowest. This depicts a situation in which the strongest inequality is observed in the South subgroup. The ranking changes only for the upper part of the distribution. Considering, for instance, the first 95% of the households in each geographical area, we observe that the South uniformity ratio becomes the greatest. This means that the ratio between the mean of the equivalent income of the poorest 95% of South households represents a larger fraction of the mean of the 5% of the richest households in the South subgroup if compared with the 5% upper group in the North and Center subgroups. In other words, within

the South subgroup there is a higher uniformity for high values of income. If we observe the micro data in detail, we note that in the South subgroup the greatest equivalent income is € 168,187.63, which is not far from the previous values, whereas four highest values are found in the North: € 216,969.24, € 230,317.86, € 284,264.91 and € 507,556.55, which lie quite far from the median and mean incomes of the North subgroup, as reported in Table 3. Similarly for the Center subgroup, the four extreme values are € 221,438.43, € 278,429.61, € 315,767.34 and € 811,087.83. The heavy tailed distributions of the North and Center subgroups are responsible for the crossover between the curves in the last 5^{th} percentile.

To proceed with calculations of ${}_{j,j}U_i$ for $j = 1, 2, 3$, we need to specify the i values. First we note that the value $i = 4{,}303$ is the minimum integer such that the relative cumulative frequency ${}_1N_{i=4,303}/n_{\cdot 1}$ is not lower than 0.5 for the North distribution; hence, $x_{4,303} =$ € 18,859.53 is the median of the equivalent income for the North subgroup. Analogously, for $i = 4{,}131$ we obtain $x_{4,131} =$ € 18,282.82, which is the median income of the Center subgroup. Finally, for $i = 1{,}480$ we have the median income of the South subgroup, which is $x_{1,480} =$ € 11,000.00. Analogously, we arrive at the values $i = 7{,}087$, 7,051 and 6,108 corresponding to the 95^{th} percentiles of the North, Center and South subgroups. The ordinates of the three uniformity curves corresponding to the medians and the 95^{th} percentiles, are:

$$
\begin{aligned}
{}_{1,1}U_{i=4,303} &= 0.454582 \qquad {}_{1,1}U_{i=7,087} = 0.159641\\
{}_{2,2}U_{i=4,131} &= 0.417704 \qquad {}_{2,2}U_{i=7,051} = 0.282822\\
{}_{3,3}U_{i=1,480} &= 0.411023 \qquad {}_{3,3}U_{i=6,108} = 0.304303.
\end{aligned}
$$

These values show different rankings between the three curves. Furthermore, note that the overall uniformity curve is almost always lower than the subgroup uniformity curves: it happens in the central 90% of the distribution. When the overall curve is the lowest one, this means that income is more uniformly distributed within each geographic area than when joining the groups.

Note 11. One might be interested in comparing the Lorenz curves within subgroups (Fig. 5) similarly to our above comparison in the case of the uniformity curves (Fig. 4). However, it seems more awkward to derive our earlier observations from the Lorenz curves. We refer to Radaelli [75] for detailed remarks on comparing respective decompositions by subgroups of the Gini index and the Zenga uniformity and inequality indices.

We next evaluate the uniformities within and between subgroups. To this end, we draw the cross uniformity curves $(j, h = 1, 2, 3)$ as the stepwise functions with the value ${}_{j,h}U_i$ in $(x_{i-1}; x_i]$ for $i = 1, \ldots, 7{,}363$ and $x_0 = 0$. Specifically, Fig. 6 depicts the uniformity curves within each subgroup

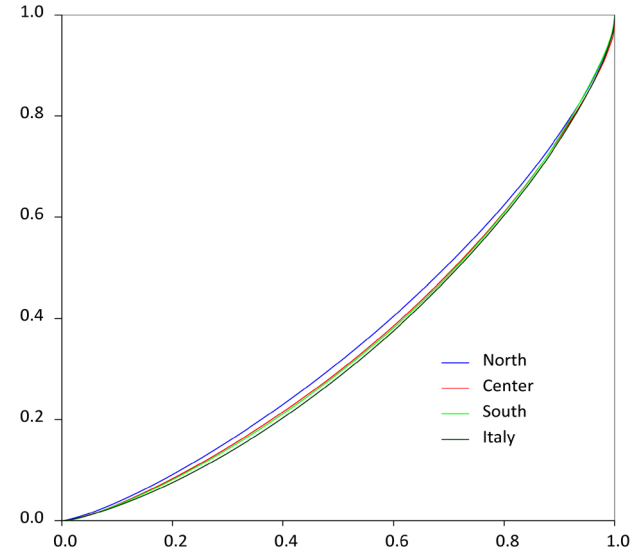

Figure 5. Lorenz curves of the three geographical areas.

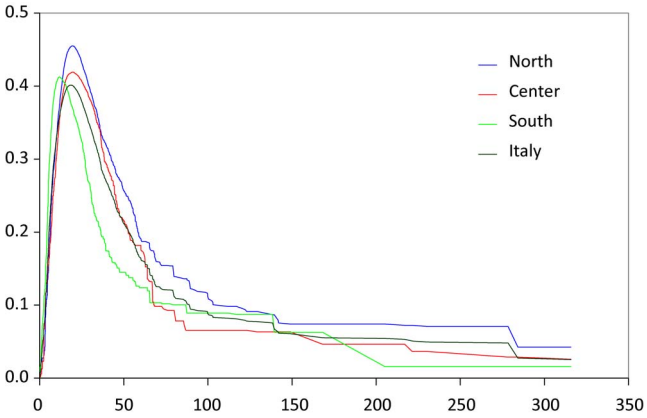

Figure 6. 'Within' uniformity curves with values on the x-axis in thousands of €.

$j = 1, 2, 3$, with the top 0.0002% of the distribution (which corresponds to the two most extreme incomes) discarded.

Furthermore, the three curves in Fig. 7 represent the cross uniformity curves because 'between' curves arise whenever units of one subgroup are compared with units of another subgroup. Each 'between' curve ${}_{j,h}U_i$ provides an answer to the question 'How do we feel about ourselves when compared to those richer than us in subgroup h?' posed by the units in subgroup j with incomes lower than x. For example, the top panel in Fig. 7 has been obtained by comparing lower partial means of the North households with upper partial means of the other geographical areas. The interpretation is immediate. For example, with the reference to the top panel in Fig. 7 and focusing on the first quartile $x_{i=2,596} = Q_1(North) =$ € 14,100.00 of equivalent incomes in the North subgroup on the x axis, the ordinates of the

