



HAL
open science

Statistical Methods for Distributional Analysis

Frank A. Cowell, Emmanuel Flachaire

► **To cite this version:**

Frank A. Cowell, Emmanuel Flachaire. Statistical Methods for Distributional Analysis. 2014. halshs-01115996

HAL Id: halshs-01115996

<https://shs.hal.science/halshs-01115996>

Preprint submitted on 12 Feb 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Statistical Methods for Distributional Analysis

Frank A. Cowell
Emmanuel Flachaire

WP 2015 - Nr 07

Statistical Methods for Distributional Analysis

by

Frank A. Cowell

STICERD
London School of Economics
Houghton Street
London, W2CA 2AE, UK
f.cowell@lse.ac.uk

and

Emmanuel Flachaire

Aix-Marseille University
(Aix-Marseille School of Economics), CNRS & EHESS
Institut Universitaire de France
2 rue de la charité,
13002 Marseille, France,
emmanuel.flachaire@univ-amu.fr

April 2014

We thank Russell Davidson, Stephen Jenkins, Michel Lubrano, Andre-Marie Taptue and the editors, Tony Atkinson and Francois Bourguignon for helpful comments. Flachaire would like to thank McGill University and CIRANO for their hospitality, and the Institut Universitaire de France and the Aix-Marseille School of Economics for financial support.

Abstract

This Chapter is about the techniques, formal and informal, that are commonly used to give quantitative answers in the field of distributional analysis – covering subjects including inequality, poverty and the modelling of income distributions. It deals with parametric and non-parametric approaches and the way in which imperfections in data may be handled in practice

JEL Codes: D31, D63, C10

Keywords: Goodness of fit, parametric modelling, non-parametric methods, dominance criteria, welfare indices, inequality measure, poverty measure, influence function, hypothesis testing, confidence intervals, bootstrap

Contents

1	Introduction	3
1.1	Why Statistical Methods?	3
1.2	Basic notation and terminology	4
1.3	A guide to the chapter	5
2	Data	6
2.1	Data sources	6
2.2	Data structure	7
2.3	Data problems	8
2.3.1	Measurement error and data contamination	8
2.3.2	Incomplete information	8
2.4	Grouped data	10
3	Density estimation	10
3.1	Parametric estimation	12
3.1.1	Pareto	13
3.1.2	Lognormal	14
3.1.3	Generalized Beta	15
3.1.4	Goodness of fit	17
3.2	Kernel method	18
3.2.1	From histogram to kernel estimator	18
3.2.2	Bandwidth selection	21
3.2.3	Adaptive kernel estimator	22
3.2.4	Multivariate and conditional density	23
3.3	Finite-mixture models	24
3.3.1	A group decomposition approach	24
3.3.2	Number of components and number of groups	25
3.3.3	Group profiles explanation	27
3.3.4	Finite mixture of regressions	30
3.4	Finite sample properties	32
4	Welfare indices	34
4.1	Basic cases	34
4.2	Asymptotic inference	36
4.2.1	The influence function	36
4.2.2	Background results	38
4.2.3	QAD Welfare indices	39
4.3	Application: Inequality measures	40
4.3.1	The generalised entropy class	40
4.3.2	The mean deviation and its relatives	42
4.3.3	The Gini coefficient	43
4.4	Application: Poverty measures	45
4.4.1	Foster-Greer-Thorbecke (FGT)	47
4.4.2	Sen poverty index	48

4.4.3	Sen-Shorrocks-Thon (SST) poverty index	49
4.5	Finite sample properties	50
4.5.1	Asymptotic and bootstrap methods	50
4.5.2	Simulation evidence	51
4.5.3	Inference with heavy-tailed distributions	53
4.5.4	Testing equality of inequality measures	56
4.6	Parametric approaches	59
5	Distributional comparisons	60
5.1	Ranking and dominance: principles	61
5.1.1	Dominance and welfare indices	61
5.1.2	Stochastic dominance	63
5.2	Ranking and dominance: implementation	63
5.2.1	Asymptotic distributions	64
5.2.2	Dominance: an intuitive application	67
5.2.3	The null hypothesis: dominance or non-dominance	70
5.2.4	Hypothesis testing	72
6	Other estimation problems	74
6.1	Contamination	74
6.1.1	The concept of robustness	74
6.1.2	Robustness, welfare indices and distributional comparisons	75
6.1.3	Model estimation	77
6.2	Incomplete data	78
6.2.1	Censored and truncated data	78
6.2.2	Trimmed data	80
6.3	Semi-parametric methods	84
6.3.1	The model	85
6.3.2	Model estimation	85
6.3.3	Choice of β	86
6.3.4	Inequality and dominance	86
7	Conclusions	90
7.1	Important lessons: a round-up	90
7.2	A worked example	91
7.3	A cribsheet	92
	References	95

1 Introduction

This Chapter is about the techniques, formal and informal, that are commonly used to give quantitative answers in the field of distributional analysis – covering subjects such as inequality, poverty and the modelling of income distributions.

At first sight this might not appear to be the most exciting of topics. Discussing statistical and econometric techniques could appear to be purely secondary to the important questions in distributional analysis. However, this is not so. At a very basic level, without data what could be done? Clearly if there were a complete dearth of quantitative information about income and wealth distributions we could still talk about inequality, poverty and principles of economic justice. But theories of inequality and of social welfare would stay as theories without practical content. Knowing how to use empirical evidence in an appropriate manner is essential to the discourse about the welfare economics of income distribution and to the formulation of policy. Furthermore, understanding the nature and the limitations of the data that are available – or that may become available – may help to shape one’s understanding of quite deep points about economic inequality and related topics; good practice in quantitative analysis can foster the development of good theory.

1.1 Why Statistical Methods?

If we carry out a simple computation of the values of an inequality or poverty measure, computed from two different samples, we will usually find greater inequality or poverty in one sample, *even if the two samples come from the same population*. Clearly simple computation alone is not enough in order to draw useful conclusions from the raw data: statistical methods are required to test the hypothesis that the two values are not statistically different. For instance, Table 1 reports the values of the Gini and Theil inequality indices,¹ with confidence intervals at 95%, computed from two samples of 1000 observations drawn from the same distribution: The two samples are independent, with observations drawn independently from the Singh-Maddala distribution with parameters $a = 2.8$, $b = 0.193$ and $q = 1.7$, which closely mimics the net income of German households, up to a scale factor (Brachmann et al. 1996). Clearly the values of the Gini and Theil indices are greater in sample 1 than in sample 2. However, the confidence intervals (in brackets) intersect for both inequality measures, which leads us to not reject the hypothesis that the level of inequality is the same in the two samples.

There is a wide variety of inequality indices in common use. Different indices, with different properties, could lead to opposite conclusions in practice. Lorenz curves comparisons can be very useful, since a (relative) Lorenz curve always lying above another one implies that any comparisons of relative inequality measures would lead to similar conclusions – a result that holds for any inequality measures respecting anonymity, scale invariance, replication invariance and

¹For formal definitions see equations (51), (69), (70) below.

	sample 1	sample 2
Gini	0.303	0.285
	$[0.286; 0.320]$	$[0.271; 0.299]$
Theil	0.158	0.135
	$[0.133; 0.183]$	$[0.120; 0.151]$

Table 1: Inequality indices with confidence intervals at 95%

the transfer principle (Atkinson 1970). In practice, we have on hand a finite number of observations and, empirical Lorenz dominance can be observed many times when the two samples come *from the same population*. In the case of two independent samples of 1 000 observations drawn from the same Singh-Maddala distribution, we obtain sample Lorenz dominance 22% of cases. Dardanoni and Forcina (1999) argue that it can be as high as 50% of cases due to the fact that empirical Lorenz curve ordinates are typically strongly correlated. This demonstrates the need to use statistical methods.

The point about simple computation being insufficient is also easily demonstrated in the case of Lorenz curves: Figure 1 shows the difference between the empirical Lorenz curves obtained from two independent samples drawn from the same distribution, with confidence intervals at 95% calculated at the population proportions $q = 0.01, 0.02, \dots, 0.99$. The ordinates are always positive: so it is clear that one empirical Lorenz curve always dominates the other. However, the confidence intervals show that *each* Lorenz curve ordinate difference is never significantly different from zero; as a result Lorenz dominance in the population is not as clear as simple computation from the sample might suggest. To be able to make conclusions on dominance or non-dominance, we need to test simultaneously that *all* ordinate differences are statistically greater than zero, or not less than zero. Appropriate test statistics need to be used to make such multiple comparisons.

In this chapter we will provide a survey of the theory and methods underlying good practice in the statistical analysis of income distribution. We also offer a guide to the tools that are available to the practitioner in this field.

1.2 Basic notation and terminology

Throughout the chapter certain concepts are used repeatedly and so it is convenient to collect here some of the terms that are used repeatedly.

- *Income* y . Here “income” this is merely a convenient shorthand for what in reality may be earnings, wealth, consumption, or something else. We will suppose that y belongs to a set $\mathbb{Y} = [\underline{y}, \bar{y})$, an interval on the real line \mathbb{R} .
- *Population proportion* q . For convenience we will write $q \in \mathbb{Q} := [0, 1]$.
- *Distribution* F . This is the cumulative distribution function (CDF) so that, for any $y \in \mathbb{Y}$, $F(y)$ denotes the proportion of the population that

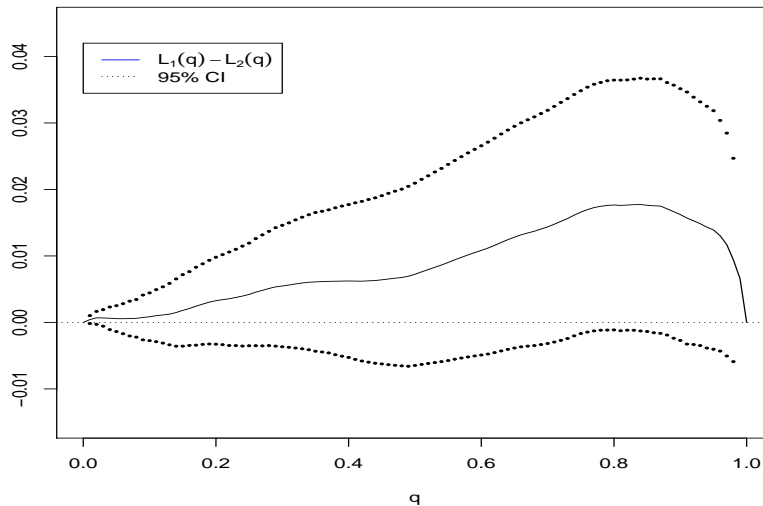


Figure 1: Difference between two empirical Lorenz curves, $\hat{L}_1(q) - \hat{L}_2(q)$, with 95% confidence intervals. The samples are drawn from the same distribution.

has income y or less. Where the density is defined we will write the density at $y \in \mathbb{Y}$ as $f(y)$. The set of all distribution functions will be denoted \mathbb{F} .

- *Indicator function* $\iota(\cdot)$. Suppose there is some logical condition D which may or may not be true. Then $\iota(\cdot)$ is defined as:

$$\iota(D) = \begin{cases} 1 & \text{if } D \text{ is true} \\ 0 & \text{if } D \text{ is not true} \end{cases} \quad (1)$$

1.3 A guide to the chapter

We begin with a discussion of some of the general data issues that researchers should bear in mind (Section 2). Section 3 deals with the issues that arise if we want to try to “model” an income distribution: the motivation for this is that sometimes it makes sense to approach the analysis of income distributions in two stages, (1) using a specific functional form or other mathematical technique to capture the evidence about the income distribution in an explicit model and (2) making inequality comparisons in terms of the modelled distributions. Section 4 deals with the general class of problem touched on in our little example outlined in Table 1: the emphasis is on hypothesis testing using sample data and we cover both inequality and poverty indices. As a complement to this Section 5 deals with the class of problem highlighted just now in Figure 1: we look

at a number of “dominance” questions that have a similarity with the Lorenz problem described there. Section 6 returns to mainly data-related questions: how one may deal with some of the practical issues relating to imperfections in data sets. Finally, in Section 7, we draw together some of the main themes that emerge from our survey of the field.

2 Data

2.1 Data sources

It is not difficult to imagine instances where there is a known, finite set of income receivers and where the income associated with each income-receiver is observable (example: if there are 50 states in a federation and one wishes to analyse the change in the distribution of announced military contracts awarded by the federal government among those 50 states). Under those circumstances a complete enumeration of the relevant “population” (the 50 states) is possible and the income of each member of this “population” is measured with complete accuracy (from the federal government announcement). There is very little to do in terms of statistical analysis and no data problem. But this kind of example is rarely encountered in practice and might be dismissed as somewhat contrived. It is much more common to have to deal with situations where an enumeration of the population is impossible and we have to rely on some sort of sample.

Administrative data

Governments and government agencies have long published summaries of income distributions in grouped form; in many countries official data providers have gone further and made available to researchers micro-data from official sources which could be used, for example, to analyse the distribution of income and wealth. The data made available in this way used to be of similar size to sample surveys (discussed below). However, it is increasingly the case that very large data sets have been opened up for research, an order of magnitude larger – effectively complete collections of administrative data rather than official samples from them. It might be tempting to treat these as methodologically equivalent to the complete-enumeration case described above. But this would be to overlook two points. First, administrative data will only contain what is legally permissible and what government agencies find convenient to release: if, for example, one is interested in the distribution of personal incomes a very large dataset of tax records could be extremely useful but it will miss out many of those persons who are not required to file tax returns. Second, the design of the data-set may not match what the social scientist or economist would wish: for example, if one wishes to adjust the data to allow for differences in need according to the type of household or family in which each person lives, the required information for constructing an appropriate equivalence scale may not be present in the same data set.

Survey data

The problems from administrative data stem largely from the fact that the data are the by-product of information gathered for other purposes. It is clear that specially designed surveys have a potential advantage in this respect. However, although surveys are usually purpose-built (and often purpose-built using advice from social scientists) one also has to be cautious about their limitations. This concerns not only the smaller size and worse response rate than the administrative-data counterparts. Once again the survey design may exclude some sections of the population (a survey based on households would obviously miss out people who are homeless and those in institutions) and, where there is an attempt to create longer series of surveys, the criteria for the design of contemporary surveys may follow a standardised format determined by conventions that are no longer relevant.

2.2 Data structure

In implementing the statistical criteria discussed in this chapter one needs to be clear about the relevant assumptions concerning the way the sample was drawn.

Simple design

In the majority of this chapter we will take it that simple random sampling is an appropriate assumption. By this we mean that the sample has been designed in such a way that each member of the population has an equal probability of being included in the sample. This can be taken as an ideal case that enables one to focus on the central issues of statistical inference. Even the supposedly “ideal” case may not be ideal in practice if the sampling frame is inappropriate – it could be out of date, it could be specified in such a way that part of the population is excluded (see the remarks above about homeless people).

Complex design

In practice there are often simple practical reasons why something other than simple random sampling is used.² Two features in particular are often built into the design of the sample. *Clustering* the observations by geographical location may reduce the costs of running the survey, both in terms of initial visits to carry out the survey and in follow-up visits for monitoring and completing missing information. *Stratification* is a common technique for deliberately oversampling certain categories of respondent in order to ensure that there is adequate representation in the combined sample of certain types of individuals or households that are of particular interest but that are likely to show up comparatively rarely either because they are genuinely rare in the population or because they are less likely to respond to the survey (for example it is commonly found that richer households are over-represented in the “non-response” category and if one were

²See Deaton (1997) for a full discussion of the issues involved.

just to ignore that possibility there would be the danger of having a biased sample). In effect one divides up the population of interest into subpopulations and chooses a sample for each subpopulation – each stratum – at an appropriate sampling rate.

Although the assumption of a simple random sample sweeps aside practical problems associated with the design of the survey, this idealised case gives us a good base for explaining the core issues in estimation and inference. At appropriate points in sections 4 and 5 we will comment on the extensions to the complex-data case and other related issues.³

Other problems with the data merit special discussion. We briefly outline the nature of these problems here and then return to a formal analysis of them (in section 6) after we have extensively discussed conventional inference problems in the preceding sections.

2.3 Data problems

2.3.1 Measurement error and data contamination

Measurement error in income-distribution analysis can be handled in a way similar to measurement error in other contexts. Observed income is true income adjusted by an error term (Chesher and Schluter 2002) and the resulting model resembles the problem of decomposition by factor source; *data contamination* can be represented as a mixture of a true distribution and a contamination distribution: the resulting model resembles the problem of decomposition by population subgroup (Cowell 2000, Cowell and Fiorio 2011). However, the appropriate model for analysing this second type of problem uses tools which are useful for the analysis of questions beyond the narrow data-contamination question. This will be discussed in sections 4 to 6.

2.3.2 Incomplete information

In many practical applications we need to deal with situations in which some parts of the sample space are excluded completely from the sample data or where information in part of the sample is missing; for convenience we will refer to this part of the sample as the “excluded” subset, even though some information may be available. The exclusion of information may be imposed by the data provider, for example because of reasons of confidentiality, or it may be introduced by the researcher in order to deal pragmatically with some other problem in the data.

Table 2, taken from Cowell and Victoria-Feser (2003), sets out the main cases that are of interest. There are two principal issues for the researcher to consider, as follows:

³An example: if the data are based on a simple survey of households, but one wants to infer something about the distribution of individuals one needs to weight each household observation by an amount proportional to the number of persons in the household; this structure is similar to the weights introduced by stratification.