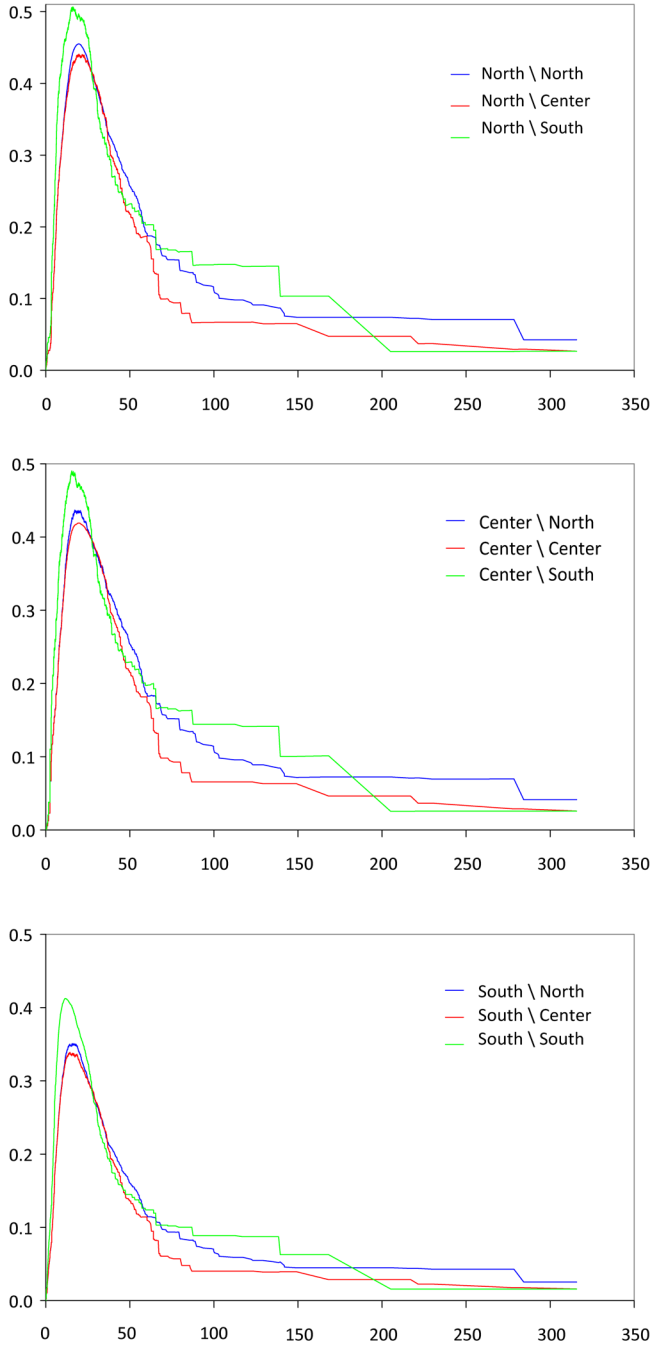

Figure 7. Uniformity curves within and between subgroups with values on the x-axis in thousands of €.

three curves are

$$
\begin{aligned}
{}_{1,1}U_{i=2596} &= 0.419443,\\
{}_{1,2}U_{i=2596} &= 0.405694,\\
{}_{1,3}U_{i=2596} &= 0.488432,
\end{aligned}
$$

respectively. That is, the mean income of the poorest 25% North households represents the 41.94%, 40.57%, and 48.84% of the mean income of the richest households in North (within), Center, and South, respectively. Because of the bijective relation between the index i and the income values x_i, each point in the curve can be equivalently interpreted in terms of income: all northern household members having incomes lower than € 14,100.00 own in average 41.94% of the mean income of people having incomes greater than € 14,100.00 in the same North subgroup. Analogous interpretation can be given for the other two 'between' uniformity curves.

A drastic change in the ordinates in these graphs can be explained as in the following example, corresponding to the uniformity curve between the North and South subgroups in proximity of the income $x =$ € 140,000. A careful inspection of the micro data shows that in the North subgroup the lower mean attains the value ${}_1\bar{M}_i =$ € 21,197.92 in the abscissa $x =$ € 141,844.88, and it increases very slowly to its final value € 21,554.51 at $x =$ € 315,767.34. The upper mean ${}_3\dot{M}_i$ of the South subgroup changes from € 145,665.73 to € 205,073.02 when the last but one income is reached. This fact results in a strong slope of the 'between' uniformity curve.

As we have anticipated, each uniformity index U_i can be decomposed according to eq. (24) into 'within' and 'between' components, and this leads to an analogous decomposition for the overall uniformity index U (see eq. (25)). The results of the decomposition of the 'within' and 'between' components on the overall uniformity are (see eq. (25))

$$\sum_j {}_{j,j}U_i \, \frac{{}_jN_i}{N_i} \, {}_jw_i \frac{n_{i\cdot}}{N} = 0.118713$$

and

$$\sum_i \sum_j \sum_{h\neq j} {}_{j,h}U_i \, \frac{{}_jN_i}{N_i} \, {}_hw_i \, \frac{n_{i\cdot}}{N} = 0.222605,$$

respectively. The overall uniformity index is $U = 0.341318$. Hence, the 'within' and 'between' components account for 34.78% and 65.22% of the overall uniformity, respectively. The main contribution to the overall uniformity is therefore due to the 'between' component.

By expressing the overall uniformity index U as the weighted average of the quantities ${}_WU$ and ${}_BU$ (see eq. (26)), we obtain

$$
\begin{aligned}
U &= {}_WU \cdot A + {}_BU \cdot (1 - A)\\
&= 0.377236 \cdot 0.314692 + 0.324824 \cdot 0.685308\\
&= 0.341318.
\end{aligned}
$$

The decomposition of the overall inequality index (see

eq. (27)) becomes

$$\begin{aligned} Z &= 1 - 0.341318\,(= 0.658682) \\ &= {}_W I\ A + {}_B I\ \ (1 - A) \\ &= 0.622764 \cdot 0.314692 + 0.675176 \cdot 0.685308 \\ &= 0.195979 + 0.462703, \end{aligned}$$

which shows that the overall inequality has contributed 29.75% to the inequality within subgroups and 70.25% to the inequality between subgroups.

4. INDICES OF ECONOMIC INEQUALITY AND THEIR NORMALIZATION

In this section we discuss how indices of inequality and, likewise, of equality can be constructed using functions noted in previous sections. To this end we shall mainly use the Lorenz curve (LC), but our considerations can easily be adjusted to accommodate, for example, Zenga's indices of inequality and equality.

To start with, note that the absolute Lorenz curve $ALC_F(t)$ is convex and takes on values 0 and μ_F at $t = 0$ and $t = 1$, respectively. Hence, normalizing the ALC by the mean μ_F confines the curve to the triangle $\{(s,t) \in [0,1]^2 : s \geq t\}$, with the resulting curve being the LC as defined in equation (13). This geometric interpretation implies that every LC lies between two 'extreme' LC's: $O(t)$ from below and $I(t)$ from above, which are defined as follows. The lower LC

$$O(t) = \begin{cases} 0, & t \in [0,1), \\ 1, & t = 1 \end{cases}$$

corresponds to the case of extreme inequality as, using economic terminology, no one possesses any wealth except one individual, represented by $t = 1$, who possesses everything, that is, 100%. On the other hand, the upper LC

$$I(t) = t,$$

which is known in the econometric literature as the egalitarian Lorenz curve, corresponds to the case of absolute equality as it is produced by every rv that takes on a constant (non-negative) value.

Hence, in view of the bounds $O(t) \leq LC_F(t) \leq I(t)$ that hold irrespectively of F (with non-negative support), we may compare the actual Lorenz curve $LC_F(t)$ to the egalitarian one $I(t)$ and arrive at a measure of inequality. To formulate the idea rigorously, denote the class of all Lorenz curves by $\mathcal{L}$ and let $\mathcal{L} \times I = \{(LC, I) : LC \in \mathcal{L}\}$. We call

$$D : \mathcal{L} \times I \to [0, \infty]$$

the 'functional of inequality' if the bound $D(LC_F, I) \geq D(LC_G, I)$ holds for all $LC_F, LC_G \in \mathcal{L}$ such that $LC_F(t) \leq LC_G(t)$ for all $t \in [0,1]$.

When thinking about examples of the functional $D : \mathcal{L} \times I \to [0, \infty]$, we immediately realize that the maximal distance between $LC_F(t)$ and $I(t)$ conveys little information about the convex body encompassed by the two curves. The area $\int_0^1 (t - LC_F(t))dt$ and its various modifications, on the other hand, have been of particular use and interest when measuring economic inequality and thus has given rise to a wealth of indices in the area. The aforementioned modifications are based on emphasizing or de-emphasizing (depending on the problem under consideration) the difference $t - LC_F(t)$ in some regions of the unit interval $[0,1]$, as follows.

In the classical utility theory and in the Yaari's dual theory of choice under risk, the quantile function $F^{-1}(t)$ is transformed into either $v(F^{-1}(t))$ with a 'utility' function $v(t)$ or $F^{-1}(t)w(t)$ with a 'weight' function (see, e.g., Kaas, Goovaerts, Dhaene and Denuit [53], Denuit, Dhaene, Goovaerts and Kaas [27]). Similarly, many indices of economic inequality are obtained by transforming the function $t - LC_F(t)$ into either $v(t - LC_F(t))$ or $(t - LC_F(t))w(t)$, or possibly into the combination $v(t - LC_F(t))w(t)$ of the two.

A frequent example of $v(t)$ is t^p. Indeed, a large number of indices of economic inequality fall into the class defined by

$$D_{p,w}(LC_F, I) = \left(\int_0^1 (t - LC_F(t))^p w(t) dt \right)^{1/p}. \tag{28}$$

Example 4. When $p = 1$ and $w(t) \equiv 2$ (we shall understand the meaning of the '2' later in this section), then the index $D_{p,w}(LC_F, I)$ becomes the Gini index G_F, which, after integration by parts, can be written as follows:

$$G_F = \frac{1}{\mu_F} \int_0^1 F^{-1}(s)(2s - 1)ds. \tag{29}$$

The Gini index has played a central role in measuring economic inequality since its development by Corrado Gini at the beginning of the 20th century (for a translation Gini's original work, see [36]; also Giorgi [40]). For historical and bibliographical notes on the subject we refer to Giorgi [37–39]. Beyond econometric applications, the Gini index has been used in many other areas of research and application, including actuarial science, geography, mathematics, health sciences, law and public policy.

Note 12. Comparing equations (10) and (29), we see that $G_F = AG_F/\mu_F$, which suggests calling G_F the *relative* Gini index, as opposed to the *absolute* Gini index AG_F. This interpretation of the Gini index G_F invites the notion of the ratio L-statistic, or the RL-statistic for short, which is a special case of the earlier defined RL-function when the kernel $K(s,t)$ does not depend on t. The RL-statistic has been introduced and analyzed by Tarsitano [90]; for special cases, see, e.g., Atkinson [2], Maesono [58], Mimoto and Zitikis [61].

Note 13. The empirical Gini index is obtained by replacing the population cdf $F(x)$ on the right-hand side of equation (29) by the empirical cdf $F_n(x)$:

$$G_n = \frac{1}{\bar{X}n}\sum_{i=1}^{n}\left(\frac{2i-1}{n}-1\right)X_{i:n}. \tag{30}$$

Note that $G_n = D_{p,w}(LC_n, I)$, where $LC_n(t)$ is the empirical Lorenz curve. For statistical inferential results related to the empirical Gini index G_n, we refer to Gastwirth [30, 31]; see also Zitikis and Gastwirth [101] for results concerning the S-Gini index (see Kakwani [54], Donaldson and Weymark [28], Weymark [92]).

Example 5. When $p = 1$ and the weight function $w(t)$ is any, then $D_{p,w}(LC_F, I)$ is the Mehran [60] index, which also appears in Nygård and Sandström [66, 67] and is called the weighted Lorenz area. When $p = 1$ and $w(t) = w_1(t)/\int_0^1 sw_1(s)ds$ for a function $w_1 : [0,1] \to [0,\infty]$, then $D_{p,w}(LC_F, I)$ becomes the generalized Gini index (see Shorrocks and Slottje [88]; Sen [81], p. 142).

Example 6. When $w(t) \equiv 2^p$, then $D_{p,w}(LC_F, I)$ is the E-Gini index

$$E_{F,p} = 2\left(\int_0^1 (t - LC_F(t))^p dt\right)^{1/p}$$

of Chakravarty [12]. In fact, Chakravarty [12] introduces a more general index by replacing t^p and its inverse $t^{1/p}$ by any strictly increasing function $v(t)$ and its inverse $v^{-1}(t)$, respectively. For statistical inferential theory for the E-Gini index $E_{F,p}$, see Zitikis [100] and references therein.

The reason why the Gini index G_F is defined as twice the area between the two curves $I(t)$ and $LC_F(t)$ is to force the index G_F into the interval $[0,1]$ irrespectively of the cdf $F(x)$. Inspired by this observation, we next normalize the index $D_{p,w}(LC_F, I)$ by its maximal value, which is achieved when $LC_F(t) = O(t)$, and obtain the normalized index of inequality

$$ND_{p,w}(O, LC_F, I) = \frac{D_{p,w}(LC_F, I)}{D_{p,w}(O, I)}.$$

An illustrative example follows. (Note that from now on till the end of this section, all the weight functions $w(t)$ are identically equal to 1.)

Example 7. When $w(t) = 1$, then $ND_{p,w}(O, LC_F, I)$ is the following modification

$$NE_{F,p} = (p+1)^{1/p}\left(\int_0^1 (t - LC_F(t))^p\, dt\right)^{1/p}$$

of Chakravarty's [12] E-Gini index $E_{F,p}$. We call $NE_{F,p}$ the normalized E-Gini index. When $p = 1$, the normalized E-Gini index $NE_{F,p}$ reduces to the Gini index G_F.