	<i>Information about Excluded Sample</i>		
	<i>None</i>	<i>Sample proportion</i>	<i>Multiple statistics</i>
<i>incomes below \underline{z} and above \bar{z} excluded</i>	A	B	C
<i>lowest $100\beta\%$ and highest $100\bar{\beta}\%$ excluded</i>	D	(E)	(F)

Table 2: Types of incomplete information

Boundaries of the excluded subset What determines the boundaries of the excluded subset of the sample space? There are two possible cases summarised in the rows of Table 2: (i) a subset of \mathbb{Y} is specified, or (ii) a subset of \mathbb{Q} is specified. In the first case the income-boundaries of the excluded subset (\underline{z}, \bar{z}) are fixed but the proportions of the excluded subsets $(\underline{\beta}, \bar{\beta})$ are unknown, although these proportions can be estimated if enough information is available. In the second case the boundaries of the excluded sample are fixed by the trimming proportions in the lower and upper tail $(\underline{\beta}, \bar{\beta})$ and the incomes at the boundary of the excluded samples (\underline{z}, \bar{z}) are unknown.

Information in the excluded subset There are several assumptions about the availability of information in the excluded part of the sample. The situation is going to depend on the particular problem in hand and the principal cases are summarised in the columns of Table 2. At one extreme the excluded subset is just *terra incognita* (left-hand column). On the other hand it may be that the data-provider makes available several summary statistics related to excluded subset (right-hand column).

So, in principle there are altogether six possible cases, but in practice only four are relevant:⁴

- Case A is the standard form of truncation.
- Case B represents “censoring”: in this case there are point masses at the boundaries (\underline{z}, \bar{z}) that estimate the population-share of the excluded part.⁵
- Case C is an extension of standard estimation problem with grouped data (Gastwirth et al. 1986).

⁴If, as is usual, trimming is something that is done voluntarily by the user rather than being imposed by the data-provider, then Cases E and F are not relevant in practice.

⁵A standard example of this “top-coding” of some income components in the Current Population Survey – observations above a given value \bar{z} are recorded as \bar{z} (Polivka 1998). In practical applications of distributional analysis, such as inferring inequality trends, researchers have adopted a number of work-rounds such as multiplying top-coded values by a given factor (Lemieux 2006, Autor et al. 2008) or attempting imputations for missing data (Burkhauser et al. 2010, Jenkins et al. 2011).

- Case D represents the case of trimming.

The implications of these issues for distributional analysis are considered in section 6.2 below.

2.4 Grouped data

For reasons of economy and convenience it used to be common practice for statistical offices to make income-distribution data available only in grouped form (case C in Table 2 above). Typically this would involve a simple table with a comprehensive set of pre-set income intervals, the numbers of individuals or households falling into each interval and (sometimes) the average income associated with each interval. Tabulated data are less usual today, although researchers are increasingly using historical data to construct long run time series. So it is useful to consider briefly the analytical issues that arise in connection with this type of data.

One way of using such data effectively is to estimate the underlying income distribution using parametric modelling. This can be done either by using interpolation methods in each of the intervals (see, for example, Cowell 2011) or by fitting a distribution to the bulk of the data – suitable parametric methods are discussed in Section 3.1 below. Non-parametric methods are necessarily quite limited, because of the restrictions imposed by the data.⁶ However, an interesting problem presented by any sort of grouped data is to compute bounds on inequality indices. One uses the available information to compute a maximum-inequality distribution and a minimum-inequality distribution by making alternative extreme assumptions about the way the data are distributed within each interval (Gastwirth 1975, 1972; Cowell 1991).

3 Density estimation

The analysis of a probability density function is a powerful tool to describe several properties of a variable of interest. For instance, Figure 2 shows the estimated density function of GDP per capita in 121 countries across the world in 1988.⁷ We can see that the density function is bimodal. The existence of two modes suggests that there are two distinct groups: one composed of the “richest” countries, and another consisting of the “poorest”. The second mode is much less pronounced than the first, which indicates that the two groups are not of the same size: there are relatively few “rich” countries, and distinctly more “poor” countries. Further, the first mode is located just to the left of the value 0.5 on the X-axis, while the second is found at around 3. We can thus conclude from this figure that, on average in 1988, “rich” countries enjoyed

⁶The problem of statistical inference with grouped data is discussed in Hajargasht et al. (2012).

⁷The data are unweighted (each country as equal weight) and are taken from the *Penn World Table* of Summers and Heston (1991). The horizontal axis is the per capita GDP for each country normalised by the (unweighted) mean over all countries.

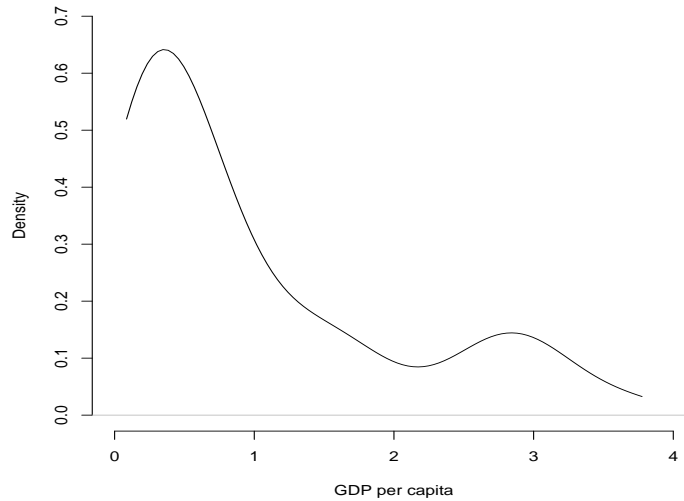


Figure 2: Kernel density estimation of GDP per capita in 121 countries normalized by the global mean.

a level of GDP per capita that was around three times the average, whereas that of “poor” countries was only half of the average level. It is clear from this example that much more information is available from the full distribution of a variable, than the restricted information provided by standard descriptive statistics, as the mean, variance, skewness or kurtosis, which summarize each limited properties of the distribution on single values.

In the multivariate case, the conditional density function can provide useful insights on the relationship between several variables. For instance, Figure 3 shows the estimated density functions of wages conditional to experience, for individuals with the same level of education.⁸ We can see that, as experience increases, the conditional distribution becomes bimodal and the gap between the two modes increases. It suggests that the population is composed by two subgroups and the marginal impact of experience on wages is not the same for the two groups. A standard regression tracks the dynamics of the first moment of the conditional distribution only, and then cannot highlights the features just described. Here, a linear regression of wages on experience would estimate the marginal impact of experience on the average of wages for all individuals, while the graphical analysis suggests that experience does not affect individual’s wages identically. Mixture models of regressions would be more appropriate in such cases.

In practice, the functional form of the density function is often unknown and

⁸The data are simulated.

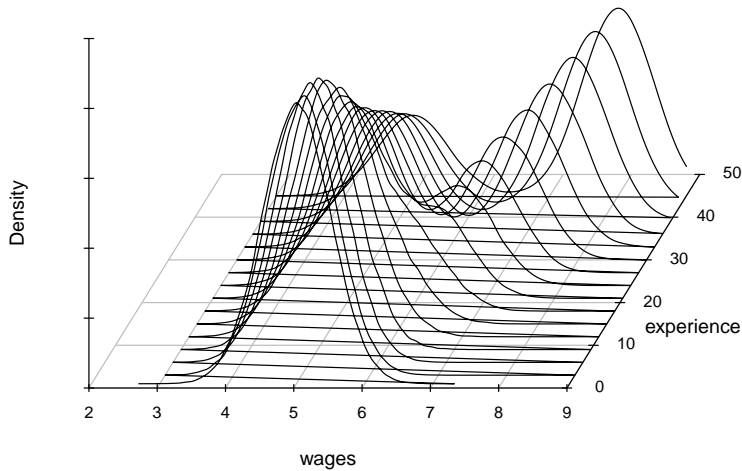


Figure 3: Conditional density estimates of wages on experience

has to be estimated. For a long time, the main estimation method was mainly parametric. However, a parametric density estimation requires the choice of a functional form a priori, and most of them do not fit multimodal distribution. In the last two decades, nonparametric and semi-parametric estimation methods have been extensively developed. They are often used in empirical studies now and allow us to relax the specific assumptions underlying parametric estimation method, but they require in general more data on hand.

In this section, we will present parametric, nonparametric and semi-parametric density estimation methods. Standard parametric methods are presented in section 3.1, kernel density methods in section 3.2 and finite-mixture models in section 3.3.

3.1 Parametric estimation

We say that a random variable Y has a probability density function f , if the probability that Y falls between a and b is defined as:

$$P(a < Y < b) = \int_a^b f(y) dy,$$

where a and b are real values and $a < b$. The density function f is defined as non-negative everywhere and its integral over the entire space is equal to one.

A parametric estimation requires to specify a priori a functional form of the density function, that is, to know the density function up to some parameters. The density function can then be written $f(y; \theta)$, where θ is a vector of k unknown parameters and y a vector of n observations. The estimation remains to estimate θ , it is usually done by maximizing the likelihood to observe the actual values in a sample. If the data are independent and identically distributed, the joint density function of n observations y_1, y_2, \dots, y_n is equal to the product of the individual densities:

$$f(y; \theta) = \prod_{i=1}^n f(y_i; \theta).$$

The estimation of the density function therefore requires the maximization of this function with respect to θ . Since the logarithmic function is positive and monotone, it is equivalent to maximize

$$\ell(y; \theta) = \log f(y; \theta) = \sum_{i=1}^n \log f(y_i; \theta),$$

from which the resolution is often much simpler. This estimation method is known as the maximum likelihood method.

3.1.1 Pareto

Pareto (1895) initiated the modelling of income distribution with a probability density function⁹ that is still in common use for modelling the upper tail of income and wealth distributions. Beyond a minimal income, he observed a linear relationship between the logarithm of the proportion of individuals with incomes above a given level and the logarithm of this given level of income. This observation has been made in many situations and suggests a distribution that decays like a power function, such behaviour characterizes a heavy-tailed distribution.¹⁰ The Pareto CDF is given by

$$F(y; \alpha) = 1 - \left[\frac{y}{y_0} \right]^{-\alpha}, \quad y > y_0 \quad (2)$$

with density

$$f(y; \alpha) = \alpha y^{-\alpha-1} y_0^\alpha \quad (3)$$

If p_y denotes the proportion of the population with incomes greater than or equal to y (for $y \geq y_0$) then we have

$$\log p_y = \log A - \alpha \log y \quad (4)$$

⁹What is now commonly described as *the* Pareto distribution is more precisely referred to as “Pareto type I.” For other, more general forms introduced by Pareto and their relationship to the Pareto type I see Cowell (2011), Kleiber and Kotz (2003).

¹⁰A “heavy-tailed” distribution F is one for which $\lim_{y \rightarrow \infty} e^{\lambda y} [1 - F(y)] = \infty$, for all $\lambda > 0$: it has a tail that is heavier than the exponential.

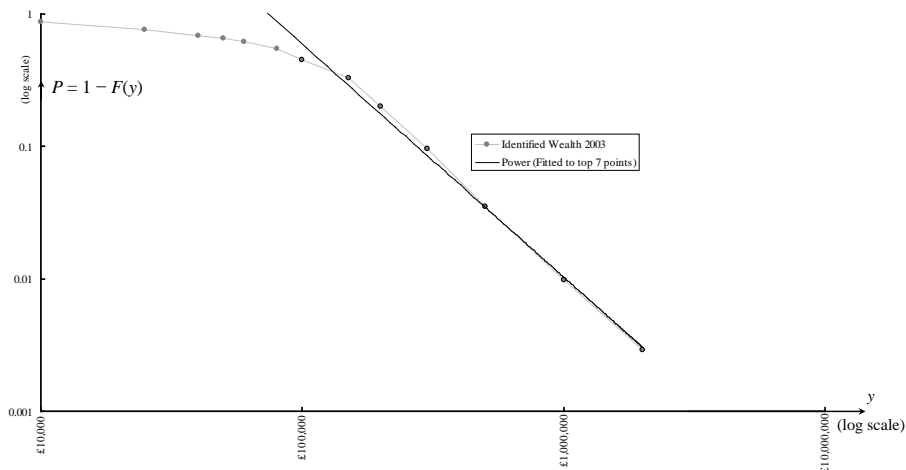


Figure 4: Pareto distribution. UK Identified wealth 2003. Source: Inland Revenue Statistics 2006 Table 13.1

where $A := y_0^\alpha$. The Pareto index α is the elasticity of a reduction in the number of income-receiving units when moving to a higher income class. The larger is the Pareto index, the smaller the proportion of very high-income people. The Pareto distribution often fits wealth distributions and high levels of income well – see Figure 4 – but it is not designed to fit low levels of income. Other distributions have then been proposed in the literature.

3.1.2 Lognormal

Gibrat (1931) highlighted the central place of the lognormal distribution in many economic situations. His *law of proportionate effect* says that if the variation of a variable between two successive periods of time is a random proportion of the variable at the first period, then the variable follows a lognormal distribution.¹¹ He successfully fitted lognormal distributions with many different datasets, as for instance income, food expenditures, wages, legacy, rents, real estate, firm profits, firm size, family size and city size. The lognormal distribution has then been very popular in empirical work¹² and is often appropriate for studies of

¹¹If $X_t - X_{t-1} = \varepsilon_t X_{t-1}$, then $\sum_{i=1}^n \varepsilon_t = \sum_{i=1}^n \frac{X_t - X_{t-1}}{X_{t-1}} \approx \log X_n - \log X_0$. From a central limit theorem (CLT), $\log X_n$ follows asymptotically a Normal distribution.

¹²In addition, it has nice properties related to the measurement of inequality (Cowell 2011), it is closely related to Normal distribution and it fits homogeneous subpopulations quite well

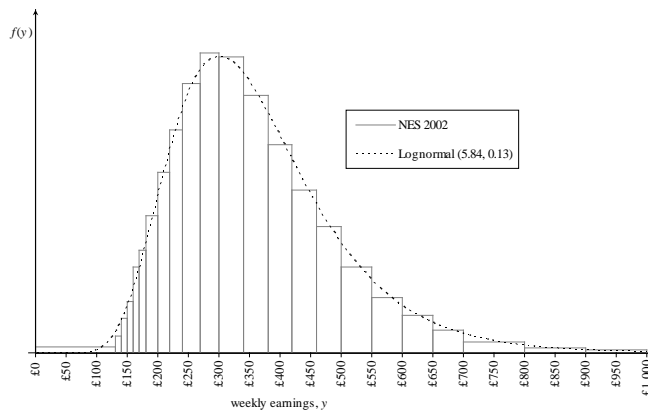


Figure 5: Lognormal distribution. UK Male Manual Workers on Full-Time Adult Rates. Source: New Earnings Survey 2002, Part A Table A35.

wages – see Figure 5. However, the fit of the upper tail of more broadly based income distributions appears to be quite poor. The tail of the lognormal distribution decays faster than the Pareto distribution, at the rate of an exponential function rather than of a power function. It has led to the use of other distributions with two to five parameters to get better fits of the data over the entire distribution.

3.1.3 Generalized Beta

The gamma and Weibull distributions have shown good fit in empirical studies.¹³ The lognormal, gamma and Weibull density functions are two-parameter distributions, they share the property that Lorenz curves do not intersect, contrary to what is observed in several data. To allow intersecting Lorenz curves, three-parameter distributions should be used, as the generalized gamma (GG), Singh-Maddala (SM) and Dagum distributions.¹⁴ As shown by McDonald and Xu (1995), all the previously mentioned distributions are special or limiting

(Aitchison and Brown 1957).

¹³Among others, Salem and Mount (1974) show that the gamma distribution fits better than the lognormal for income data in the United States for the years 1960 to 1969 ; Bandourian, McDonald, and Turley (2003) found the Weibull distribution as the best two-parameter distribution for income distribution in many countries

¹⁴See Stacy (1962), Singh and Maddala (1976), Dagum (1977). The Singh-Maddala and Dagum distributions are also known as, respectively, the Burr 12 and Burr 3 distributions

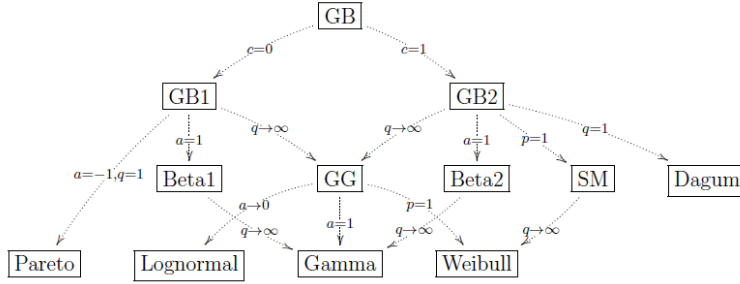


Figure 6: Parametric distributions tree (*Source: Bandourian et al. 2003*)

cases of the five-parameter generalized beta distribution, defined by the following density function:

$$GB(y; a, b, c, p, q) = \frac{|a|y^{ap-1}[1 - (1-c)(y/b)^a]^{q-1}}{b^{ap}B(p, q)[1 + c(y/b)^a]^{p+q}}, \quad (5)$$

for $0 < y^a < b^a/(1-c)$, and is equal to zero otherwise. $B(p, y) := \int_0^1 t^{p-1}(1-t)^{q-1} dt$ is the beta function, $0 \leq c \leq 1$ and b, p and q are positive. Figure 6 shows graphically the relationships between distributions.¹⁵ As an example of the paths through this diagram take the case where $c = 0$ in (5): we find the Generalised Beta of the first kind

$$GB1(y; a, b, p, q) = \frac{|a|y^{ap-1}[1 - (y/b)^a]^{q-1}}{b^{ap}B(p, q)}. \quad (6)$$

Going a stage further, the special case of (6) with $a = 1$ gives the Beta distribution of the first kind

$$B1(y; b, p, q) = \frac{y^{p-1}[1 - y/b]^{q-1}}{b^p B(p, q)}, \quad (7)$$

As an alternative route from (6), setting $a = -1$ and $q = 1$ we obtain $b^p y^{-p-1}$ which, with a change of notation, is clearly the density function of the Pareto type I distribution (3). For more details on continuous univariate distributions, see Johnson et al. (1994) and Kleiber and Kotz (2003).