Since neither $O(t)$ nor $I(t)$ depend on any cdf, it might seem natural to construct an empirical estimator for the index $ND_{p,w}(O, LC_F, I)$ by replacing the population Lorenz curve LC_F by its empirical counterpart $LC_n(t)$. This route may not, however, lead to a natural estimator since in the case of extreme inequality, the empirical Lorenz curve is

$$O_n(t) = \begin{cases} 0, & t \in \left[0, 1-n^{-1}\right), \\ 1-n(1-t), & t \in \left[1-n^{-1}, 1\right]. \end{cases}$$

In other words, for every empirical Lorenz curve $LC_n(t)$, we have the bounds $O_n(t) \le LC_n(t) \le I(t)$ for every set of non-negative rv's of size n. Hence, it is natural to use $ND_{p,w}(O_n, LC_n, I)$ as an estimator of $ND_{p,w}(O, LC_F, I)$. Two illustrative examples follow (see Zitikis [99] for additional examples).

Example 8. When $p = 1$ and $w(t) = 1$, then $D_{p,w}(O_n, I) = (1 - n^{-1})/2$ and so $ND_{p,w}(O_n, LC_n, I)$ is the normalized empirical Gini index

$$NG_n = \frac{1}{1-n^{-1}}G_n,$$

where G_n is defined in equation (30).

Example 9. When $w(t) = 1$, then $D_{p,w}(O_n, I) = (1 - n^{-1})/(p+1)^{1/p}$. Consequently, $ND_{p,w}(O_n, LC_n, I)$ is the empirical normalized E-Gini index

$$NE_{n,p} = \frac{(p+1)^{1/p}}{1-n^{-1}}\left(\int_0^1 (t - LC_n(t))^p\, dt\right)^{1/p},$$

which is an estimator of the earlier defined normalized E-Gini index $NE_{F,p}$.

5. L-PROCESS

We have by now seen a number of examples of the L-function $L_F(t)$, which is unknown in practice since the cdf $F(x)$ is generally unknown. Hence, we construct an empirical estimator for $L_F(t)$ using one of the many available methods, such as parametric, Bayesian, nonparametric, or some other one. The choice of the method depends on the data set, sample size, researchers point of view, etc. We shall next discuss a non-parametric estimator $L_n(t)$ of the L-function which is defined by replacing the population quantile $F^{-1}(s)$ in the definition of $L_F(t)$ by its empirical counterpart $F_n^{-1}(s)$. This gives

$$L_n(t) = \frac{1}{n}\sum_{i=1}^{n} c_{i,n}(t)X_{i:n} \tag{31}$$

with the coefficients

$$c_{i,n}(t) = n\int_{(i-1)/n}^{i/n} K(s,t)ds.$$

When t is fixed, or when the kernel $K(s,t)$ does not depend on t, then $L_n(t)$ is a classical L-statistic (see equation (3)). Various asymptotic results for the L-statistic have been established in the literature. Indeed, essentially all the major results that are available in the case of the sample mean $\bar{X}$ have been extended to L-statistics.

In the context of the present discussion, however, we are particularly interested in statistical inference for the L-function $L_F(t)$. For this we need to establish weak convergence of the L-process

$$\Lambda_n(t) = \sqrt{n}\,(L_n(t) - L_F(t)).$$

The task relies on resolving two problems. First, we need to verify convergence of finite dimensional distributions (fdd's) of the L-process, which can be done using the Cramér-Wold device and the pointwise (i.e., for each $t \in [0,1]$) asymptotic representation

$$\eta_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \lambda_i(t) + o_{\mathbf{P}}(1), \tag{32}$$

which can be found in many proofs dealing with the central limit theorem for the L-statistic (see Chernoff, Gastwirth and Johns [13], Serfling [84], Shorack [87], Shao [86], references therein).

Resolving the second problem, i.e., establishing tightness of the measures generated by the L-processes $\Lambda_n(t)$, $n \geq 1$, is a very complex task (see, e.g., Goldie [41]). We shall next discuss two (among many other possible) ways for achieving this goal, depending on whether the L-process has continuous paths or not, which depends on the kernel $K(s,t)$.

In the case of continuous paths, the following general theorem provides a most convenient way to establish weak convergence of the L-process as it reduces the problem (and, in particular, establishing tightness) to the verification of a moment-type condition.

Theorem 1 (cf., e.g., Ibragimov and Has'minskii [49])**.** *Let $\eta_n(t)$, $n \geq 1$, be continuous stochastic processes defined on $[0,1]$. Let all the fdd's of $\eta_n(t)$ converge to the corresponding fdd's of a process $\eta(t)$ when $n \to \infty$. Assume that there are constants $\alpha \geq \beta > 1$ and $c \in (0,\infty)$ such that, for all $n \geq 1$, we have $\mathbf{E}[|\eta_n(0)|^\alpha] \leq c$ and*

$$\mathbf{E}\left[|\eta_n(t) - \eta_n(u)|^\alpha\right] \leq c|t-u|^\beta \tag{33}$$

for all $t,u \in [0,1]$. Then the process $\eta_n(t)$ converges weakly to $\eta(t)$ when $n \to \infty$, and the limiting process $\eta(t)$ has continuous paths almost surely.

In view of the above theorem, when the L-process $\Lambda_n(t)$ has continuous paths, the main task is to verify condition (33) with $\eta_n(t) = \Lambda_n(t)$. Naturally, the verification is easy (at least in the i.i.d. case) when $\Lambda_n(t)$ admits representation (32) with the remainder term $o_{\mathbf{P}}(1)$ identically equal to zero. This happens, for example, in the case of the uniform empirical process

$$e_n(t) = \sqrt{n}\,(F_n(x) - F(x)), \quad x = F^{-1}(t). \tag{34}$$

The remainder term $o_{\mathbf{P}}(1)$ is not, however, equal to zero in the case of the absolute Lorenz process $\sqrt{n}\,(ALC_n(t) - ALC_F(t))$, which is $\sqrt{n}\int_0^t (F_n^{-1}(s) - F^{-1}(s))ds$. A useful technique to circumvent the problem is to employ the general Vervaat process (see, e.g., Zitikis [98])

$$V_n(t) = \int_0^t (F_n^{-1}(s) - F^{-1}(s))ds + \int_0^{F^{-1}(t)} (F_n(x) - F(x))dx. \tag{35}$$

This is due to the fact (see statement (36) below) that $V_n(t)$ is asymptotically smaller than any of the two integrals on the right-hand side of equation (35). Hence, $V_n(t)$ can be viewed as a remainder term in the representation of the first integral on the right-hand side of equation (35) in terms of the second one. Note that the second integral is equal to $-n^{-1}\sum_{i=1}^n \lambda_i(t)$, where

$$\lambda_i(t) = -\int_0^{F^{-1}(t)} \big(\mathbf{1}\{X_i \leq x\} - F(x)\big)dx.$$

Hence, we have representation (32) for $\eta_n(t) = \sqrt{n}\,(ALC_n(t) - ALC_F(t))$ with the remainder term $o_{\mathbf{P}}(1)$ uniformly in t, provided that

$$\sqrt{n} \sup_{0<t<1} |V_n(t)| = o_{\mathbf{P}}(1), \tag{36}$$

which needs to be verified. In the next note we show how this can be done and also provide additional insights into the Vervaat process $V_n(t)$.

Note 14. The Vervaat process $V_n(t)$ has been investigated in great detail (see, e.g., Zitikis [98], Davydov and Zitikis [23, 24], and references therein). For a use of the Vervaat process in a statistical analysis of the Lorenz and Bonferroni curves, we refer to Csörgő, Gastwirth and Zitikis [14]. We know in particular that $V_n(t)$ is non-negative for all $t \in [0,1]$ and satisfies the bound

$$V_n(t) \leq -\big(F_n(F^{-1}(t)) - t\big)\big(F_n^{-1}(t) - F^{-1}(t)\big) \tag{37}$$

for every cdf F. If, however, the cdf $F(x)$ is continuous at the point $x = F^{-1}(t)$, then $t = F(F^{-1}(t))$ and so, with $e_n(t)$ defined in equation (34) and a function $q(t) > 0$, we have the bound

$$\sqrt{n}\,V_n(t) \leq \sup_{0<t<1} \frac{|e_n(t)|}{q(t)} \sup_{0<t<1} q(t)\big|F_n^{-1}(t) - F^{-1}(t)\big|. \tag{38}$$

Choose, for example, $q(t) = t^{1/2-\epsilon}(1-t)^{1/2-\epsilon}$ with any $\epsilon > 0$. Then the first supremum on the right-hand side of bound (38) is of the order $O_{\mathbf{P}}(1)$. Furthermore, assuming

that the quantile function $F^{-1}(t)$ is continuous and the moment $\mathbf{E}[X^{2+\delta}]$ is finite for some $\delta > 0$, we can find $\epsilon > 0$ such that the second supremum on the right-hand side of bound (38) is of the order $o_{\mathbf{P}}(1)$ (see Mason [59] for a general result). Statement (36) follows.

As we have already hinted at above, some kernels $K(s,t)$ may not lead to L-processes with continuous paths. Since some type of continuity assumption on the paths is needed to at least ensure that, for example, the supremum of the process is a rv, we assume that the paths of the the L-process $\Lambda_n(t)$ are right-continuous. In this case, the following theorem by Davydov [20] is particularly helpful (see also Davydov and Zitikis [26] for related results).

Theorem 2 (Davydov [20]). *Let the processes $\xi_n(t)$, $n \geq 1$, and ξ be defined on the interval $[0,1]$, and let the paths of $\xi_n(t)$ be elements of the space $D[0,1]$ with probability 1. Furthermore, let all the fdd's of $\xi_n(t)$ converge (when $n \to \infty$) to the corresponding fdd's of the process $\xi(t)$. Assume that there are constants $\alpha \geq \beta > 1$ and $c \in (0,\infty)$ such that $\mathbf{E}[|\xi_n(t)|^\alpha] \leq c$ for all $t \in [0,1]$ and*

$$\mathbf{E}[|\xi_n(t) - \xi_n(u)|^\alpha] \leq c|t-u|^\beta \tag{39}$$

for all $t,u \in [0,1]$ such that $|t-u| \geq a_n$, where a_n is a sequence converging to 0. Furthermore, assume that $\xi_n(t)$ can be written as the difference $\xi_n^\circ(t) - \xi_n^*(t)$ of two non-decreasing processes $\xi_n^\circ(t)$ and $\xi_n^*(t)$, with the process $\xi_n^*(t)$ such that

$$\max_{k=1,\ldots,k_n} |\xi_n^*(t_{k+1}) - \xi_n^*(t_k)| = o_{\mathbf{P}}(1) \tag{40}$$

when $n \to \infty$, where $k_n = [1/a_n]$, $t_k = ka_n$ for all $k = 1,\ldots,k_n$, and $t_{k_n+1} = 1$. Then the process $\xi_n(t)$ converge weakly to $\xi(t)$ when $n \to \infty$.

When applying Theorem 2 with $\xi_n(t) = \Lambda_n(t)$, we may not always be able to easily reduce the L-process $\Lambda_n(t)$ to the sum of elementary processes like we have done earlier in the case of the empirical and absolute Lorenz processes. When facing this difficulty, we may explore the following route. Start with the equation

$$\begin{aligned}\Lambda_n(t) - \Lambda_n(u) &= \int_0^1 \sqrt{n}\,\big(F_n^{-1}(s) - F^{-1}(s)\big)\big(K(s,t) - K(s,u)\big)ds\end{aligned} \tag{41}$$

and then use Hölder's inequality. This reduces condition (39) to showing that, for some $p,\ q > 1$ such that $p^{-1} + q^{-1} = 1$, there exist constants $c, \delta \in [0,\infty)$ such that, for all $n \geq 1$,

$$\mathbf{E}\left[\left(\int_0^1 \left|\sqrt{n}\,(F_n^{-1}(s) - F^{-1}(s))\right|^p ds\right)^{\alpha/p}\right] \leq cn^\delta \tag{42}$$

and

$$\frac{1}{n^\delta}\left(\int_0^1 \big|K(s,t) - K(s,u)\big|^q ds\right)^{1/q} \leq c|t-u|^{\,\beta} \tag{43}$$

for all $u,t \in [0,1]$ such that $|t-u| \geq a_n$. This completes our general description of how to possibly verify conditions of Theorem 2 in the context of the L-process $\Lambda_n(t)$.

Example 10. To illustrate conditions (42) and (43), consider the absolute Lorenz process, which is the L-process $\Lambda_n(t)$ with the kernel $K(s,t) = \mathbf{1}\{s \leq t\}$. Condition (43) becomes equivalent to the requirement that $|t-s| \geq a_n$ with $a_n = n^{-\delta/(\beta-1+1/p)}$. The value of $\delta \geq 0$ depends on an upper bound that we can derive for the expectation on the right-hand side of bound (42).

In addition to the above discussed estimation of the L-function $L_F(t)$, there is a variety of other related problems of interest, notably the one about comparing several L-functions $L_{F_1}(t),\ldots,L_{F_K}(t)$. The L-functions can be compared at a given point t, simultaneously for all t in the interval $[0,1]$, or its subinterval. Solutions can be formulated in the form of confidence intervals, confidence bands, hypothesis tests. The underlying K populations may or may not be independent.