Income distributions have been extensively estimated with parametric density functions in the literature, see for instance Singh and Maddala (1976),

¹⁵GB1 and GB2 are, respectively, the generalized beta of the first and second kinds introduced by McDonald (1984). Beta1 and Beta2 are, respectively, the beta of first and second kinds. An alternative three-parameter approach that giving a good representation of income distributions in practice is provided by the Pareto-Lévy class (Mandelbrot 1960, Dagsvik et al. 2013); unfortunately, except in a few cases, the probability distributions associated with this class cannot be represented in closed form.

Dagum (1977, 1980, 1983), McDonald (1984), Butler and McDonald (1989), Majumder and Chakravarty (1990), McDonald and Xu (1995), Bantilan et al. (1995), Victoria-Feser (1995, 2000), Brachmann et al. (1996), Bordley et al. (1997), Tachibanaki et al. (1997) and Bandourian et al. (2003). In most of these empirical studies, the generalized beta of the second kind, the Singh-Maddala and the Dagum distributions perform better than other two/three parameter distributions.

3.1.4 Goodness of fit

Goodness-of-fit test statistics are used to test whether a given sample of data is drawn from an estimated probability distribution. They are used to know if an estimated density function is appropriate and fits well the data. Several statistics have been proposed in the literature. The well-known Pearson chi-squared statistic is defined as: $\chi^2 = \sum_{i=1}^m (O_i - E_i)^2 / E_i$, where O_i is the observed percentage in the i^{th} histogram interval, E_i is the expected percentage in the i^{th} histogram interval and m is the number of histogram intervals. This measure summarizes discrepancies between frequencies given by an histogram obtained from the data and those expected from the estimated density function. A statistic not significantly different from zero suggests that the estimated density function fits well the unknown density function that generated the data. In finite sample, this statistic is known to have poor finite sample power properties, that is, to under-reject when the estimated density function is not appropriate, see Stephens (1986). Then the Pearson chi-square test is usually not recommended as a goodness-of-fit test. Empirical Distribution Function (EDF) based statistics perform better. Given a set of observations $\{y_1, y_2, \dots, y_n\}$, the EDF is defined as

$$F^{(n)}(y) = \frac{1}{n} \sum_{i=1}^n \iota(y_i \leq y) \quad (8)$$

where $\iota(\cdot)$ is the indicator function defined in (1). When the observations are independently and identically distributed it is a consistent estimator of the Cumulative Distribution Function (CDF). EDF-based statistics measure the discrepancy between the EDF and the estimated CDF. They are not sensitive to the choice of the histogram's bins as in the Pearson chi-squared statistic. For instance, the Kolmogorov-Smirnov statistic is equal to

$$\sup_y |F^{(n)}(y) - F(y, \hat{\theta})| \quad (9)$$

where $F(y, \hat{\theta})$ is an estimated CDF from a parametric family function with parameters θ . Other statistics can be expressed as

$$n \int_{-\infty}^{\infty} \left(F^{(n)}(y) - F(y, \hat{\theta}) \right)^2 w(y) dF(y, \hat{\theta}), \quad (10)$$

where $w(y)$ is a weighting function. The Cramér-von Mises statistic corresponds to the special case $w(y) = 1$, while the Anderson-Darling statistic puts more

weight in the tails, with $w(y) = [F(y, \hat{\theta})(1 - F(y, \hat{\theta}))]^{-1}$. In finite sample, the Anderson-Darling test statistic outperforms the Cramér-von-Mises statistic, which in turns outperform the Kolmogorov-Smirnov test statistic, see Stephens (1986).

In distributional analysis goodness-of-fit test statistics and inequality measures do not usually share the same intellectual foundation. The former are based on purely statistical criteria while the latter are typically based on axiomatics that may be associated with social-welfare analysis or other formal representations of inequality in the abstract. Cowell et al. (2011) developed a family of goodness-of-fit tests founded on standard tools from the economic analysis of income distributions, defined as:

$$G_\xi = \frac{1}{(\xi^2 - \xi)} \sum_{i=1}^n \left[\left[\frac{u_i}{\mu_u} \right]^\xi \left[\frac{2i}{n+1} \right]^{1-\xi} - 1 \right], \quad (11)$$

where $\xi \in \mathbb{R} \setminus \{0, 1\}$ is a parameter, $u_i = F(y_{(i)}; \hat{\theta})$, $\mu_u = \frac{1}{n} \sum_{i=1}^n u_i$ and $y_{(i)}$ is the i th order statistic (the i th smallest observation). G_ξ is closely related to Generalized Entropy (GE) inequality indices – see equations (49)-(51) below. GE inequality measures are divergence measures between the EDF and the most equal distribution, where everybody gets the same income. They tell us how far an empirical distribution is from the most equal distribution.¹⁶ Goodness of fit test statistics G_ξ are divergence measures between the EDF and an estimated parametric CDF. They tell us how far an empirical distribution is from an estimated parametric distribution.¹⁷ It has excellent size and power properties as compared with other, commonly used, goodness-of-fit tests. It has the further advantage that the profile of the G_ξ statistic as a function of ξ can provide valuable information about the nature of the departure from the target family of distributions, when that family is wrongly specified.

3.2 Kernel method

3.2.1 From histogram to kernel estimator

Histograms are the most widely used nonparametric density estimators. However, they have several drawbacks that kernel density method allows us to handle.

¹⁶See Cowell et al. (2013) for an extension to the choice of any other “reference” distribution, giving for instance inequality measures telling us how far an empirical distribution is from the most *unequal* distribution.

¹⁷The term in the first bracket in (11) is related to the CDF, $u_i/\mu_u = F(y_{(i)}, \hat{\theta})/\mu_u$, while the term in the second bracket is related to the EDF, $2i/(n+1) = v_i/\mu_v$ where $v_i = \hat{F}^{(n)}(y_{(i)}) = i/n$ and $\mu_v = n^{-1} \sum_{i=1}^n v_i = \sum_{i=1}^n i/n^2 = (n+1)/(2n)$. Using the most equal distribution and replacing u_i and v_i by their q -quantile counterparts, $F^{-1}(q_i = i/n, \hat{\theta}) = \hat{\mu}_y$ and $\hat{F}^{(n)-1}(q_i = i/n) = y_{(i)}$, give GE measures. Note that u_i and v_i have bounded support ($u_i, v_i \in [0, 1]$), a property required to show that the asymptotic distribution of G_ξ exists, see Davidson (2012).

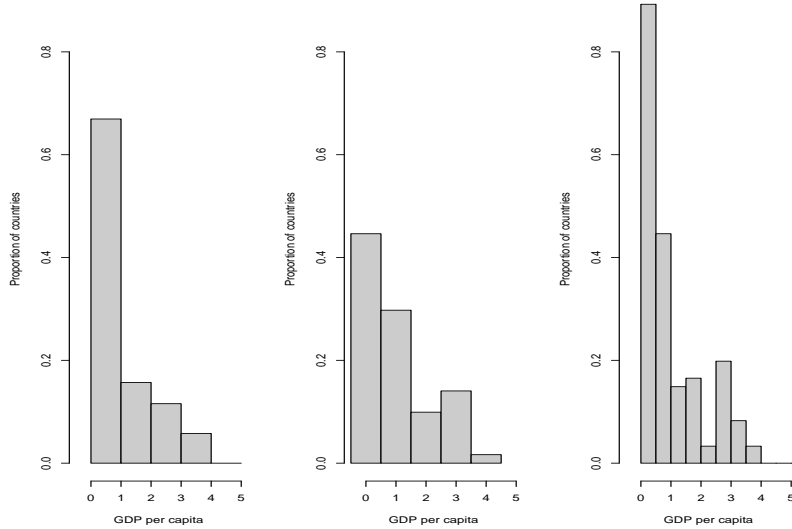


Figure 7: Histogram’s sensitivity to the position and the number of bins

Figure 7 illustrates several problems arising with histograms, using GDP per capita in 121 countries in 1988 (solid line). The left plot is given with 5 bins of the same length between 0 and 5. The middle plot is similar but the position of the bins changed, they are between -0.5 and 4.5. The two pictures are very different, even if they estimate the same distribution. The left panel shows a unimodal distribution, while the middle panel shows a bimodal distribution. Histograms are then sensitive to the point at which we start drawing bins. The right panel is given with 10 bins of the same length between 0 and 5. Once again, it gives a different picture of the same distribution. Histograms are then sensitive to the number of bins used, which is also relatively arbitrary. Moreover, and most obviously, the pictures given by histograms provide discontinuities at the edge of each bin, which may not be an appropriate property of the true underlying distribution.

To avoid having to make arbitrary choice on the position and the number of bins, we can use intervals that may overlap, rather than being separate from each other. The principle here is to estimate a density function at one point by counting the number of observations which are close to this evaluation point. For a sample of n observations, y_1, \dots, y_n , the naive density estimator is given by:

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^n \iota \left(y - \frac{h}{2} < y_i < y + \frac{h}{2} \right), \quad (12)$$

where h is the width of the intervals and $\iota(\cdot)$ is the indicator function (1). In

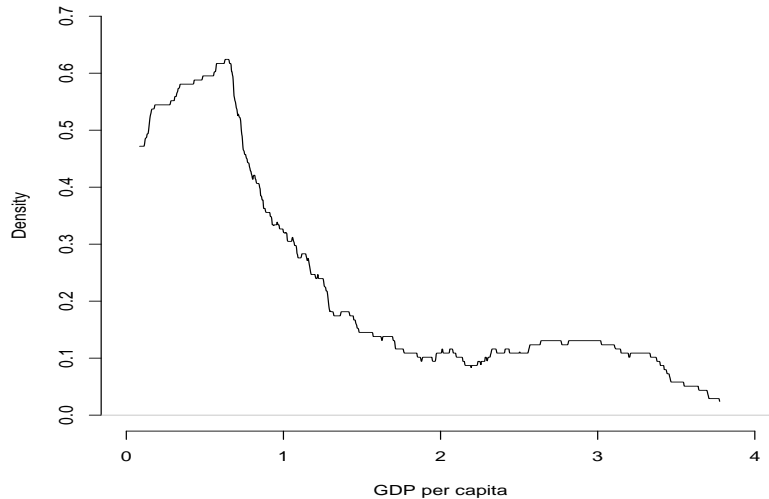


Figure 8: Naive estimator of GDP per capita

this equation, the estimate of the density at point y is given by the proportion of observations which are within a distance of $h/2$ or less from point y . The global density is obtained by sliding this window of width h along all of the evaluation points.

Figure 8 presents the naive estimation of the density of GDP per capita across different countries in 1988. Compared to histograms, the naive estimator reveals much more detail about the curvature of the density function. However, discontinuities are still present.

The kernel estimator is a generalization of the naive estimator which allows us to overcome the problem of differentiability at all points. The discontinuity problem comes from the indicator function $\iota(\cdot)$, which allocates a weight of 1 to all of the observations which belong to the interval centered on y , and zero weight to the other observations. The principle of kernel estimation is simple: rather than giving all observations in the interval the same weight, the allocated weight is greater the closer the observation is to y . The transition from 1 to 0 in the weights is then carried out gradually, rather than abruptly. The kernel estimator is obtained by replacing the indicator function by a kernel function $K(\cdot)$:

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{y - y_i}{h}\right). \quad (13)$$

In order for $\hat{f}(y)$ to conserve the properties of a density function, the integral of the kernel function over the entire space has to be equal to one. Any probability distribution can then be used as kernel function. The Gaussian and the

Epanechnikov distributions are two kernels commonly used in practice.¹⁸ Kernel density estimation is known to be sensitive to the choice of the bandwidth h , while it is not really affected by the choice of the kernel function.

3.2.2 Bandwidth selection

The question of which value of h is the most appropriate is particularly a thorny one, even if automatic bandwidth selection procedures are often used in practice. Silverman’s rule of thumb is mostly used, which is defined as follows:¹⁹

$$\hat{h}_{opt} = 0.9 \min \left(\hat{\sigma}; \frac{\hat{q}_3 - \hat{q}_1}{1.349} \right) n^{-\frac{1}{5}}, \quad (14)$$

where $\hat{\sigma}$ is the standard deviation of the data, and \hat{q}_3 and \hat{q}_1 are respectively the third and first quartiles calculated from the data. This rule boils down to using the minimum of two estimated measures of dispersion: the variance, which is sensitive to outliers, and the interquartile range. It is derived from the minimization of an approximation of the mean integrated squared error (MISE), a measure of discrepancy between the estimated and the true densities, where the Gaussian distribution is used as a reference distribution. This rule works well in numerous cases. Nonetheless, it tends to over-smooth the distribution when the true density is far from the Gaussian distribution, as multimodal and highly skewed. Figure 2 is a kernel density estimation of GDP per capita with Silverman’s rule-of-thumb bandwidth selection. It appears as a smoothed version of the naive estimator in Figure 8. Several other data-driven methods for selecting the bandwidth have been developed such as cross-validation (Stone 1974, Rudemo 1982, Bowman 1984) and plug-in methods (Sheather and Jones 1991, Ruppert et al. 1995), among others.

Rather than using a Gaussian reference distribution in the approximation of the MISE, the *plug-in* approach consists of using a prior non-parametric estimate, and then choosing the h that minimizes this function. This choice of bandwidth does not then produce an empirical rule as simple as that proposed by Silverman, as it requires numerical calculation. For more details, see Sheather and Jones (1991).

Rather than minimizing the MISE, the underlying idea of *cross-validation by least squares* is to minimize the integrated squared error (ISE). In other words, we use the same criterion, but not expressed in terms of expectations. The advantage of the ISE criterion is that it provides an optimal formula for h for a given sample. The counterpart is that two samples drawn from the same density will lead to two different optimal bandwidth choices. The ISE solution consists in finding the value of h that minimizes: $\text{ISE}(h) = \int [\hat{f} - f]^2 dy = \int \hat{f}^2 dy - 2 \int \hat{f} f dy + \int f^2 dy$, where, for simplicity, f and \hat{f} correspond to $f(y)$ and $\hat{f}(y)$.

¹⁸The Gaussian kernel corresponds to the choice of the standard Normal distribution: $K(x) = e^{-x^2/2}/\sqrt{2\pi}$. Epanechnikov (1969) proposed a second-degree polynomial, adjusted to satisfy the properties of a density function: $K(x) = 3(1 - x^2/5)/(4\sqrt{5})$ if $|x| < \sqrt{5}$ and 0 otherwise.

¹⁹See equation (3.31), page 48, in Silverman (1986).

The last term in this equation does not contain h and thus plays no role in the minimization. Furthermore, the term $\int \hat{f} f \, dy$ is exactly $E(\hat{f})$. Let \hat{f}_{-i} be the estimator of the density based on the sample containing all of the observations except for y_i . An unbiased estimator of $E(\hat{f})$ is given by $n^{-1} \sum_{i=1}^n \hat{f}_{-i}$. The minimization of the ISE criterion thus requires us to minimize the following expression:

$$\text{CV}(h) = \int \hat{f}^2(y) \, dy - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(y_i).$$

This method is also called *unbiased cross-validation*, as $\text{CV}(h) + \int f^2 \, dy$ is an unbiased estimator of MISE. The value of h which minimizes this expression converges asymptotically to the value that minimizes the MISE.

3.2.3 Adaptive kernel estimator

In the kernel density estimation presented above, the bandwidth remains constant at all points where the distribution is estimated. This constraint can be particularly onerous when the concentration of data is markedly heterogeneous in the sample. There would advantages to use narrower bandwidth in dense parts of the distribution (the middle) and wider ones in the more sparse parts (the tails). The adaptive kernel estimator is defined as follows:

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h\lambda_i} K\left(\frac{y - y_i}{h\lambda_i}\right),$$

where λ_i is a parameter that varies with the local concentration of the data. An estimate of the density at point y_i , denoted by $\tilde{f}(y_i)$, measures the concentration of the data around this point: a higher value of $\tilde{f}(y_i)$ denotes a greater concentration of data, while smaller values indicate lighter concentrations. The parameter λ_i can thus be defined as being inversely proportional to this estimated value: $\lambda_i = [g/\tilde{f}(y_i)]^\theta$, where g is the geometric mean of $\tilde{f}(y_i)$ and θ is a parameter that takes on values between 0 and 1.²⁰ The parameter λ_i is smaller when the density is greater (notably towards the middle of the distribution), and larger when the density is lighter (in the tails of the distribution).

Figure 9 presents the adaptive kernel density estimation of GDP per capita across different countries in 1988. Compared to the simple kernel density estimation, with fixed-bandwidth (dashed line), the first mode is higher and the second mode lower.

Several empirical studies on income distribution have used kernel density estimation, among other see Marron and Schmitz (1992), Jenkins (1995), Cowell et al. (1996), Daly et al. (1997), Quah (1997), Burkhauser et al. (1999), Bourguignon and Morrisson (2002), Pittau and Zelli (2004), Jenkins and Van Kerm (2005) and Sala-i-Martin (2006).

²⁰In practice, an initial fixed-bandwidth kernel estimator can be employed as $\tilde{f}(y_i)$, with $\theta = 1/2$ and λ obtained with Silverman's rule of thumb.

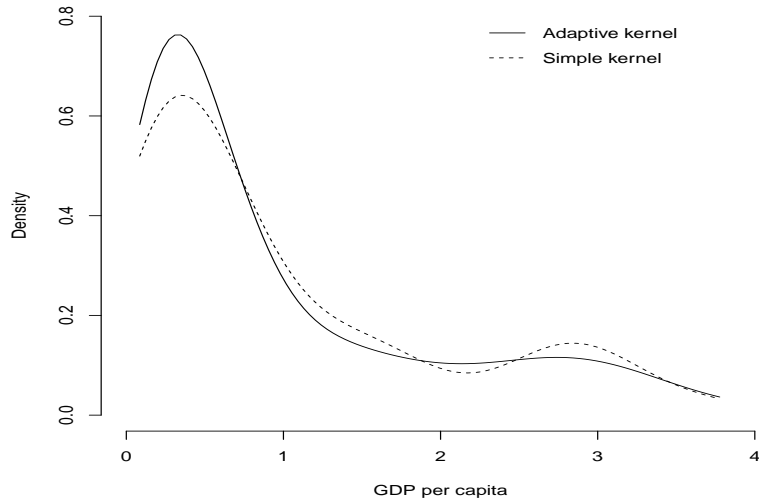


Figure 9: Adaptive kernel estimation of GDP per capita

3.2.4 Multivariate and conditional density

The extension to the multivariate case is straightforward. The joint density of two variables y and x , for which we have n observations, can be estimated with a bivariate kernel function

$$\hat{f}(y, x) = \frac{1}{nh_1h_2} \sum_{i=1}^n K\left(\frac{y_i - y}{h_1}; \frac{x_i - x}{h_2}\right), \quad (15)$$

which is equivalent to the product of two univariate kernels in the Gaussian case. The extension to the d -dimensional case is immediate, via the use of multivariate kernels in d -dimensions. Scott (1992) extends the Silverman's rule of thumb as follows: $h_j = n^{-1/(d+4)}\hat{\sigma}_j$, where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the sample standard deviations of, respectively, y and x . In practice, kernel density estimation is rarely used with more than 2 dimensions. With three or more dimensions, not only may the graphical representation be problematic, but also the precision of the estimation. Silverman (1986) shows that the number of observations required to guarantee a certain degree of reliability rises explosively with the number of dimensions. This problem is known as the *curse of dimensionality*.

A conditional density function is equal to the ratio of a joint distribution and a marginal distribution, $f(y|x) = f(x, y)/f(x)$. A kernel conditional density

estimation is then given by

$$\hat{f}(y|x) = \frac{\frac{1}{h_1 h_2} \sum_{i=1}^n K\left(\frac{y_i - y}{h_1}, \frac{x_i - x}{h_2}\right)}{\frac{1}{h_3} \sum_{i=1}^n K\left(\frac{x_i - x}{h_2}\right)}. \quad (16)$$

When several conditional variables are considered, $x = \{x_1, \dots, x_k\}$, the bandwidth selection obtained by cross-validation can mitigate the curse of dimensionality problem if some of them are irrelevant (Hall et al. 2004, Fan and Yim 2004). Several recent studies have been focused on conditional analysis in a nonparametric framework. To evaluate policy effects, the impact of a counterfactual change in the distribution of some covariates on the unconditional distribution of some variable of interest has been investigated in DiNardo et al. (1996), Donald et al. (2000), Chernozhukov et al. (2009), Rothe (2010), Donald et al. (2012). For more details on kernel density estimation, see Silverman (1986), Paul (1999), Li and Racine (2006), Ahamada and Flachaire (2011).

3.3 Finite-mixture models

3.3.1 A group decomposition approach

A population can be decomposed into several distinct groups in many different ways. The density function of the population is then equal to the sum of the densities associated with each of the different groups. Let us consider κ groups, with the density function of each group being parametric, $f_k(y; \theta_k)$ for $k = 1, \dots, \kappa$ where θ is a set of parameters. Then, the density function of the population can be written,

$$f(y; \theta) = \sum_{k=1}^{\kappa} \pi_k f_k(y; \theta_k) \quad (17)$$

where π_k is the proportion of the population belonging to subgroup k . The conditions $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{\kappa} \pi_k = 1$ are required to guarantee that the population density integrates to one over the support. A density estimation by mixture models is obtained by replacing the unknown parameters by estimated parameters. In finite-mixture models, the group to which each individual belongs is not observed.²¹ They thus allows us to capture the effect of unobserved heterogeneity. They can also be used for classification purpose. Bayes' theorem allows us to deduce the *a posteriori* probability that an observation i belongs to the group k :

$$\pi_{ik} = \frac{\pi_k f_k(y_i; \theta_k)}{\sum_{k=1}^{\kappa} \pi_k f_k(y_i; \theta_k)}. \quad (18)$$

Replacing the unknown parameters by consistent estimates, these individual probabilities can be used to classify the observations into the different groups.

²¹When the groups are known and also the densities associated to each groups, the mixture model is entirely parametric and can be estimated by maximum likelihood (see section 3.1)

The estimation of a density by a mixture model allows us to bring out the link between parametric and non-parametric estimation. If we consider one single group ($\kappa = 1$), then the mixture models amount to just one parametric function. Adding additional groups allows us to estimate more complicated densities, which cannot be modeled with one sole group; adding more groups allows us to reflect the heterogeneity of the population. Mixture models thus permit much greater modeling flexibility. In the extreme case, where we have as many groups as we do observations ($\kappa = n$), the mixture is equivalent to the estimation of a density by kernel methods (see section 3.2).²² For values of κ between 1 and the size of the sample n , the mixture model can thus be seen as a semi-parametric compromise between parametric estimation and non-parametric kernel estimation. The parametric aspect is reflected in the fact that the density is expressed as a sum of parametric density functions; the non-parametric aspect is captured by the presence of a number of different groups.

The theory of mixture models tells us that, under regularity conditions, any probability density can be consistently estimated as a mixture of Normal distributions.²³ Figure 10 depicts a number of different mixtures of two Normal distributions, of which the density can be written as $\pi_1\phi(y; \mu_1, \sigma_1) + \pi_2\phi(y; \mu_2, \sigma_2)$, where $\phi(\cdot)$ is the density of the Normal distribution, with mean μ_k and variance σ_k , for $k = 1, 2$. The global density and the two individual components are represented in the same figure.²⁴ From global densities (solid lines), we can see that a wide variety of densities can be represented by a mixture of only two Normal distributions, as top flat (panel a), bimodal (panel b), skewed (panel c) and thick upper-tailed (panel d) distributions. Many further examples can be provided to illustrate the very wide variety of distributions that can be characterized by a mixture of κ Normal distributions: see, amongst others, Marron and Wand (1992). All of these examples reveal the great flexibility of finite-mixture models in estimating densities.

3.3.2 Number of components and number of groups

For a given number of κ , we can estimate the unknown parameters by maximum likelihood.²⁵ The number κ , known as the number of components, can be selected by minimizing a criterion, as the Bayesian Information Criterion,

$$\text{BIC} = -2\hat{\ell} + \#\text{param} \log n, \quad (19)$$

where $\hat{\ell}$ is the estimated log-likelihood, $\#\text{param}$ is the number of parameters to estimate and n is the number of observations. If the main concern is the best

²²If $K = n$ and $\pi_1 = \dots = \pi_K = 1/n$, equation (17) is then equivalent to equation (13) where the $f_k(\cdot)$ function is the kernel function $K(\cdot)$.

²³See Escobar and West (1995), Ferguson (1983), Titterington, Makov, and Smith (1985), McLachlan and Peel (2000), Ghosal and van der Vaart (2001).

²⁴In panel (a), $\pi_1 = \pi_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 0$, $\mu_2 = 2$. In panel (b), $\pi_1 = \pi_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 0$, $\mu_2 = 4$. In panel (c), $\pi_1 = 0.6$, $\pi_2 = 0.4$, $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 0$, $\mu_2 = 2$. In panel (d), $\pi_1 = 0.75$, $\pi_2 = 0.25$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\mu_1 = 0$, $\mu_2 = 2$.

²⁵The EM algorithm of Dempster et al. (1977) is often used. Bayesian methods can also be employed, see Robert and Casella (2005), Frühwirth-Schnatter (2006).

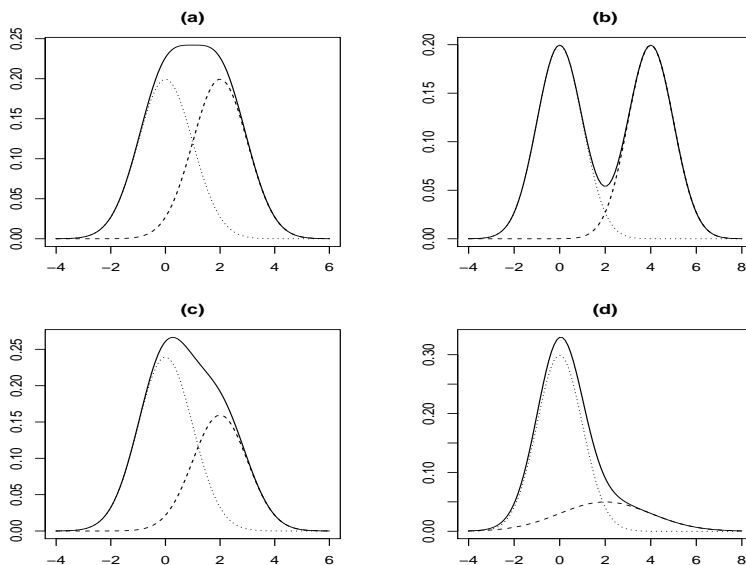


Figure 10: Mixtures of two Normal distributions

fit of the overall density, this selection criterion is appropriate. However, if the main concern is the detection of distinct groups, the choice of κ is less simple. Indeed, there is no automatic correspondence between the choice of κ and the number of underlying groups in the population. For instance, the panel (d) in Figure 10 shows that the second component is required to fit a thick upper tail, but it does not clearly identify a distinct group from the first component. Indeed, the two distributions of the groups intersect a lot. Here, the number of component κ is not necessarily equivalent to the number of groups. It illustrates that the definition of what constitutes a distinct groups and its detection can be a difficult task in finite-mixture models.²⁶

Figure 11 shows kernel density estimation (on the left) and estimation by a mixture of Lognormal distributions (on the right) of the income distribution in the United Kingdom in 1973. A lognormal distribution is heavy-tailed, in the sense that it has an upper-tail that is heavier than an exponential. Then, it is more appropriate to use finite mixture of lognormals rather than finite mixture of normals to fit income distributions, which are typically heavy-tailed. The estimation of the density by a mixture of Lognormal distributions is obtained from an estimation of the density of log-incomes by a mixture of Normal distributions.²⁷ The value of the bandwidth given by Silverman's empirical rule ($h = 0.08559$) allows us to reproduce the kernel density estimation results in

²⁶A more appropriate method could be to test the number of modes of the distribution, see Ray and Lindsay (2005)

²⁷From $f(x; \Theta) = \sum_{k=1}^K \pi_k \phi(x; \mu_k, \sigma_k)$, we have $f(y; \Theta) = \sum_{k=1}^K \pi_k \Lambda(y; \mu_k, \sigma_k)$ where $x = \log y$.

Marron and Schmitz (1992). Kernel estimation with a smaller value of h (0.01) are overlaid in the same figure. The comparison of the two estimators reveals that the results differ significantly: with $h = 0.08559$, the first mode is smaller than the second, while with $h = 0.01$ the reverse holds. This confirms that the kernel estimation with Silverman’s rule of thumb does indeed tend to over-smooth the function when the underlying distribution is multimodal and highly skewed distribution (see section 3.2). In our example, the Silverman selection choice tends to flatten out considerably the first mode relative to the second. The right panel shows the density estimation with a mixture of lognormal distributions, obtained by minimizing the BIC criterion. The overall distribution appears to be a smoothed representation of the kernel density estimation with $h = 0.01$. In addition, the mixture estimation identifies three separate components. The first and the third components do not overlap a lot, they can be associated to two distinct modes. The second component overlaps to a considerable extent with the third, and to a lesser extent with the first. The presence of this second component allows a better fit of the right-hand side tail of the distribution, but cannot be clearly associated to a distinct group.

A very few empirical studies have used finite-mixture models to estimate income distributions. Flachaire and Nuñez (2007) studied the distribution of household income in the United Kingdom with a mixture of lognormal distributions. Pittau and Zelli (2006) and Pittau et al. (2010) studied the evolution of per capita income distributions across EU regions and countries. Chotikapanch and Griffiths (2008) estimate the Canadian income distribution using a mixture of Gamma distributions. Lubrano and Ndoye (2011) model the income distribution using a Bayesian approach and a mixture of lognormal densities.²⁸

3.3.3 Group profiles explanation

In addition to the estimate of a density function of any form, finite-mixture estimation can be used to explain the profiles of the different groups underlying the overall population. This can be done by introducing covariates in the probabilities π_k :

$$f(y|z; \Theta) = \sum_{k=1}^{\kappa} \pi_k(z, \alpha_k) f_k(y; \theta_k), \quad (20)$$

where $z = \{z_1, \dots, z_l\}$ is a vector of l observed variables and $\alpha_k = \{\alpha_{1k}, \dots, \alpha_{lk}\}$ is a vector of l unknown parameters. This model defines a conditional density function, which takes into account directly the fact that the probability of group membership may be a function of individual characteristics (a white collar worker has a greater probability of belonging to the group of the richest households than do a blue collar worker). As well as the non-parametric estimation of the density and the decomposition into different groups, covariates

²⁸Paap and Van Dijk (1998) considered mixtures of two distributions, using Normal, Log-normal, Gamma and Weibull distributions. However, their approach is entirely parametric, with the number of components and the densities of each groups fixed *a priori*.

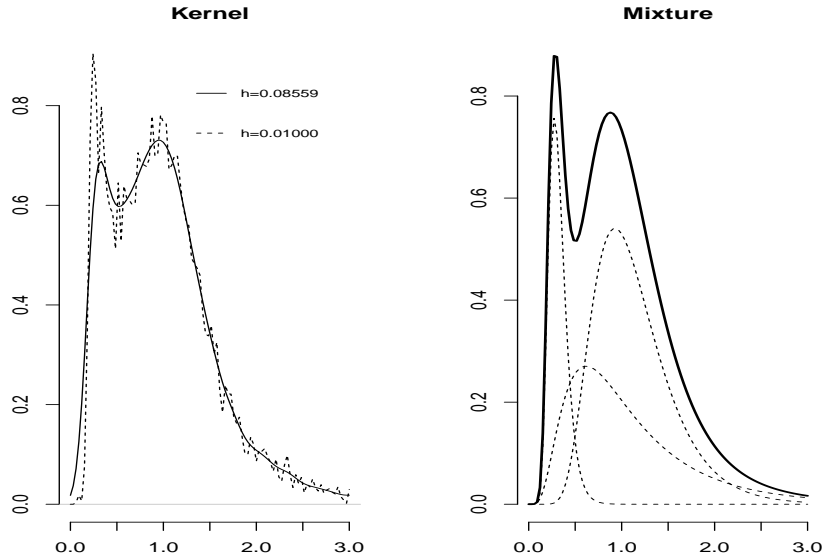


Figure 11: Income distribution in the United Kingdom in 1973 (incomes normalised by contemporary mean)

also explain the variability between groups. The relationship between the probabilities π_k and the covariates z can be specified with an ordered logit/probit or multinomial regression model and the unconditional density can be obtained as follows:

$$f(y; \Theta) = \sum_{k=1}^{\kappa} \bar{\pi}_k \phi(y; \mu_k, \sigma_k) \quad \text{with} \quad \bar{\pi}_k = \frac{1}{n} \sum_{i=1}^n \pi_k(z_i, \alpha_k),$$

where z_i represents the vector of characteristics of the i^{th} observation and n is the number of observations. In other words, $\pi_k(z_i, \alpha_k)$ is the probability that the individual i with characteristics z_i belong to the group k . For more details, see Ahumada and Flachaire (2011).

Figure 12 reproduces the results of the estimation of the distribution of household income in the United Kingdom in 1979 and 1988, obtained in Flachaire and Nuñez (2007), by a mixture of Lognormal distributions.²⁹ The decomposition into groups of the mixture estimator does emphasize clear changes over time, that would be difficult to see from the comparisons of the overall distributions. The analysis by groups shows that, in 1988, a small separate group had formed to the extreme left of the distribution, while that situated to the far right of the distribution had grown in size. Table 3 reproduces the estimated

²⁹The analysis here uses the same data as Marron and Schmitz (1992), with the exception that the incomes are normalized via an equivalence scale in order to account for differences in household size.

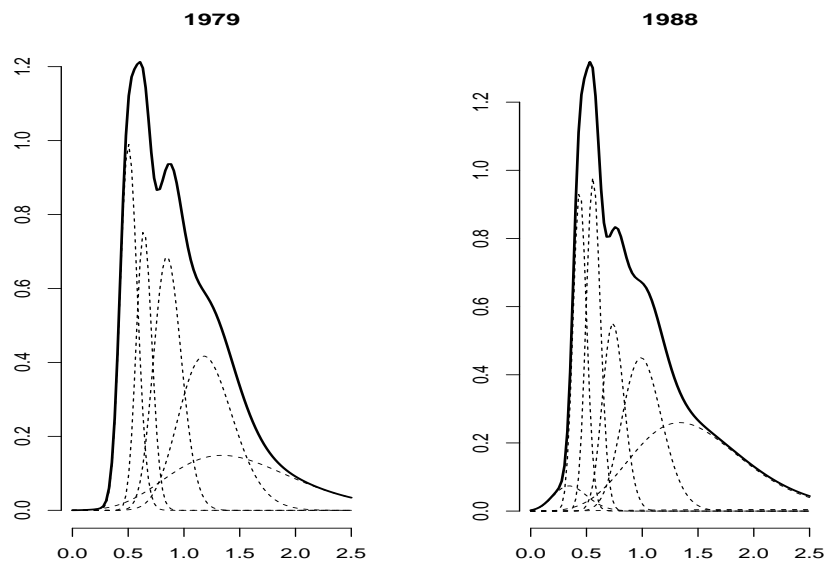


Figure 12: Income distribution in the United Kingdom in 1979 and 1988 (incomes normalised by contemporary mean)

coefficients (with standard errors in parentheses) associated to the following covariates: z_1 for a retired household, z_2 for single-parent families, z_3 for households where all of the adults work, z_4 if no adult works in the household (in a non-retired household), and z_5 for the number of children.

	z_1	z_2	z_3	z_4	z_5
1979	-1.77 (0.059)	-0.67 (0.106)	0.61 (0.050)	-1.16 (0.086)	-0.44 (0.020)
1988	-1.33 (0.058)	-0.69 (0.106)	0.78 (0.053)	-1.44 (0.068)	-0.35 (0.022)

Table 3: Coefficient estimates of covariates

An ordered probit model is used to specify the relationship between the probabilities and the covariates. If a coefficient is positive (negative), then the position of an observation with the associated variable moves to the right (left) of the distribution as the variable z_l increases. On the other hand, a value of α_l which is not significantly different from zero indicates that the characteristic z_l does not help us to explain the decomposition of the sample into the different groups. From the results, we can see that Retired (z_1) and Non-working (z_4) households are more likely to be found towards the bottom of the income distribution, and households where all adults work (z_3), on the contrary, are more likely to be found towards the right of the distribution. In addition, the position of retired households has improved over this period, while that of households where no-one works has deteriorated. These results emphasize the usefulness of

Variables	Linear models		Mixture model	
	Men	Women	Group 1 ($\pi_1 = 0.406$)	Group 2 ($\pi_2 = 0.594$)
<i>Explanatory Variables</i>				
Constant	0.66194*	0.2225	0.67517*	0.34909*
Education	0.07941*	0.10915*	0.08202*	0.09844*
Experience	0.04484*	0.02597*	0.05147*	0.02590*
(Experience) ²	-0.00066*	-0.00038*	-0.00078*	-0.00040*
<i>Concomitant Variables</i>				
Constant			-	-0.49423*
Female			-	9.03215*
Union Member			-	-8.16128 [†]

* $p < 0.05$, [†] $p = 0.057$

Table 4: Mincer Earnings Equations

mixture models, which yield an overall picture of the distribution of income and how this has changed over time, with richer results than those obtained from other commonly-employed techniques.