When t is fixed, the problem reduces to comparing $K \geq 2$ parameters, whose empirical estimators are L-statistics in the non-parametric case, or functions of parameter estimators in the parametric case. These are complex problems whose solutions in various special cases can be found, for example, in Puri [71–73], Tryon and Hettmansperger [91], and references therein. In the case of actuarial risk measures, these problems have been discussed by Jones and Zitikis [50], Jones, Puri and Zitikis [51], and Brazauskas, Jones, Puri and Zitikis [10].

Comparing $K \geq 2$, or just $K = 2$, functions simultaneously for all $t \in [0,1]$ is a considerably more complex problem. Its resolution heavily relies on the theory of stochastic and, in particular, empirical processes. For results, techniques of proof, and references in the area, we refer to, e.g., Davidson and Duclos [19], Barrett and Donald [5], Hall and Yatchew [43], Linton, Maasoumi and Whang [56], Horváth, Kokoszka and Zitikis [48], Schechtman, Shelef, Yitzhaki and Zitikis [80], and Brazauskas, Jones, Puri and Zitikis [11].

6. SUMMARY

In this paper we have argued that the herein introduced L-function, which is a generalization of the L-statistic, is a natural and useful object encompassing numerous indices of economic inequality, actuarial risk measures, and curves appearing in econometric and actuarial literature. We have illustrated and justified our theoretical considerations with a thorough analysis of a data set from the Bank of Italy year 2006 sample survey on household budgets, and in this way opened up a number of research avenues for practically relevant theoretical investigations of various properties of the L-function. Furthermore, we have noted a number of routes for developing desired statistical inferential results

about the population L-function and, as a by-product, introduced the L-process. When discussing the asymptotic behaviour of the L-process, we have highlighted a particularly useful role of the general Vervaat process as well as of two general results on weak convergence provided by Ibragimov and Has'minskii [49] (for continuous processes) and Davydov [20] (for possibly discontinuous processes); see also Davydov and Zitikis [26] for a generalization of Davydov [20] to multiparameter stochastic processes.

ACKNOWLEDGEMENTS

Our sincere thanks are due to Zhaohai Li, an anonymous referee, and Editor-in-Chief Heping Zhang for suggestions that have resulted in a substantial improvement of the paper. The third author (RZ) is also indebted to Abdelhakim Necir for sharing with him results (published and in progress) concerning statistical inference for actuarial risk measures in the case of heavy-tailed distributions.

Received 28 February 2009

REFERENCES

[1] Aghevly, B. B. and Mehran, F. (1981). Optimal grouping of income distribution data. *J. Amer. Statist. Assoc.* **76** 22–26. MR0608174

[2] Atkinson, A. B. (1970). On the measurement of inequality. *J. Econom. Theory* **2** 244–263. MR0449508

[3] Baklanov, E. A. and Borisov, I. S. (2003). Probability inequalities and limit theorems for generalized L-statistics. (Russian) *Liet. Mat. Rink.* **43** 149–168; translation in *Lithuanian Math. J.* **43** 125–140. MR1996759

[4] Banca d'Italia (2006). *Household Income and Wealth in 2004*. Supplements to the Statistical Bulletin - Sample Surveys, XVI, n.7.

[5] Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica* **71** 71–104. MR1956856

[6] Bhattacharya, P. K. (1974). Convergence of sample paths of normalized sums of induced order statistics. *Ann. Statist.* **2** 1034–1039. MR0386100

[7] Borisov, I. S. and Baklanov, E. A. (1998). Moment inequalities for generalized L-statistics. (Russian) *Sibirsk. Mat. Zh.* **39** 483–489; translation in *Siberian Math. J.* **39** 415–421. MR1639472

[8] Borisov, I. S. and Baklanov, E. A. (2001). Probability inequalities for the generalized L-statistics. (Russian) *Sibirsk. Mat. Zh.* **42** 258–274; translation in *Siberian Math. J.* **42** 217–231. MR1833156

[9] Brandolini, A. and Cipollone, P. (2002). *Urban Poverty in Developed Countries*. Working Paper No. 329, Luxembourg Income Studies Working Paper Series.

[10] Brazauskas, V., Jones, B. L., Puri, M. L., and Zitikis, R. (2007). Nested L-statistics and their use in comparing the riskiness of portfolios. *Scand. Actuar. J.* **2007** 162–179. MR2361124

[11] Brazauskas, V., Jones, B. L., Puri, M. L., and Zitikis, R. (2008). Estimating conditional tail expectation with actuarial applications in view. *J. Statist. Plann. Inference* **138** 3590–3604. MR2450099

[12] Chakravarty, S. R. (1988). Extended Gini indices of inequality. *Internat. Econom. Rev.* **29** 147–156. MR0954119

[13] Chernoff, H., Gastwirth, J. L., and Johns, M. V., Jr. (1967). Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist* **38** 52–72. MR0203874

[14] Csörgő, M., Gastwirth, J. L., and Zitikis, R. (1998). Asymptotic confidence bands for the Lorenz and Bonferroni curves based on the empirical Lorenz curve. *J. Statist. Plann. Inference* **74** 65–91. MR1665121

[15] Cutler, D. M. and Katz, L. F. (1992). *Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980s*. National Bureau of Economic Research Working Paper No. 3964, available online: http://www.nber.org/papers/w3964.

[16] David, H. A. (1973). Concomitants of order statistics. *Bull. Inst. Internat. Statist.* **45** 295–300. MR0373149

[17] David, H. A. (1993). Concomitants of order statistics: review and recent developments. In *Multiple Comparisons, Selection, and Applications in Biometry* (Hamilton, ON, 1991), Statist. Textbooks Monogr. **134**, pp. 507–518, Dekker, New York. MR1241016

[18] David, H. A. and Nagaraja, H. N. (1998). Concomitants of order statistics. In *Order Statistics: Theory & Methods*, Handbook of Statist. **16**, pp. 487–513, North-Holland, Amsterdam. MR1668757

[19] Davidson, R. and Duclos, J.-Y. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* **68** 1435–1464. MR1793365

[20] Davydov, Y. (1996). Weak convergence of discontinuous processes to continuous ones. In *Probability Theory and Mathematical Statistics* (St. Petersburg, 1993), pp. 15–18, Gordon and Breach, Amsterdam. MR1661689

[21] Davydov, Y. and Egorov, V. (2000). Functional limit theorems for induced order statistics. *Math. Methods Statist.* **9** 297–313. MR1807096

[22] Davydov, Y. and Egorov, V. (2001). Functional CLT and LIL for induced order statistics. In *Asymptotic Methods in Probability and Statistics with Applications* (St. Petersburg, 1998), Stat. Ind. Technol., pp. 333–349, Birkhäuser Boston, Boston, MA. MR1890337

[23] Davydov, Y. and Zitikis, R. (2003). Generalized Lorenz curves and convexifications of stochastic processes. *J. Appl. Probab.* **40** 906–925. MR2012676

[24] Davydov, Y. and Zitikis, R. (2004). Convex rearrangements of random elements. In *Asymptotic Methods in Stochastics*, Fields Inst. Commun. **44**, pp. 141–171, Amer. Math. Soc., Providence, RI. MR2106853

[25] Davydov, Y. and Zitikis, R. (2005). An index of monotonicity and its estimation: a step beyond econometric applications of the Gini index. *Metron* **63** 351–372. MR2276056

[26] Davydov, Y. and Zitikis, R. (2008). On weak convergence of random fields. *Ann. Inst. Stat. Math.* **60** 345–365. MR2403523

[27] Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2005). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Chichester, Wiley.