3.3.4 Finite mixture of regressions

Covariates have been introduced in the probabilities to characterized group profiles. They can be also introduced into the modeling of the densities in each of the groups, leading us to consider mixture of regression models. Let us consider a mixture of Normal distributions with variance σ_k^2 and mean being conditional on some covariates, $\mu_k = x\beta_k$, which can be written as:

$$f(y|x; \Theta) = \sum_{k=1}^{\kappa} \pi_k \phi(y|x; \beta_k, \sigma_k). \quad (21)$$

If there are two groups ($\kappa = 2$), it remains to consider the following model:

$$\begin{aligned} \text{Group 1: } & y = x\beta_1 + \varepsilon_1, & \varepsilon_1 & \sim N(0, \sigma_1^2), \\ \text{Group 2: } & y = x\beta_2 + \varepsilon_2, & \varepsilon_2 & \sim N(0, \sigma_2^2), \end{aligned}$$

where ε_1 and ε_2 are independent and identical Normally-distributed error terms within each group, with variances of σ_1^2 and σ_2^2 respectively. In this model, we consider that the population is composed of two different groups, for which the relationship between the dependent and explanatory variables is different, and the observations come from the different groups in the population in unknown proportions. This specification would be particularly appropriate if we assume that the marginal impact of covariates may be different in each of the groups, as suggested in Figure 3. Covariates could be introduced at the same time in the probabilities, to explain group profiles.

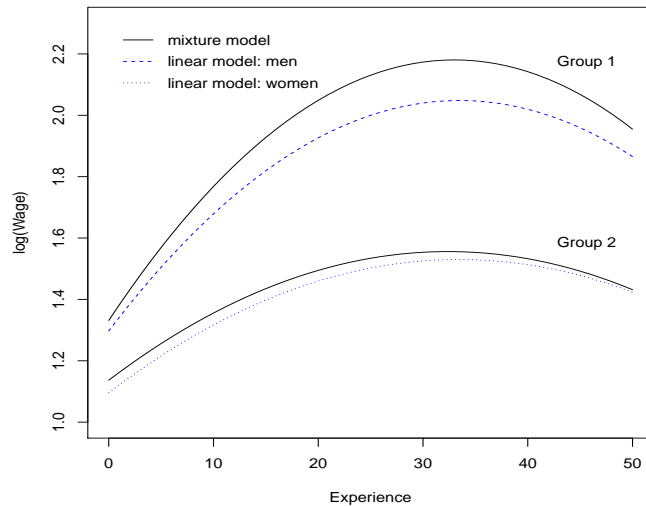


Figure 13: The relationship between earnings and labor-market experience

To illustrate, consider a simple Mincer earnings equation, which explains the logarithm of an individual’s earnings by their number of years of education and number of years of labor-market experience. One way of testing for earnings differences between men and women is to test if the parameters of the earnings equation are statistically significantly different between the two groups of individuals, via a Chow test (Chow 1960). Table 4 shows OLS estimation results from linear regression models for the groups of men and of women (column 1 and 2), using data from a household survey carried out by the US census Bureau in May 1985.³⁰ A Chow test, equals to 14.19, rejects the null hypothesis that the two sets of coefficients are identical. As the dependent variable is the log of earnings, the estimation results show that one additional year of education increases earnings by around 7.9% for men and 10.9% for women on average. The earnings profiles as a function of labor-market experience are different between the two groups. These are traced out in Figure 13 for eight years of education, from which we can see that the gender gap is sharply increasing with labor-market experience during the first 30-years.³¹

In the linear model approach, we define *a priori* two groups of individuals - men and women. To the opposite, in a mixture model approach, we do not specify *a priori* the groups, we let the data identifying homogeneous groups with

³⁰The data come from Chapter 5 of Berndt (1990)

³¹The curve for the group of men corresponds to the polynomial $y = 0.66194 + 8 * 0.07941 + 0.04484 * x - 0.00066 * x^2$ and that for the group of women to $y = 0.22254 + 8 * 0.10915 + 0.02597 * x - 0.00038 * x^2$. The gap between the two curves widens with experience at first, before narrowing again after around 30 years of labor-market experience.

respect to the relationship between the dependent and explanatory variables. Table 4 shows estimation results from a mixture model (columns 3 and 4), the BIC criterion suggests that there is two groups. The estimation results show that one additional year of education increases earnings by around 8.2% for individuals in the first group and 9.8% for individuals in the second, on average. The impact of experience on earnings are traced out in Figure 13 for eight years of education (solid lines).³² From this figure, the gap between the two groups is much larger than the gap between the groups of men and of women obtained from linear models.

The use of concomitant variables in the mixture model allows us to characterize the profile of the groups. Two dummy variables are taken into account as concomitant variables, the first is for the individual being a woman (Female), and the second for the individual being a union member (Union Member). In Table 4, column 4, the positive and significant coefficient on the “Female” variable indicates that women are more likely to belong to the second group than to the first group. The negative significant coefficient on union member shows that unionized workers are less likely to belong to the second than to the first group.³³ A classification shows that 96.3% of women belong to group 2, while the analogous percentage of men is only 19%.³⁴ Equally, the percentage of union members classified in group 1 is 80.2%. Last, the results of this analysis suggest that, for the vast majority of women, the relationship between earnings and experience is much flatter than that for most men and union members, holding everything else equal. The gap obtained is much larger than those obtained by considering all men and all women in two distinct groups.

For more details on mixture models of regression, see McLachlan and Peel (2000), Frühwirth-Schnatter (2006), Ahamada and Flachaire (2011).

3.4 Finite sample properties

In this section, we study the quality of the fit of nonparametric density estimation in finite sample. To assess the quality of the density estimation, we need to use of a distance measure between the estimated density and the true density. We use the mean integrated absolute errors (MIAE) measure,

$$\text{MIAE} = E \left(\int_0^\infty |\hat{f}(y) - f(y)| dy \right). \quad (22)$$

In our experiments, data are generated from two unimodal distributions: Lognormal, $\Lambda(y; 0, \sigma)$, and Singh-Maddala distributions, $\text{SM}(y; 2.8, 0.193, q)$. We also use a bimodal distribution: a mixture of two Singh-Maddala distributions: $\frac{2}{5} \text{SM}(y; 2.8, 0.193, 1.7) + \frac{3}{5} \text{SM}(y; 5.8, 0.593, q)$, plotted in Figure 14.

³²The curve for the first group corresponds to the polynomial $y = 0.67517 + 8 * 0.08202 + 0.05147 * x - 0.00078 * x^2$, and that of the second to $y = 0.34909 + 8 * 0.09844 + 0.02590 * x - 0.0004 * x^2$.

³³This coefficient is significant at the one per cent level according to a LR test.

³⁴An individual is assigned to a group when his individual’s *a posteriori* probability to belong to this group is higher than the probabilities to belong to other groups, see (18).

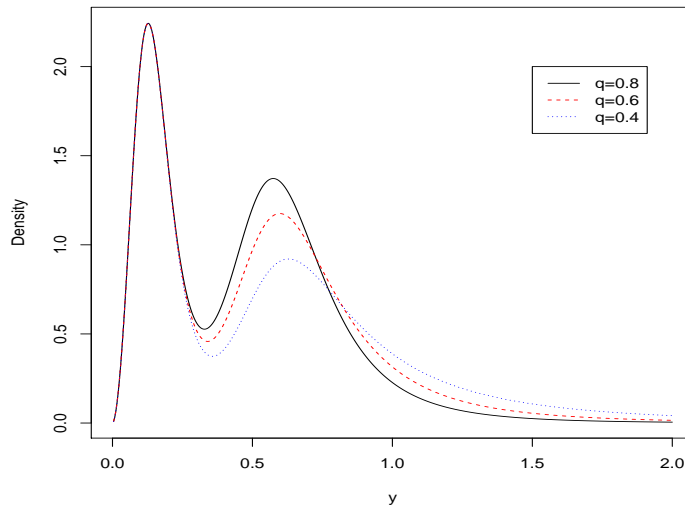


Figure 14: Mixture of two Singh-Maddala distributions

As σ increases and q decreases the upper tail of the distribution decays more slowly. The sample size is $n = 500$ and the MIAE criterion is calculated as the average of $\int_0^\infty |\hat{f}(y) - f(y)| dy$ computed for 1 000 samples.

Table 5 shows the quality of the fit for several density estimation methods. We first consider standard kernel estimator, with fixed bandwidth selected by the Silverman's rule-of-thumb (Silv.), by cross-validation (CV) and by the plug-in (Plug-in) methods. Then, we consider adaptive kernel methods based on each of the previous fixed bandwidths. Finally, we consider density estimation based on mixture of lognormal distributions.³⁵

The results in Table 5 show that, for standard and adaptive kernel methods, the quality of the fit deteriorates as the upper tail becomes more heavy (as σ increases and q decreases, MIAE increases). Moreover, standard kernel method with the Silverman's rule-of-thumb bandwidth fails when the distribution is multimodal and highly skewed (case of two Singh-Maddala distributions), compared to other methods. Finally, our results suggest that, in the cases of heavier-tailed distributions, the adaptive kernel based on the plug-in bandwidth and the mixture of lognormals perform better than standard kernel methods.

³⁵The density function of the logarithmic transformation of the data is estimated by a mixture of normal distributions and the number of components is selected with the BIC.

	Standard kernel			Adaptive kernel			Mixture lognormal
	Silv.	CV	Plug-in	Silv.	CV	Plug-in	
<i>Lognormal</i>							
$\sigma = 0.5$	0.1044	0.1094	0.1033	0.0982	0.1098	0.1028	0.0407
$\sigma = 0.75$	0.1326	0.1326	0.1252	0.1098	0.1283	0.1179	0.0407
$\sigma = 1$	0.1643	0.1716	0.1522	0.1262	0.1609	0.1362	0.0407
<i>Singh-Maddala</i>							
$q = 1.7$	0.0942	0.1009	0.0951	0.0915	0.0994	0.0934	0.0840
$q = 1.2$	0.1039	0.1100	0.1048	0.0947	0.1050	0.0994	0.0920
$q = 0.7$	0.1346	0.1482	0.1326	0.1049	0.1349	0.1175	0.0873
<i>Mixture of two Singh-Maddala</i>							
$q = 0.8$	0.2080	0.1390	0.1328	0.1577	0.1356	0.1224	0.1367
$q = 0.6$	0.2458	0.1528	0.1463	0.1896	0.1457	0.1293	0.1464
$q = 0.4$	0.2885	0.1953	0.1733	0.2234	0.1812	0.1450	0.1366

Table 5: Quality of density estimation (MIAE), $n = 500$.

4 Welfare indices

We can use the term “welfare indices” to cover a number of specific tools of distributional analysis that are of interest to economists and social scientists. These include social-welfare functions, inequality measures and poverty indices. Our approach is to characterise some basic classes of indices, to introduce some standard results that enable us to describe the statistical properties of these indices and then to apply the analysis to particular welfare indices that are of interest to students of income distribution. The applications here will be to inequality and poverty indices.

4.1 Basic cases

It is useful begin with two of the simplest welfare indices, the quantile and the income cumulation. Quantiles and income cumulations are themselves incomes and so belong to the interval $\mathbb{Y} = [\underline{y}, \bar{y})$, introduced in section 1.2. Once again we work with distribution functions $F \in \mathbb{F}$ so that the total population in the distribution is implicitly normalised to 1.

Let $q \in \mathbb{Q}$ denote an arbitrary population proportion; then we may define Q the *quantile functional* from $\mathbb{F} \times \mathbb{Q}$ to \mathbb{Y} as

$$Q(F; q) := \inf\{y | F(y) \geq q\} \tag{23}$$

(Gastwirth 1971). For any distribution F the quantile functional gives the smallest income in \mathbb{Y} such that $100q$ percent of the population have exactly that income or less. In cases where the distribution F is understood, we can use a

shorthand form for the q th quantile

$$y_q := Q(F; q). \quad (24)$$

The functional Q provides the basis for several intuitive approaches to the analysis of income distribution. For example, commonly used quantile ratios — such as the “90/10” ratio, the “90/50” ratio (Alvaredo and Saez 2009, Autor et al. 2008, Burkhauser et al. 2009) – are found by taking pairs of instances of (23) with appropriate q -values: y_{90}/y_{10} , y_{90}/y_{50} and so on.

Likewise C , the *cumulative income functional* from $\mathbb{F} \times \mathbb{Q}$ to \mathbb{Y} is defined as

$$C(F; q) := \int_y^{y_q} y \, dF(y) \quad (25)$$

(Cowell and Victoria-Feser 2002); for any distribution F the cumulative functional gives the total income received by the bottom $100q$ percent of the population. Again, in cases where the distribution F is understood, we can use the shorthand form for the q th cumulation: $c_q := C(F; q)$. A word of caution here: remember that the population is normalised to one; this convention is also embedded in the income cumulations (25). In particular, if we set $q = 1$ in (25), we get

$$c_1 = C(F; 1) = \mu(F), \quad (26)$$

the mean of the distribution F . We can find other intuitive approaches to the analysis of income distribution using C : for example, the income share of the poorest $100q$ percent of the population is obtained from two cumulants defined in (25) as

$$\frac{c_q}{c_1} = \frac{C(F; q)}{C(F; 1)}. \quad (27)$$

However, this is just a beginning. The indices generated by Q and C in (23) and (25) are but two well-known examples of a large class of welfare indices that can be expressed in *additively decomposable* form

$$W_{AD}(F) := \int \phi(y) \, dF(y), \quad (28)$$

up to a transformation involving $\mu(F)$, where $\phi : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ is piecewise differentiable. Decomposability here means decomposable by population subgroups (Cowell and Fiorio 2011). This property can be seen more intuitively in the special case of a discrete distribution. If F consists of m point masses consisting of mass f_i located at income y_i $i = 1, \dots, m$ then (28) becomes

$$\sum_{i=1}^m f_i \phi(y_i). \quad (29)$$

It is clear that the form of (29) implies that the welfare index can be found by evaluating income y_i in each of the m separate groups, weighting by the population of the group and aggregating.

Of course it is not only the rather restrictive class W_{AD} that is interesting for distributional analysis. Many welfare indices can be conveniently expressed in the more general *quasi-additively decomposable* form

$$W_{\text{QAD}}(F) := \int \varphi(y, \mu(F)) \, dF(y) \quad (30)$$

where $\varphi : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ is piecewise differentiable; and most of the other commonly-used welfare indices that cannot be expressed in the form (30) can be expressed in the *rank-dependent* form

$$W_{\text{RD}}(F) := \int \psi(y, \mu(F), F(y)) \, dF(y), \quad (31)$$

where ψ is piecewise differentiable. We will discuss specific examples from the W_{QAD} and W_{RD} classes of functionals in subsections 4.3 and 4.4 on inequality and poverty measures.

4.2 Asymptotic inference

In this section and sections 4.3, 4.4 we focus on estimation and inference problems for cases where the sample size n may be considered to be arbitrarily large.³⁶ The small sample problem is discussed in sections 4.5. Furthermore, for the moment we will concentrate only on *distribution-free* approaches, that do not require any estimation of the density function, parametric or even non-parametric; the parametric approach is considered in section 4.6.

There are several methods that we can use to derive the tools that we need. Here we will make extensive use of an approach that enables us to derive the asymptotic results quickly and simply and that lays the basis for further discussions in section 6 below.³⁷

4.2.1 The influence function

The principal analytical tool employed here is the *influence function* (IF) which can be used here as a device to quantify the effect of a perturbation on some given theoretical distribution. So, assume that $F \in \mathbb{F}$ is the distribution in question and that $H^{(z)} \in \mathbb{F}$ is another distribution that consists just of a single point mass at z

$$H^{(z)}(y) = \iota(y \geq z), \quad (32)$$

where $\iota(\cdot)$ is the indicator function (1). Then the mixture distribution

$$G := [1 - \delta]F + \delta H^{(z)}, \quad 0 \leq \delta \leq 1 \quad (33)$$

can be taken as a representation of the perturbation of the distribution F by the point mass, where δ represents the relative size of the perturbation. Now

³⁶For an overview of literature see Cowell (1999).

³⁷This approach draws heavily on Cowell and Victoria-Feser (2003).

we need a way of quantifying the importance of this perturbation of F : consider a functional $T : \mathbb{F} \rightarrow \mathbb{R}^m$ that represents some statistic in which we are interested. The IF measures the impact of the perturbation on the statistic T for infinitesimal δ , namely

$$IF(z; T, F) := \lim_{\delta \downarrow 0} \left[\frac{T(G) - T(F)}{\delta} \right] \quad (34)$$

which becomes $\frac{\partial}{\partial \delta} T(G)|_{\delta \rightarrow 0}$ if T is differentiable.

The IF is particularly useful in analysing the problem of data contamination (see section 6.1). But the IF has other convenient applications: its relevance to this part of our discussion is that it may be used to derive asymptotic results such as asymptotic covariance matrices. If the distribution G is “near” F (as in equation 33 for small δ) then the first-order von-Mises expansion of T at F evaluated in G is given by

$$T(G) = T(F) + \int IF(y; T, F) d(G - F)(y) + \text{remainder}$$

When the observations are independently and identically distributed according to F then, by the Glivenko-Cantelli theorem, the empirical distribution $F^{(n)} \rightarrow F$. So we may replace G by $F^{(n)}$ for sufficiently large n and obtain

$$T(F^{(n)}) \approx T(F) + \frac{1}{n} \sum_{i=1}^n IF(y_i; T, F) + \text{remainder}$$

from which we obtain (Hampel et al. 1986, p. 85):

Lemma 1 *When the remainder becomes negligible as $n \rightarrow \infty$, by the central limit theorem, $\sqrt{n} (T(F^{(n)}) - T(F))$ is asymptotically normal with asymptotic covariance matrix*

$$\int IF(y; T, F) IF^\top(y; T, F) dF(y) \quad (35)$$

Regularity conditions can be found in Reeds (1976), Boos and Serfling (1980) and Fernholz (1983).

Lemma 1 constitutes the basis of the results which follow. Given a statistic T , one just needs to compute its IF to obtain the asymptotic covariance matrix. For inequality and poverty measures (uni-dimensional statistic), T is a functional $\mathbb{F} \rightarrow \mathbb{R}$. In many cases, we can express the IF as a random variable Z minus its expectation,

$$IF(y, T, F) = Z - E(Z) \quad (36)$$

For uni-dimensional statistic, from Lemma 1, $\sqrt{n} (T(F^{(n)}) - T(F))$ is then asymptotically normal with asymptotic variance equal to

$$\int IF(y, T, F)^2 dF(y) = \int (Z - E(Z))^2 dF(Z), \quad (37)$$

which is nothing but the variance of Z . This result allows us to estimate the asymptotic variance of the statistic from a sample using

$$\widehat{\text{var}}\left(T(F^{(n)})\right) = \frac{1}{n} \widehat{\text{var}}(Z) = \frac{1}{n^2} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (38)$$

where Z_i , for $i = 1, \dots, n$ are sample realizations of Z and $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$. From a sample, the asymptotic variance of the statistic is simple to estimate: it is the empirical variance of Z_1, \dots, Z_n divided by n .

The main issue here is to provide IF s and to express it as a function of Z as in (36), for a wide range of welfare indices and ranking tools and as well for different forms of data. Moreover, for some important cases, we will also develop analytically the formula in (35) so that the approach for computing asymptotic covariances matrices based on the IF can be compared to those from other approaches in the literature.³⁸

4.2.2 Background results

Several useful results in income-distribution analysis can be found from a simple application of the IF ; in particular we have two key properties for the fundamental functionals introduced in section 4.1 (Cowell and Victoria-Feser 2002). Applying (23) to the distribution in (33) we get the q th quantile in the mixture distribution:

$$Q(G, q) = Q\left(F, \frac{q - \iota(y_q \geq z) \delta}{1 - \delta}\right) \quad (39)$$

where $y_q = Q(F, q)$ is the q th quantile for the (unmixed) income distribution. Let f be the density function for the distribution function F ; then, differentiating (39) with respect to δ and setting $\delta = 0$, we obtain the following result.

Lemma 2 *The IF for the quantile functional is:*

$$IF(z; Q(\cdot, q), F) = \frac{q - \iota(Q(F; q) \geq z)}{f(Q(F; q))} = \frac{q - \iota(y_q \geq z)}{f(y_q)}. \quad (40)$$

Likewise if we apply (25) to the distribution in (33) we get the q th income cumulation in the mixture distribution:

$$C(G; q) = [1 - \delta] \int_{\underline{y}}^{Q(G; q)} y dF(y) + \delta z \quad (41)$$

where $Q(G, q)$ is given by (39). Once again, differentiating (41) with respect to δ and setting $\delta = 0$ we obtain another basic result.

³⁸For previous suggestions on the use of the IF for estimating asymptotic variances see e.g. Efron (1982), Deville (1999).

Lemma 3 *The IF for the cumulative income functional is:*

$$\begin{aligned} IF(z; C(\cdot; q), F) &= qQ(F; q) - C(F; q) + \iota(q \geq F(z))[z - Q(F; q)] \\ &= qy_q - c_q + \iota(y_q \geq z)[z - y_q]. \end{aligned} \quad (42)$$

We will find that these results are useful not only for welfare indices considered in this section but also for distributional comparisons treated in section 5 below.

4.2.3 QAD Welfare indices

Let us first deal with the broad W_{QAD} class, the welfare indices that are quasi-additively decomposable; we will turn to the important, but more difficult, rank-dependent class W_{RD} later. Fortunately this class covers a great number of commonly used tools of distributional analysis; fortunately also the properties are straightforward. Given a sample y_1, \dots, y_n , the sample analogues of W_{QAD} defined in (30) are given by

$$\widehat{W}_{\text{QAD}} := W_{\text{QAD}}(F^{(n)}) = \frac{1}{n} \sum_{i=1}^n \varphi(y_i, \hat{\mu}) \quad (43)$$

where $F^{(n)}$ is the EDF defined in equation (8) and $\hat{\mu}$ is the sample mean:

$$\hat{\mu} := \mu(F^{(n)}) = \frac{1}{n} \sum_{i=1}^n y_i. \quad (44)$$

Substituting the mixture distribution (33) into (30), differentiating with respect to δ and evaluating at $\delta = 0$ we find the influence function for the QAD class as:

$$IF(z; W_{\text{QAD}}, F) = \varphi(z, \mu(F)) - W_{\text{QAD}}(F) + [z - \mu(F)] \int \varphi_\mu(z, \mu(F)) dF(z) \quad (45)$$

where φ_μ denotes the partial derivative with respect to the second argument. This influence function can be expressed as in (36), that is, as a random variable Z minus its expectation,

$$IF(y, W_{\text{QAD}}, F) = Z - E(Z), \quad (46)$$

where

$$Z = \varphi(y, \mu(F)) + y \int \varphi_\mu(y, \mu(F)) dF(y). \quad (47)$$

From (36) and (37), the asymptotic variance of $\sqrt{n}(\widehat{W}_{\text{QAD}} - W_{\text{QAD}})$ is then equal to the variance of Z . From (38), the asymptotic variance of \widehat{W}_{QAD} can be estimated from a sample using

$$\widehat{\text{var}}(\widehat{W}_{\text{QAD}}) = \frac{1}{n} \widehat{\text{var}}(Z). \quad (48)$$

4.3 Application: Inequality measures

Almost all commonly-used inequality indices other than the Gini coefficient can be written in the form $\Psi(W_{\text{QAD}}(F), \mu(F))$ where $\Psi : \mathbb{R}^2 \rightarrow \mathbb{R}$. So we can use the results on the broad W_{QAD} class to derive the sampling distribution for a large range of inequality measures. We consider two leading examples in sections 4.3.1 and 4.3.2.

4.3.1 The generalised entropy class

We consider first an important family of inequality measures that belongs to the additively-decomposable class (28). Members of the Generalised Entropy (GE) class (characterised by the parameter ξ) are defined by equations (49)-(51)

$$I_{\text{GE}}^{\xi}(F) = \frac{1}{\xi^2 - \xi} \left[\int_{\underline{y}}^{\bar{y}} \left[\frac{y}{\mu(F)} \right]^{\xi} dF(y) - 1 \right], \quad \xi \in \mathbb{R}, \xi \neq 0, 1 \quad (49)$$

$$I_{\text{GE}}^0(F) = - \int_{\underline{y}}^{\bar{y}} \log \left(\frac{y}{\mu(F)} \right) dF(y) \quad (50)$$

$$I_{\text{GE}}^1(F) = \int_{\underline{y}}^{\bar{y}} \frac{y}{\mu(F)} \log \left(\frac{y}{\mu(F)} \right) dF(y) \quad (51)$$

Clearly the GE class belongs to the decomposable class of indices given in (28), a subset of the broad W_{QAD} class. The parameter ξ of the GE class characterizes the sensitivity to income differences in different parts of the income distribution. The more positive (negative) ξ is, the more sensitive is the inequality measure to income differences at the top (bottom) of the distribution. The Mean Logarithmic Deviation (MLD) index, $I_{\text{GE}}^0(F)$, is the limiting case when $\xi = 0$. The Theil index, $I_{\text{GE}}^1(F)$, is the limiting case of the GE when $\xi = 1$. The sample analogues of these indices are given by

$$\hat{I}_{\text{GE}}^{\xi} := I_{\text{GE}}^{\xi}(F^{(n)}) = \begin{cases} [n(\xi^2 - \xi)]^{-1} \sum_{i=1}^n [(y_i/\hat{\mu})^{\xi} - 1] & \text{for } \xi \neq 0, 1 \\ -n^{-1} \sum_{i=1}^n \log(y_i/\hat{\mu}) & \text{for } \xi = 0 \\ n^{-1} \sum_{i=1}^n (y_i/\hat{\mu}) \log(y_i/\hat{\mu}) & \text{for } \xi = 1 \end{cases} \quad (52)$$

Using (46), (47) and (48), we find that the variance of the generalised entropy measures can be estimated by

$$\widehat{\text{var}}(\hat{I}_{\text{GE}}^{\xi}) = \frac{1}{n^2} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (53)$$

where

$$Z_i = \begin{cases} (\xi^2 - \xi)^{-1} (y_i/\hat{\mu})^{\xi} - \xi (y_i/\hat{\mu}) \left[\hat{I}_{\text{GE}}^{\xi} + (\xi^2 - \xi)^{-1} \right] & \text{for } \xi \neq 0, 1 \\ (y_i/\hat{\mu}) - \log y_i & \text{for } \xi = 0 \\ (y_i/\hat{\mu}) \left[\log(y_i/\hat{\mu}) - \hat{I}_{\text{GE}}^1 - 1 \right] & \text{for } \xi = 1 \end{cases} \quad (54)$$

From a sample y_1, \dots, y_n , the values of Z_1, \dots, Z_n can be calculated for a fixed ξ . The variance estimate of the generalised index is then computed as the empirical variance of Z_1, \dots, Z_n divided by n .

To show the results in (54), let us consider the case $\xi \neq 0, 1$, from which we have

$$\varphi(y, \mu(F)) = \frac{1}{\xi^2 - \xi} \left[\left[\frac{y}{\mu(F)} \right]^\xi - 1 \right] \quad (55)$$

$$\varphi_\mu(y, \mu(F)) = \frac{-\xi}{\xi^2 - \xi} \left[\frac{y^\xi}{\mu(F)^{\xi+1}} \right] = -\frac{\xi}{\mu} \left(\varphi(y, \mu(F)) + \frac{1}{\xi^2 - \xi} \right) \quad (56)$$

Substitute (55) and (56) into (47) gives the result in (54) where y is replaced by its sample realization y_i . The same methodology can be applied for the cases $\xi = 0$ and $\xi = 1$.

Clearly the same approach can be applied to functions of moments of the distribution such as the coefficient of variation. Likewise the statistical properties of the Atkinson class of inequality indices (Atkinson 1970)

$$I_{\text{Atk}}^\xi(F) = 1 - \left[\int_{\underline{y}}^{\bar{y}} \left[\frac{y}{\mu(F)} \right]^\xi dF(y) \right]^{1/\xi}, \quad \xi < 1 \quad (57)$$

can easily be derived from (53).³⁹

The standard approach to obtain the results for the GE class and associated indices is by expressing the indices as a function of moments of the distribution and using the delta method. We can show that both IF and delta approaches give the same results. Indeed, from (28) and (29) a decomposable inequality measure can written as a function of two moments,

$$I = \psi(\nu; \mu) \quad \text{with } \mu = E(y) \text{ and } \nu = E(\phi(y)), \quad (58)$$

where ϕ and ψ are functions $\mathbb{R}^2 \rightarrow \mathbb{R}$ and ψ is monotonic increasing in its first argument; in particular this is true for the I_{GE}^ξ and I_{Atk}^ξ families. The estimation of inequality indices is usually obtained by replacing the unknown moments of the distribution by consistent estimates. The moments are directly estimated by their sample counterparts. Let us consider y_i , for $i = 1, \dots, n$, a sample of IID observations drawn from F . The estimator of the inequality measure can be expressed as a non-linear function of two consistently estimated moments,

$$\hat{I} = \psi(\hat{\nu}; \hat{\mu}) \quad \text{with } \hat{\mu} = \frac{1}{n} \sum_{i=1}^N y_i \text{ and } \hat{\nu} = \frac{1}{n} \sum_{i=1}^N \phi(y_i). \quad (59)$$

³⁹From (49), (50) and (57) we have $I_{\text{Atk}}^\xi(F) = 1 - [(\xi^2 - \xi)I_{\text{GE}}^\xi(F) + 1]^{1/\xi}$, $\xi \neq 0$, and $I_{\text{Atk}}^0(F) = 1 - \exp(-I_{\text{GE}}^0(F))$. Variances of Atkinson indices can thus be written as functions of variances of GE indices using the delta method.

From the Central Limit Theorem (CLT), this estimator is also consistent and asymptotically Normal, with asymptotic variance that can be calculated by the delta method. Specifically, the asymptotic variance is equal to

$$\text{var}(\hat{I}) = \left(\frac{\partial\psi}{\partial\nu}\right)^2 \text{var}(\hat{\nu}) + 2 \left(\frac{\partial\psi}{\partial\nu} \frac{\partial\psi}{\partial\mu}\right) \text{cov}(\hat{\nu}, \hat{\mu}) + \left(\frac{\partial\psi}{\partial\mu}\right)^2 \text{var}(\hat{\mu}). \quad (60)$$

An estimate of the asymptotic variance can be obtained by replacing the moments and their variances and covariance by consistent estimates. For the case $\xi = 0$, the MLD index can be written $I_{\text{GE}}^0 = \log \mu - \nu$ where $\nu = \int \log y dF(y)$. From (60), the asymptotic variance given by the delta method is equal to

$$\frac{1}{\mu^2} \text{var}(\hat{\mu}) - \frac{2}{\mu} \text{cov}(\hat{\mu}, \hat{\nu}) + \text{var}(\hat{\nu}). \quad (61)$$

From the influence function approach, we have $Z = y/\mu - \log y$ and, the asymptotic variance of the MLD index is the variance of Z divided by n ,

$$\frac{1}{n} \text{var}(Z) = \frac{1}{n} \left[\frac{1}{\mu^2} \text{var}(y) - \frac{2}{\mu} \text{cov}(y, \log y) + \text{var}(\log y) \right]. \quad (62)$$

The two equations (61) and (62) are identical, which demonstrates that the delta and the IF methods give the same results. It can be also demonstrated for $\xi \neq 0$.⁴⁰

4.3.2 The mean deviation and its relatives

Now consider the mean deviation, an inequality index which does not belong to the class of decomposable indices (28), but does belong to the quasi-additive class (30).

$$I_{\text{MD}}(F) := \int |y - \mu(F)| dF(y).$$

Noting that $I_{\text{MD}}(F)$ can be rewritten as

$$I_{\text{MD}}(F) = 2 \int \iota_y [y - \mu(F)] dF(y) \quad (63)$$

where $\iota_y := \iota(y \geq \mu(F))$, the influence function is

$$IF(z; I_{\text{MD}}, F) = 2[\iota_z + \bar{q} - 1][z - \mu(F)] - I_{\text{MD}}(F) \quad (64)$$

where $\bar{q} := F(\mu)$. The asymptotic variance of the MD index can be obtained, rewriting the influence function as a random variable minus its expectation $IF(y; I_{\text{MD}}, F) = Z - E(Z)$. From (64), we have

$$Z = 2(\bar{q} - 1)y + 2\iota_y[y - \mu(F)] \quad (65)$$

⁴⁰For extensions to the case of weighted data and complex survey design see Zheng and Cushing (2001), Cowell and Jenkins (2003), Biewen and Jenkins (2006), Verma and Betti (2011). For an alternative approach to estimation of Atkinson indices using a Box-Cox transformation, see Guerrero (1987).

From Lemma 1, (36), (37), the asymptotic variance of $\sqrt{n}(I_{\text{MD}}(F^{(n)}) - I_{\text{MD}}(F))$ is then equal to the variance of Z .

From a sample (y_1, \dots, y_n) , the Mean Deviation index can be estimated as,

$$\hat{I}_{\text{MD}} := I_{\text{MD}}(F^{(n)}) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{\mu}|. \quad (66)$$

The asymptotic variance can be estimated as the empirical variance of (Z_1, \dots, Z_n) divided by n ,

$$\widehat{\text{var}}(\hat{I}_{\text{MD}}) = \frac{1}{n^2} \sum_{i=1}^n (Z_i - \bar{Z})^2, \quad (67)$$

where

$$Z_i = 2(\hat{q} - 1)y_i + 2(y_i - \hat{\mu}) \iota(y_i \geq \hat{\mu}), \quad (68)$$

and $\hat{q} := F^{(n)}(\hat{\mu}) = n^{-1} \sum_{i=1}^n \iota(y_i \leq \hat{\mu})$.