[28] Donaldson, D. and Weymark, J. A. (1980). A single-parameter generalization of the Gini indices of inequality. *J. Econom. Theory* **22** 67–86. MR0568769

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

[30] Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *Rev. Econom. Statist.* **54** 306–316. MR0314429

[31] Gastwirth, J. L. (1974). Large sample theory of some measures of income inequality. *Econometrica* **42** 191–196. MR0411100

[32] Gastwirth, J. L. (1988). *Statistical Reasoning in Law and Public Policy* **1**: *Statistical Concepts and Issues of Fairness*. Statistical Modeling and Decision Science. Academic Press, Boston, MA.

[33] Gastwirth, J. L. (1988). *Statistical Reasoning in Law and Public Policy* **2**: *Tort Law, Evidence, and Health*. Statistical Modeling and Decision Science. Academic Press, Boston, MA. MR0971984

[34] Gastwirth, J., Modarres, R., and Bura, E. (2005). The use of the Lorenz curve, Gini index and related measures of relative inequality and uniformity in securities law. *Metron* **63** 451–469.

MR2252510
[35] Gilat, D. and Helmers, R. (1997). On strong laws for generalized L-statistics with dependent data. *Comment. Math. Univ. Carolin.* **38** 187–192. MR1455483
[36] Gini, C. (2005). On the measurement of concentration and variability of characters. Translated from the 1914 Italian original by Fulvio De Santis. *Metron* **63** 3–38. MR2200970
[37] Giorgi, G. M. (1990). Bibliographic portrait of the Gini concentration ratio. *Metron* **48** 183–221. MR1159665
[38] Giorgi, G. M. (1993). A fresh look at the topical interest of the Gini concentration ratio. *Metron* **51** 83–98.
[39] Giorgi, G. M. (2005). Gini's scientific work: an evergreen. *Metron* **63** 299–315. MR2276053
[40] Giorgi, G. M. (2005). Reasons for a translation. *Metron* **63** 1–2. MR2196347
[41] Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Appl. Probability* **9** 765–791. MR0478267
[42] Greselin, F. and Pasquazzi, L. (2008). *Asymptotic Confidence Intervals for a New Inequality Measure*. Quaderno di Ricerca n.143, Dipartimento di Metodi Quantitativi, Università di Milano Bicocca, Milano.
[43] Hall, P. and Yatchew, A. (2005). Unified approach to testing functional hypotheses in semiparametric contexts. *J. Econometrics* **127** 225–252. MR2156334
[44] Helmers, R. (1982). *Edgeworth Expansions for Linear Combinations of Order Statistics*, Mathematical Centre Tracts **105**, Mathematisch Centrum, Amsterdam. MR0665747
[45] Helmers, R. and Ruymgaart, F. H. (1988). Asymptotic normality of generalized L-statistics with unbounded scores. *J. Statist. Plann. Inference* **19** 43–53. MR0944195
[46] Helmers, R., Janssen, P., and Serfling, R. (1988). Glivenko-Cantelli properties of some generalized empirical DF's and strong convergence of generalized L-statistics. *Probab. Theory Related Fields* **79** 75–93. MR0952995
[47] Helmers, R., Janssen, P., and Serfling, R. (1990). Berry-Esséen and bootstrap results for generalized L-statistics. *Scand. J. Statist.* **17** 65–77. MR1062846
[48] Horváth, L., Kokoszka, P., and Zitikis, R. (2006). Testing for stochastic dominance using the weighted McFadden-type statistic. *J. Econometrics* **133** 191–205. MR2250178
[49] Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York. MR0620321
[50] Jones, B. L. and Zitikis, R. (2005). Testing for the order of risk measures: an application of L-statistics in actuarial science. *Metron* **63** 193–211. MR2210654
[51] Jones, B. L., Puri, M. L., and Zitikis, R. (2006). Testing hypotheses about the equality of several risk measure values with applications in insurance. *Insurance Math. Econom.* **38** 253–270. MR2212526
[52] Jones, B. L. and Zitikis, R. (2007). Risk measures, distortion parameters, and their empirical estimation. *Insurance Math. Econom.* **41** 279–297. MR2339569
[53] Kaas, R., Goovaerts, M. J., Dhaene, J., and Denuit, M. (2001). *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Dordrecht.
[54] Kakwani, N. (1980). On a class of poverty measures. *Econometrica* **48** 437–446. MR0560520
[55] Lehmann, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153. MR0202228
[56] Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Rev. Econom. Stud.* **72** 735–765. MR2148141
[57] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *J. Amer. Statist. Assoc.* **9** 209–219.
[58] Maesono, Y. (2005). Asymptotic representation of ratio statistics and their mean squared errors. *J. Japan Statist. Soc.* **35** 73–97. MR2183501
[59] Mason, D. M. (1982). Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a Glivenko-Cantelli theorem for sample quantiles. *Z. Wahrsch. Verw. Gebiete* **59** 505–513. MR0656513
[60] Mehran, F. (1976). Linear measures of income inequality. *Econometrica* **44** 805–809. MR0455258
[61] Mimoto, N. and Zitikis, R. (2008). The Atkinson index, the Moran statistic, and testing exponentiality. *J. Japan Statist. Soc.* **38** 187–205. MR2458927
[62] Necir, A. and Boukhetala, K. (2004). Estimating the risk-adjusted premium for the largest claims reinsurance covers. In *COMPSTAT 2004–Proceedings in Computational Statistics*, pp. 1577–1584, Physica, Heidelberg. MR2173177
[63] Necir, A., Meraghni, D., and Meddi, F. (2007). Statistical estimate of the proportional hazard premium of loss. *Scand. Actuar. J.* **2007** 147–161. MR2361123
[64] Necir, A. and Meraghni, D. (2009). *Estimating L-functionals for Heavy-Tailed Distributions and Applications*. Laboratory of Applied Mathematics Working Paper. University Mohamed Khider, Biskra.
[65] Nygård, F. and Sandström, A. (1981). *Measuring Income Inequality*. Almqvist & Wiksell, Stockholm.
[66] Nygård, F. and Sandström, A. (1988). The weighted mean difference. *Metron* **46** 21–31. MR1032951
[67] Nygård, F. and Sandström, A. (1989). Income inequality measures based on sample surveys. *J. Econometrics* **42** 81–95.
[68] Pietra, G. (1915). Delle relazioni tra gli indici di variabilità (I, II). *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, a.a. 1914–1915, **LXXIV** 775–804.
[69] Polisicchio, M. (2008). The continuous random variable with uniform point inequality measure I(p). *Statistica & Applicazioni* **6** 136–151.
[70] Polisicchio, M. and Porro, F. (2008). *The I(p) Curve for Some Classical Income Models*. Rapporto di Ricerca del Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali, Universit degli Studi di Milano - Bicocca No. 159, available online: http://www.dimequant.unimib.it/_ricerca/pubblicazione.jsp?id=169.
[71] Puri, M. L. (1965). Some distribution-free k-sample rank tests of homogeneity against ordered alternatives. *Comm. Pure Appl. Math.* **18** 51–63. MR0175212
[72] Puri, M. L. (1965). On the combination of independent two sample tests of a general class. *Rev. Inst. Internat. Statist.* **33** 229–241. MR0182091
[73] Puri, M. L. (1967). Combining independent one-sample tests of significance. *Ann. Inst. Statist. Math.* **19** 285–300. MR0217932
[74] Putt, M. E. and Chinchilli, V. M. (2002). Estimating the asymptotic variance of generalized L-statistics. *Comm. Statist. Theory Methods* **31** 733–751. MR1905142
[75] Radaelli, P. (2008). *On the Decomposition by Subgroups of the Gini's Index and Zenga's Uniformity and Inequality Indexes*. Rapporto di Ricerca del Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali, Universita' degli Studi di Milano - Bicocca No. 150, available online: http://www.dimequant.unimib.it/_ricerca/pubblicazione.jsp?id=170.
[76] Radaelli, P. (2008). A subgroup decomposition of Zenga's Uniformity and Inequality indexes. *Statistica & Applicazioni* **6** 117–136.
[77] Rao, C. R. and Zhao, L. C. (1995). Convergence theorems for empirical cumulative quantile regression functions. *Math. Methods Statist.* **4** 81–91. MR1324691
[78] Rao, C. R. and Zhao, L. C. (1996). Law of the iterated logarithm for empirical cumulative quantile regression functions. *Statist. Sinica* **6** 693–702. MR1410741
[79] Schechtman, E. and Zitikis, R. (2006). Gini indices as areas and covariances: What is the difference between the two representations? *Metron* **54** 385–397. MR2354689
[80] Schechtman, E., Shelef, A., Yitzhaki, S., and Zitikis, R. (2008). Testing hypotheses about absolute concentration curves