The same methodology with some extra terms can be used to derive the asymptotic variance of the more commonly used relative mean deviation or Pietra ratio

$$\int \left| \frac{y}{\mu(F)} - 1 \right| dF(y).$$

In the literature, the asymptotic variance is usually obtained with the *IF* method without using it expressed as a function of a random variable minus its expectation. It gives similar numerical results, but formulas and implementation are more complicated. For instance, using Lemma 1 and equations (64) and (63), the asymptotic variance of the Mean Deviation can be derived as follows:

$$\begin{aligned} \int IF(z; I_{\text{MD}}, F)^2 dF(z) &= 4 \int [\iota_z + \bar{q} - 1]^2 [z - \mu(F)]^2 dF(z) + I_{\text{MD}}(F)^2 \\ &\quad - 2I_{\text{MD}}(F) \int 2\iota_z [z - \mu(F)] dF(z) \\ &= 4[\bar{q} - 1]^2 \int_{\underline{y}}^{\mu(F)} [z - \mu(F)]^2 + 4\bar{q}^2 \int_{\mu(F)}^{\bar{y}} [z - \mu(F)]^2 - I_{\text{MD}}(F)^2 \end{aligned}$$

This formula for the asymptotic variance of the mean deviation index is as the same as derived in Gastwirth (1974).

4.3.3 The Gini coefficient

The general form (31) is cumbersome but we can fairly easily derive results for the most important member of this class, namely the Gini coefficient.

The Gini index can be expressed in a number of different forms. Let us consider the following expressions,

$$I_{\text{Gini}}(F) = \frac{1}{2\mu} \iint |y - y'| dF(y) dF(y'), \quad (69)$$

$$= 1 - 2 \int_0^1 L(F; q) dq, \quad (70)$$

where $L(F; q) = C(F; q)/\mu(F)$ is the q th ordinate of the Lorenz curve – see equation (122) below. Equation (69) presents the Gini as the normalised average absolute difference between all the possible pairs of incomes in the population, while equation (70) shows that the Gini index is twice the area between the Lorenz curve and the 45° line.

Applying the influence function method to the form (70) we find the IF of I_{Gini} to be given by (Monti 1991)

$$IF(z; I_{\text{Gini}}, F) = 1 - I_{\text{Gini}}(F) - \frac{2C(F; F(z))}{\mu(F)} + z \frac{1 - I_{\text{Gini}}(F) - 2[1 - F(z)]}{\mu(F)} \quad (71)$$

The asymptotic variance of the Gini coefficient has been derived from the influence function in Cowell and Victoria-Feser (2003), Bhattacharya (2007), Barrett and Donald (2009) and Davidson (2009a, 2010). A simple formula can be obtained, noting that the IF of the Gini index can be expressed as a random variable minus its expectation, $IF(z; I_{\text{Gini}}, F) = (Z - E(Z))/\mu(F)$, where⁴¹

$$Z = [1 - I_{\text{Gini}}(F)]z - 2[C(F; F(z)) + z(1 - F(z))] \quad (72)$$

Using Lemma 1, (36) and (37), one immediately gets the asymptotic variance of $\sqrt{n}(I_{\text{Gini}}(F^{(n)}) - I_{\text{Gini}}(F))$, equal to the variance of Z divided by the square of the mean, $\text{var}(Z)/\mu(F)^2$.

The computation of the Gini index and its variance can be easily obtained in practice. If we define the “positional weight”

$$\kappa(y) := \frac{F(y^-) + F(y^+) - 1}{\mu(F)},$$

where $F(y^-) := \lim_{x \uparrow y} F(x)$ and $F(y^+) := \lim_{x \downarrow y} F(x)$, then the definition (69) can alternatively be expressed in the following convenient forms:⁴²

$$I_{\text{Gini}}(F) = \int \kappa(y)y \, dF(y) \quad (73)$$

$$= \frac{2}{\mu} \text{cov}(y, F(y)) \quad (74)$$

In other words the Gini is also equal to the weighted sum of incomes using the κ weights (73) and is equal to $2/\mu$ times the covariance between y and $F(y)$ (74). For the distribution-free approach, we replace $\mu(F)$ by the sample mean $\hat{\mu}$ and the covariance by an unbiased estimate in (74). It leads us to compute the Gini index as:

$$\hat{I}_{\text{Gini}} = I_{\text{Gini}}(F^{(n)}) = \sum_{i=1}^n \kappa_i y_{(i)} \quad (75)$$

⁴¹Note that $E[C(F; F(z))] = E[z[1 - F(z)]] = [1 - I_{\text{Gini}}(F)]\mu(F)/2$.

⁴²Using the definition of the Lorenz curve in (70), interchanging the order of the integration and simplifying give the result (Davidson 2009a). For an extensive list of alternative equivalent ways of writing the Gini coefficient – including our expressions (69), (70), (73), (74) – see Yitzhaki and Schechtman (2013).

where the $y_{(i)}$, $i = 1, \dots, n$, are the order statistics ($y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$) and⁴³

$$\kappa_i := \frac{2i - n - 1}{\hat{\mu}n(n - 1)}.$$

Davidson (2009a) shows that (75) is a bias-corrected estimator of the Gini index,⁴⁴ and proposed estimating the variance of the Gini index as:

$$\widehat{\text{var}}(\hat{I}_{\text{Gini}}) = \frac{1}{(n\hat{\mu})^2} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (76)$$

where

$$Z_i = -(\hat{I}_{\text{Gini}} + 1)y_{(i)} + \frac{2i - 1}{n}y_{(i)} - \frac{2}{n} \sum_{j=1}^i y_{(j)} \quad (77)$$

and $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$. Here the Z_i terms are estimates of the realizations of Z defined in (72), where the value of $F(y_{(i)})$ is estimated by $F^{(n)}(y_{(i)}) = (2i - 1)/(2n)$.⁴⁵ Davidson (2010) extends this approach to derive a variance estimator for the family of S-Gini indices.

4.4 Application: Poverty measures

For a poverty index we need a poverty line which may be an exogenously given constant ζ or may depend on the income distribution $\zeta(F)$. An important class of poverty indices can then be described in the following way:

$$P(F) := \int p(y, \zeta(F)) dF(y) \quad (78)$$

where p is a *poverty evaluation function* that is non-increasing in y and takes the value zero for $y \geq \zeta(F)$. Once again we need the influence function which is given by

$$IF(z; P, F) = p(z, \zeta(F)) - P(F) + \int p_\zeta(y, \zeta) dF(y) IF(z; \zeta, F) \quad (79)$$

where p_ζ is the differential of p with respect to its second argument (Cowell and Victoria-Feser 1996a). It is clear from (79) that the form for the asymptotic

⁴³Clearly (75) is the empirical counterpart of the weighted sum (73). Also, using $\text{Cov}(y, F(y)) = E(yF(y)) - E(y)E(F(y))$ in (74) and replacing $E(yF(y))$ by $(n - 1)^{-1} \sum_{i=1}^n y_{(i)}(i/n)$ and $E(y)E(F(y))$ by $\hat{\mu}(n - 1)^{-1} \sum_{i=1}^n (i/n)$ gives (75).

⁴⁴Equation (75) is equal to $n/(n - 1)$ times equation (5) in Davidson (2009a).

⁴⁵Other estimates of the variance of the Gini index have been proposed in the literature, but they are either complicated or quite unreliable. Nygård and Sandström (1985), Sandström et al. (1988), Cowell (1989), Schechtman (1991) and Bhattacharya (2007) provide formulas that are not easy to implement. A simple method based on OLS regression has been proposed by Ogwang (2000) and Giles (2004), but the standard errors obtained are unreliable, as shown by Modarres and Gastwirth (2006). For a recent review of this literature, see Langel and Tillé (2013) or Yitzhaki and Schechtman (2013); for applications to complex survey designs see Binder and Kovacevic (1995), Kovacevic and Binder (1997).

variance of the poverty index will depend on the precise way in which the poverty line depends on the income distribution. The following specifications cover almost all the versions encountered in practice

$$\zeta(F) = \zeta_0 + \gamma\mu(F), \quad (80)$$

or

$$\zeta(F) = \zeta_0 + \gamma y_q, \quad q \in \mathbb{Q}, \quad (81)$$

where y_q is defined in (24). The interpretation is that the poverty line could be tied to the mean, as in (80) in which case we have

$$IF(z; \zeta, F) = \gamma IF(z; \mu, F) = \gamma [z - \mu(F)] \quad (82)$$

or to a quantile (81), such as the median in which case we have

$$IF(z; \zeta, F) = \gamma \frac{q - \iota(y_q \geq z)}{f(y_q)}. \quad (83)$$

The asymptotic variance can be immediately calculated from (79) and (82) or (83). Let us take the simple case where $\gamma = 0$ ⁴⁶ so that one has the exogenous poverty line ζ_0 . Equation (79) yields the influence function $p(z, \zeta_0) - P(F)$ and so, using Lemma 1, we find the asymptotic variance of $P(F)$ in (78) to be:

$$\int p(z, \zeta_0)^2 dF(z) - P(F)^2.$$

The asymptotic variance of the poverty index is then equal to the variance of the poverty evaluation function, $\text{var}(p(y, \zeta_0))$. We can see that the influence function above is expressed as a function of a random variable minus its expectation,

$$IF(y; P, F) = Z - E(Z) \quad \text{where} \quad Z = p(y, \zeta_0) \quad (84)$$

From (36) and (37) the asymptotic variance is the variance of Z .

A second important class of poverty indices consists of those in the *rank-dependent* form – compare (31) above – and can be described in the following way:

$$P_{\text{RD}}(F) := \int p(y, \zeta(F), F(y)) dF(y) \quad (85)$$

Comparing (85) with (78) we see that the poverty evaluation function p has an extra argument reflecting the individual's rank in the population. The influence function for this class of poverty measures is more complicated (Cowell and Victoria-Feser 1996a) and we deal with this separately in sections 4.4.2 and 4.4.3.

⁴⁶Note that if $\gamma > 0$ then to estimate the asymptotic variance of P using (82) one needs information on the whole distribution; with (83) one needs a density estimate at y_q .

4.4.1 Foster-Greer-Thorbecke (FGT)

For a fixed poverty line ζ_0 the widely used class of poverty indices introduced by Foster et al. (1984) belongs to the class (78) and has the form

$$P_{\text{FGT}}^\xi(F) = \int_0^{\zeta_0} \left(\frac{\zeta_0 - y}{\zeta_0} \right)^\xi dF(y) \quad \xi \geq 0, \quad (86)$$

When $\xi = 0$, the FGT poverty measure is equal to the headcount ratio, which gives the proportion of individuals living in poverty, $F(\zeta_0)$. This index is insensitive to the distribution of incomes among the poor and, therefore, to the depth of poverty. When $\xi = 1$, the FGT poverty measure is the poverty gap index, which considers how far, on the average, the poor are from that poverty line. This index captures the depth of poverty, but it is insensitive to some types of transfers among the poor and, therefore, to some distributional aspect of poverty. Let $y_i, i = 1, \dots, n$, be an IID sample from the distribution F . The FGT poverty indices (86) can be estimated consistently as follows:

$$\hat{P}_{\text{FGT}}^\xi := P_{\text{FGT}}^\xi(F^{(n)}) = \frac{1}{n} \sum_{i=1}^{n_p} \left(\frac{\zeta_0 - y_{(i)}}{\zeta_0} \right)^\xi, \quad (87)$$

where n_p is the number of individuals with incomes not greater than the poverty line, that is, the number of poor, and $y_{(i)}, i = 1, \dots, n$ are the order statistics ($y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$). This estimate is asymptotically Normal, with a variance that can be estimated as:⁴⁷

$$\widehat{\text{var}}(\hat{P}_{\text{FGT}}^\xi) = \frac{1}{n} \left(\hat{P}_{\text{FGT}}^{2\xi} - \left(\hat{P}_{\text{FGT}}^\xi \right)^2 \right). \quad (88)$$

We can also estimate the asymptotic variance of the FGT index using the influence function expressed as a function of a random variable minus its expectation. From (78), (84) and (86), we have $IF(y, P_{\text{FGT}}^\xi, F) = Z - E(Z)$ where

$$Z_i = \left| 1 - \frac{y}{\zeta_0} \right|^\xi \iota(y \leq \zeta_0) \quad (89)$$

From Lemma 1, (36), (37), the asymptotic variance of $\sqrt{n}(P_{\text{FGT}}^\xi(F^{(n)}) - P_{\text{FGT}}^\xi(F))$ is then equal to the variance of Z . From a sample (y_1, \dots, y_n) , let us define

$$Z_i = \begin{cases} [(\zeta_0 - y_{(i)})/\zeta_0]^\xi & \text{for } i \leq n_p \\ 0 & \text{for } i > n_p \end{cases} \quad (90)$$

where the $y_{(i)}, i = 1, \dots, n$, are the order statistics. We can see that

$$\hat{P}_{\text{FGT}}^\xi = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{Z} \quad \text{and} \quad \widehat{\text{var}}(\hat{P}_{\text{FGT}}^\xi) = \frac{1}{n^2} \sum_{i=1}^n (Z_i - \bar{Z})^2. \quad (91)$$

⁴⁷See Kakwani (1993).

The FGT index can be estimated by the mean of Z_i , for $i = 1, \dots, n$, with a variance estimated by the empirical variance of Z_i , divided by n . The two approaches, in equations (87), (88) and in (91), give similar numerical results.⁴⁸

4.4.2 Sen poverty index

The Sen poverty index (Sen 1976) belongs to the class (85) and can be expressed as the average of the headcount ratio and the poverty gap index, weighted by the Gini coefficient among the poor,

$$P_{\text{Sen}}(F) = P_{\text{FGT}}^0 I_{\text{Gini}}^p + P_{\text{FGT}}^1 (1 - I_{\text{Gini}}^p), \quad (92)$$

where I_{Gini}^p is the Gini index computed with incomes below the poverty lines. When the distribution of incomes among the poor is equal, $I_{\text{Gini}}^p = 0$, the Sen index is equal to the headcount ration ($P_{\text{Sen}} = P_{\text{FGT}}^0$). When the distribution of incomes among the poor is extremely unequal, $I_{\text{Gini}}^p = 1$, the Sen index is equal to the poverty gap index ($P_{\text{Sen}} = P_{\text{FGT}}^1$).

The Sen poverty measure can be written

$$P_{\text{Sen}}(F) = \frac{2}{\zeta_0 F(\zeta_0)} \int_0^{\zeta_0} (\zeta_0 - y)(F(\zeta_0) - F(y)) dF(y) \quad (93)$$

From Davidson (2009a), we can derive the influence function as a function of a random variable minus its expectation, $IF(z, P_{\text{Sen}}, F) = \frac{2}{\zeta_0 F(\zeta_0)}(Z - E(Z))$, where⁴⁹

$$Z = \left[\zeta_0 F(\zeta_0) - \frac{\zeta_0 P_S}{2} - zF(\zeta_0) + zF(z) - C(F; F(z)) \right] \iota(z \leq \zeta_0) \quad (94)$$

A consistent estimate of the Sen's poverty index can be obtained by replacing F by $F^{(n)}$ in (93),⁵⁰

$$\hat{P}_{\text{Sen}} := P_{\text{Sen}}(F^{(n)}) = \frac{2}{nn_p \zeta_0} \sum_{i=1}^{n_p} (\zeta_0 - y_{(i)}) \left(n_p - i + \frac{1}{2} \right), \quad (95)$$

where the value of $F(y_{(i)})$ is estimated by $F^{(n)}(y_{(i)}) = (2i - 1)/(2n)$. This estimate is asymptotically Normal, with a variance that can be computed as follows:

$$\widehat{\text{var}}(\hat{P}_{\text{Sen}}) = \frac{4}{(\zeta_0 n_p)^2} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (96)$$

⁴⁸The problem of estimation in the presence of complex survey design is addressed in Howes and Lanjouw (1998), Zheng (2001), Berger and Skinner (2003), Verma and Betti (2011).

⁴⁹In equation (50) in Davidson (2009a), replacing y_i in the expression of the summand by z yields the influence function. The relationship to the IF is not related in the last paper, it is related in Davidson (2010) where the same method is used with S-Gini indices.

⁵⁰This expression does not coincide exactly with Sen's own definition for a discrete population. See Appendix A in Davidson (2009a) for a discussion of this point.

where

$$Z_i = \frac{\zeta_0}{2} \left(\frac{2n_p}{n} - \hat{P}_{\text{Sen}} \right) - \frac{2n_p - 2i + 1}{2n} y_{(i)} - \frac{1}{n} \sum_{j=1}^i y_{(j)} \quad (97)$$

for $i = 1, \dots, n_p$, and $Z_i = 0$ for $i = n_p + 1, \dots, n$, with $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$. Here, Z_i are estimates of the realizations of Z defined in (94).⁵¹

4.4.3 Sen-Shorrocks-Thon (SST) poverty index

The Sen-Shorrocks-Thon index is a convenient modified version of the Sen's poverty index, defined as follows,

$$P_{\text{SST}}(F) = P_{\text{FGT}}^0 P_{\text{FGT}_g}^1 (1 + I_{\text{Gini}}^{pg}) \quad (98)$$

where $P_{\text{FGT}_g}^1$ is the poverty gap index computed with incomes below the poverty line, and I_{Gini}^{pg} is the Gini coefficient computed with individual's poverty gap ratios rather than individual's incomes for the whole population ($(\zeta_0 - y_i)/\zeta_0$ rather than y_i , for $i = 1, \dots, n$).⁵² The Gini coefficient of poverty gap ratios can be viewed as a measure of poverty inequality in a society. The SST index satisfies the transfer and continuity axioms, while the Sen index does not.⁵³ This index can be decomposed into

$$\Delta \log P_{\text{SST}} = \Delta \log P_{\text{FGT}}^0 + \Delta P_{\text{FGT}_g}^1 + \Delta \log(1 + I_{\text{Gini}}^{pg}) \quad (99)$$

A percentage change in SST can then be viewed as the sum of percentage changes in the proportion of poor, the average poverty gap among the poor and one plus the Gini index of poverty gaps for the population. The poverty is decomposed in three aspects: are there more poor? are the poor poorer? is there higher poverty inequality in the society?