and marginal conditional stochastic dominance. *Econometric Theory* **24** 1044–1062.
[81] Sen, A. (1997). *On Economic Inequality*, expanded edition with a substantial annexe by J. E. Foster and A. Sen. Clarendon Press, Oxford.
[82] Seoh, M. and Puri, M. L. (1989). Central limit theorems under alternatives for a broad class of nonparametric statistics. *J. Statist. Plann. Inference* **22** 271–294. MR1006164
[83] Serfling, R. J. (1984). Generalized *L*-, *M*-, and *R*-Statistics. *Ann. Statist.* **12** 76–86 MR0733500
[84] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165
[85] Serfling, R. J. (2002). Robust estimation via generalized *L*-statistics: theory, applications, and perspectives. In *Advances on Methodological and Applied Aspects of Probability and Statistics* (Hamilton, ON, 1998), pp. 197–217, Taylor & Francis, London. MR1977510
[86] Shao, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York. MR2002723
[87] Shorack, G. R. (2000). *Probability for Statisticians*. Springer, New York. MR1762415
[88] Shorrocks, A. F. and Slottje, D. J. (1995). *Approximating Unanimity Orderings: An Application to Lorenz Dominance*, Discussion Paper. University of Essex, Colchester.
[89] Silverman, B. W. (1993). *Density estimation for Statistics and Data Analysis*. Chapman & Hall, London. MR0848134
[90] Tarsitano, A. (2004). A new class of inequality measures based on a ratio of *L*-statistics. *Metron* **62** 137–160. MR2089172
[91] Tryon, P. V. and Hettmansperger, T. P. (1973). A class of non-parametric tests for homogeneity against ordered alternatives. *Ann. Statist.* **1** 1061–1070. MR0353560
[92] Weymark, J. A. (1980/81). Generalized Gini inequality indices. *Math. Social Sci.* **1** 409–430. MR0625274
[93] Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica* **55** 95–115. MR0875518
[94] Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica & Applicazioni* **5** 3–27.
[95] Zenga, M. (2007). Application of a new inequality curve and inequality index based on the ratios between lower and upper partial arithmetic means. In *Proceedings of the 56-th Session of the International Statistical Institute*, Lisbon, Portugal.
[96] Zitikis, R. (1990). Smoothness of the distribution function of an *FL*-statistic. I. (Russian) *Litovsk. Mat. Sb.* **30** 233–246; translation in *Lithuanian Math. J.* **30** 97–106. MR1082454
[97] Zitikis, R. (1990). Smoothness of the distribution function of an *FL*-statistic. II. (Russian) *Litovsk. Mat. Sb.* **30** 500–512; translation in *Lithuanian Math. J.* **30** 231–240. MR1082476
[98] Zitikis, R. (1998). The Vervaat process. In *Asymptotic Methods in Probability and Statistics* (Ottawa, ON, 1997), pp. 667–694, North-Holland, Amsterdam. MR1661510
[99] Zitikis, R. (2002). Analysis of indices of economic inequality from a mathematical point of view. (Plenary Lecture at the 11th Indonesian Mathematics Conference, State University of Malang, Indonesia, 22–25 July 2002. http://ideas.repec.org/p/pqs/wpaper/0092005.html) *Matematika* **8** 772–782.
[100] Zitikis, R. (2003). Asymptotic estimation of the *E*-Gini index. *Econometric Theory* **19** 587–601. MR1997934
[101] Zitikis, R. and Gastwirth, J. L. (2002). The asymptotic distribution of the *S*-Gini index. *Aust. N. Z. J. Stat.* **44** 439–446. MR1934733

Francesca Greselin
Dipartimento di Metodi Quantitativi
per le Scienze Economiche e Aziendali
Università degli Studi di Milano-Bicocca
Milan
Italy
E-mail address: francesca.greselin@unimib.it

Madan L. Puri
Department of Mathematics
University of Texas
Arlington, TX 76019
U.S.A.
E-mail address: mpuri@uta.edu; puri@indiana.edu

Ričardas Zitikis
Department of Statistical and Actuarial Sciences
University of Western Ontario
London, Ontario N6A 5B7
Canada
E-mail address: zitikis@stats.uwo.ca