The SST poverty index can be written

$$P_{\text{SST}}(F) = \frac{2}{\zeta_0} \int_0^{\zeta_0} (\zeta_0 - y)(1 - F(y)) dF(y) \quad (100)$$

As in section 4.4.2 we derive the influence function as a function of a random variable minus its expectation, $IF(z, P_{\text{SST}}, F) = \frac{2}{\zeta_0} (Z - E(Z))$, where⁵⁴

$$Z = [\zeta_0(1 - F(\zeta_0)) - z(1 - F(z)) + C(F; F(\zeta_0)) - C(F; F(z))] \iota(z \leq \zeta_0) \quad (101)$$

⁵¹Another variance estimator has been proposed by Bishop et al. (1997).

⁵²The original index was proposed by Shorrocks (1995). Xu and Osberg (2002) show that it can be written as equation (98). They also show that the Sen index is equal to $S = P_{\text{FGT}}^0 P_{\text{FGT}_g}^1 (1 + I_{\text{Gini}}^{pg})$ and therefore that the SST index differs from the Sen index because it uses the Gini index of poverty gap ratios for the whole population, whereas the Sen index uses the Gini index of poverty gap ratios for the poor

⁵³The (strong upward) transfer axiom states that an increase in a poverty measure should occur if the poorer of the two individuals involved in an upward transfer of income is poor, and even if the beneficiary crosses the poverty line.

⁵⁴In the first equation of column two p.39 in Davidson (2009a), replacing y_i in the expression of the summand by z yields the influence function. The relationship to the IF is not related in the last paper, it is related in Davidson (2010) where the same method is used with S-Gini indices

The SST poverty index can be consistently estimated as:⁵⁵

$$\hat{P}_{\text{SST}} := P_{\text{SST}}(F^{(n)}) = \frac{2}{\zeta_0 n(n-1)} \sum_{i=1}^{n_p} (\zeta_0 - y_{(i)})(n-i) \quad (102)$$

It is asymptotically Normal, with an estimator of the variance given by

$$\widehat{\text{var}}(\hat{P}_{\text{SST}}) = \frac{4}{\zeta_0^2 (n-1)^2} \sum_{i=1}^n (Z_i - \bar{Z})^2 \quad (103)$$

where

$$Z_i = \zeta_0 \left(1 - \frac{n_p}{n}\right) - \frac{2n-2i+1}{2n} y_{(i)} + \frac{1}{n} \sum_{j=1}^{n_p} y_{(j)} - \frac{1}{n} \sum_{j=1}^i y_{(j)} \quad (104)$$

for $i = 1, \dots, n_p$, and $Z_i = 0$ for $i = n_p + 1, \dots, n$, with $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$. Here, Z_i are estimates of the realizations of Z defined in (101).

4.5 Finite sample properties

4.5.1 Asymptotic and bootstrap methods

Asymptotic Normality allows us to perform asymptotic inference. In practice, we are concerned with finite samples and asymptotic inference can be unreliable. When asymptotic inference does not perform well in finite sample, bootstrap methods can be used to perform accurate inference. The bootstrap appears to be an ideal method for inference with inequality and poverty indices, since the observations of the sample are often IID.

Let us consider a welfare index W and its sample counterpart \hat{W} . An asymptotic confidence interval at 95% would be computed as

$$CI_{\text{asym}} = [\hat{W} - c_{0.975} \widehat{\text{var}}(\hat{W})^{1/2}; \hat{W} - c_{0.025} \widehat{\text{var}}(\hat{W})^{1/2}] \quad (105)$$

where $c_{0.025}$ and $c_{0.975}$ are the 2.5-th and 97.5-th percentiles of the asymptotic distribution of the t -statistic, $t = (\hat{W} - W_0) / \widehat{\text{var}}(\hat{W})$, where W_0 is the true value of the welfare index. In general, the asymptotic distribution of the t -statistic is the standard Normal distribution, from which we have $c_{0.975} = -c_{0.025} \approx 1.96$.

When the bootstrap is used in combination with the asymptotic variance estimate, asymptotic refinements can be obtained over asymptotic methods.⁵⁶ To compute a bootstrap confidence interval, we generate B samples of size n by re-sampling with replacement from the observed sample. For bootstrap sample

⁵⁵The SST index defined by Shorrocks (1995) is obtained by replacing F by \hat{F} in (100). Here, (102) is a simplified version of the bias-corrected estimator in the last equation of p.37 in Davidson (2009a). In the case of complex survey data see Osberg and Xu (2000).

⁵⁶See Beran (1988). It means that the bootstrap method presented in this section provides an asymptotic refinement over the percentile bootstrap proposed in Mills and Zandvakili (1997).

b , we compute the index \hat{W}_b^* , its variance estimate $\widehat{\text{var}}(\hat{W}_b^*)$, and the t -statistic $t_b^* = (\hat{W}_b^* - \hat{W})/\widehat{\text{var}}(\hat{W}_b^*)^{1/2}$. A bootstrap confidence interval at 95% would be computed as

$$CI_{\text{boot}} = [\hat{W} - c_{0.975}^* \widehat{\text{var}}(\hat{W})^{1/2}; \hat{W} - c_{0.025}^* \widehat{\text{var}}(\hat{W})^{1/2}] \quad (106)$$

where $c_{0.025}^*$ and $c_{0.975}^*$ are the 2.5-th and 97.5-th percentiles of the EDF of the bootstrap t -statistics, that is, the $[0.025B]$ and $[0.975B]$ order statistics of the t_b^* , where $[x]$ denotes the smallest integer not smaller than x . In this approach, the unknown distribution of the population is replaced by the EDF of the original sample, from which we generate bootstrap samples and compute t -statistics testing the (true) hypothesis that the index is equal to \hat{W} . The simulated distribution of the bootstrap t -statistics is used, as an approximation of the unknown distribution of t , to calculate critical values.

The bootstrap can also be used to test hypotheses and to compute p -values. In order to test the hypothesis that the population value of the index is W_0 , for a one-tailed test, the bootstrap p -value would be the proportion of the t_b^* that are more extreme than the t -statistic computed from the observed sample t . Here, the bootstrap test is also based on the EDF of the bootstrap t -statistics t_b^* . The null hypothesis is rejected at significance level 0.05 if the bootstrap p -value is less than 0.05. In order to test the hypothesis that two indices are the same from two populations, a suitable t -statistic is $\tau = (\hat{W}_1 - \hat{W}_2)/(\widehat{\text{var}}(\hat{W}_1) + \widehat{\text{var}}(\hat{W}_2))^{1/2}$ when the samples are independent. For bootstrap sample b , the bootstrap statistic is computed then as $\tau_b^* = (\hat{W}_{1b}^* - \hat{W}_{2b}^* - \hat{W}_1 + \hat{W}_2)/(\widehat{\text{var}}(\hat{W}_{1b}^*) + \widehat{\text{var}}(\hat{W}_{2b}^*))^{1/2}$. When the samples are dependent, the statistic should take account of the covariance and the bootstrap samples should be generated by re-sampling pairs of observations with replacement. Again, the bootstrap p -value would be the proportion of the τ_b^* that are more extreme than τ .

4.5.2 Simulation evidence

We now turn to the performance in finite sample of inference based on inequality and poverty measures. The coverage rate of a confidence interval is the probability that the random interval does include, or cover, the true value of the parameter. A method of constructing confidence intervals with good finite sample properties should provide a coverage rate close to the nominal confidence level. For a confidence interval at 95%, the nominal coverage rate is equal to 95%. In this section, we use Monte-Carlo simulation to approximate the coverage rate of asymptotic and bootstrap confidence intervals in several experimental designs.

In our experiments, data are generated from Lognormal distributions, $\Lambda(y; 0, \sigma)$, and from Singh-Maddala distributions, $\text{SM}(y; 2.8, 0.193, q)$. As σ increases and q decreases the upper tail of the distribution decays more slowly. The sample size is $n = 500$, the number of bootstrap samples is $B = 499$ and the number of

	Theil		MLD		Gini		SST	
	asym	boot	asym	boot	asym	boot	asym	boot
Lognormal								
$\sigma = 0.5$	0.927	0.936	0.936	0.942	0.942	0.943	0.926	0.952
$\sigma = 1.0$	0.871	0.913	0.922	0.936	0.922	0.936	0.945	0.940
$\sigma = 1.5$	0.746	0.854	0.888	0.921	0.876	0.920	0.964	0.937
Singh-Maddala								
$q = 1.7$	0.915	0.931	0.938	0.945	0.945	0.944	0.945	0.950
$q = 1.2$	0.856	0.905	0.913	0.930	0.925	0.934	0.945	0.951
$q = 0.7$	0.647	0.802	0.820	0.890	0.847	0.906	0.939	0.946

Table 6: Coverage of asymptotic and bootstrap confidence intervals at the 95% level for the Theil, MLD, Gini and SST indices, $n = 500$.

experiments is $N = 10\,000$.⁵⁷ When poverty indices are used, the poverty line is computed as half the median.

Table 6 presents coverage of asymptotic and bootstrap confidence intervals at the 95% level for the Theil, Mean Logarithmic Deviation (MLD), Gini and Sen-Shorrocks-Thon (SST) indices. The results show that asymptotic and bootstrap confidence intervals are reliable when we consider the SST poverty index. Indeed, the coverage rates of the SST index are always close to the nominal coverage rate of 95%. In contrast, when we consider inequality measures, bootstrap confidence intervals outperform asymptotic confidence intervals, but they become less reliable as σ increases and q decreases. In other words, asymptotic and bootstrap inference deteriorate as the upper tail of the underlying distribution is more heavy. For instance, asymptotic confidence intervals cover the true value of the Theil index 64.7% of times when the underlying distribution is the Singh-Maddala with $q = 0.7$. Bootstrap confidence intervals provide better results, with a coverage rate of 80.2%, but it is still significantly different from the expected 95%. Note that the Theil index is known to be more sensitive to the upper tail of the distribution than the MLD and Gini, and confidence intervals with the Theil index are slightly less reliable than with the MLD and Gini indices.

These results illustrate that asymptotic and bootstrap inference on inequality measures is sensitive to the exact nature of the upper tail of the income distribution. Bootstrap inference on inequality measures are expected to perform reasonably well in moderate and large samples, unless the tails are quite heavy.⁵⁸ Moreover, asymptotic and bootstrap inference on poverty measures perform well in finite sample.

⁵⁷For well-known reasons - see Davison and Hinkley (1997) or Davidson and MacKinnon (2000) - the number of bootstrap re-samples B should be chosen so that $(B + 1)/100$ is an integer

⁵⁸Additional results with other distributions, other indices and hypotheses testing can be found in Davidson and Flachaire (2007), Cowell and Flachaire (2007), Davidson (2009a, 2010, 2012)

4.5.3 Inference with heavy-tailed distributions

When the distribution is one with quite a heavy upper tail, asymptotic and bootstrap inference are known to perform poorly in finite sample. Several approaches have been proposed in the literature to obtain more reliable inference.

Schluter and van Garderen (2009) and Schluter (2012) propose normalizing transformation of the index, before to use the bootstrap, in order to use a statistic with a distribution closer to the Normal. Let g denote a transformation of the index W ; a standard bootstrap confidence interval can be obtained on the transformed index $g(W)$ and, therefore, on the untransformed index by inverting the relation between the welfare index and the parameters. Let $c_{0.025}^*$ and $c_{0.975}^*$ be the 2.5-th and 97.5-th percentiles of the EDF of the bootstrap t -statistics

$$t_b^* = \frac{g(\hat{W}_b^*) - g(\hat{W})}{g'(\hat{W}_b^*) \widehat{\text{var}}(\hat{W}_b^*)^{1/2}},$$

where g' is the first derivative of g , a bootstrap confidence interval at 95% for W would then be defined as

$$\left[g^{-1} \left(g(\hat{W}) - c_{0.025}^* g'(\hat{W}) \widehat{\text{var}}(\hat{W})^{1/2} \right); g^{-1} \left(g(\hat{W}) - c_{0.975}^* g'(\hat{W}) \widehat{\text{var}}(\hat{W})^{1/2} \right) \right],$$

if g^{-1} is non-decreasing, otherwise $c_{0.025}^*$ and $c_{0.975}^*$ should be interchanged. For instance, Schluter (2012) exploits a systematic relationship between the inequality estimate and its estimated variance to propose variance stabilizing transforms of the index. He suggests to compute confidence intervals based on the following transform of the index,

$$g(W) = -\frac{2}{\gamma_2} \exp \left(-\frac{\gamma_1}{2} - \frac{\gamma_2}{2} W \right), \quad (107)$$

where γ_1 and γ_2 are the intercept and the slope of a (systematic) linear relation between the index \hat{W} and the logarithmic transformation of its variance estimates $\widehat{\text{var}}(\hat{W})$, and $\gamma_2 > 0$. The parameters γ_1 and γ_2 are estimated by ordinary least squares estimation from the regression

$$\log \widehat{\text{var}}(\hat{W}) = \gamma_1 + \gamma_2 \hat{W} + \varepsilon,$$

where realizations of $\widehat{\text{var}}(\hat{W})$ and \hat{W} are obtained by a preliminary bootstrap. For the specific transform (107), the inverse function is equal to

$$g^{-1}(x) = -\frac{2}{\gamma_2} \log \left(-\frac{\gamma_2}{2} x \right) - \frac{\gamma_1}{\gamma_2} \quad (108)$$

A bootstrap confidence interval at 95% can be computed using (107) and (108) in the confidence interval defined above.

Davidson and Flachaire (2007) and Cowell and Flachaire (2007) consider a semi-parametric bootstrap, where bootstrap samples are generated from a distribution which combines a parametric estimate of the upper tail with a non-parametric estimate of the rest of the distribution. The upper tail is modelled

by a Pareto distribution with parameter α estimated by the Hill estimator on the k greatest-order statistics of a sample of size n , for some integer $k \leq n$,

$$\hat{\alpha} := \left(\frac{1}{k} \sum_{i=0}^{k-1} \log y_{(n-i)} - \log y_{n-k+1} \right)^{-1} \quad (109)$$

where $y_{(j)}$ is the j th order statistic of the sample. Each observation of a bootstrap sample is, with probability p_{tail} , a drawing from the CDF of the Pareto distribution $F(y) = 1 - (y/y_0)^{-\hat{\alpha}}$, $y > y_0$, where y_0 is the order statistic of rank $n(1 - p_{\text{tail}})$, and, with probability $1 - p_{\text{tail}}$, a drawing from the empirical distribution of the sample of smallest $n(1 - p_{\text{tail}})$ -order statistics. In order for the bootstrap to test a true null hypothesis, we need to compute the value of the welfare index for the bootstrap distribution defined above. The CDF of the bootstrap distribution can be written as

$$F_s(y) = \frac{1}{n} \sum_{i=1}^{n(1-p_{\text{tail}})} \iota[y_{(i)} \leq y] + \iota[y \geq y_0] p_{\text{tail}} (1 - (y/y_0)^{-\hat{\alpha}}), \quad (110)$$

where $\iota(\cdot)$ is the indicator function (1). Indices of interests are functionals of the income distribution and so the index for this bootstrap distribution, \hat{W}_s , can be computed.⁵⁹ A bootstrap confidence interval can be computed as defined in (106), where $c_{0.025}^*$ and $c_{0.975}^*$ are the 2.5-th and 97.5-th percentiles of the EDF of the bootstrap t -statistics $t_b^* = (\hat{W}_b^* - \hat{W}_s) / \widehat{\text{var}}(\hat{W}_b^*)^{1/2}$. In practice, k and p_{tail} are chosen a priori. The number of observations k used to compute the Hill estimator can be selected such that $\hat{\alpha}$ does not vary significantly when more observations are taken, and p_{tail} can be chosen such that the re-sampling from the Pareto distribution is based on a smallest proportion of observation than k/n . It leads the previous authors to select $k = n^{1/2}$ and $p_{\text{tail}} = hk/n$ with $0 < h \leq 1$ in their experiments.

An alternative approach could be to generate bootstrap samples from a distribution estimated by finite mixture models. It allows us to estimate any density function, by allowing the number of components to vary, and, once the number of component is selected, to use a parametric distribution to generate bootstrap samples (see section 3.3). In order for the bootstrap to test a true null hypothesis, we need to compute the value of the welfare index for the mixture distribution, \hat{W}_m . With additively decomposable inequality measures, the index for the mixture distribution is easy to calculate, since the mixture distribution is a decomposition by groups. For instance, the class of Generalized Entropy (GE) indices can be expressed as a simple additive function of within-group and between-group inequality. Let there be K groups and let the proportion of the

⁵⁹For instance, the Theil index would be equal to $\hat{I}_s = \nu_s / \mu_s - \log \mu_s$, where $\mu_s = n^{-1} \sum_{i=1}^{n(1-p_{\text{tail}})} y_{(i)} + p_{\text{tail}} \hat{\alpha} y_0 / (\hat{\alpha} - 1)$ and $\nu_s = n^{-1} \sum_{i=1}^{n(1-p_{\text{tail}})} y_{(i)} \log y_{(i)} + p_{\text{tail}} [\log y_0 + 1 / (\hat{\alpha} - 1)] \hat{\alpha} y_0 / (\hat{\alpha} - 1)$.

	asym	boot	varstab	semip	mixture
Lognormal					
$\sigma = 0.5$	0.927	0.936	0.939	0.937	0.942
$\sigma = 1.0$	0.871	0.913	0.907	0.921	0.946
$\sigma = 1.5$	0.746	0.854	0.850	0.915	0.944
Singh-Maddala					
$q = 1.7$	0.915	0.931	0.933	0.926	0.928
$q = 1.2$	0.856	0.905	0.899	0.905	0.912
$q = 0.7$	0.647	0.802	0.796	0.871	0.789

Table 7: Coverage of asymptotic and bootstrap confidence intervals at the 95% level for the Theil index, for several bootstrap approaches, $n = 500$.

population falling in group k be p_k , the class of GE indices is equal to⁶⁰

$$I_{\text{GE}}^\xi = \sum_{k=1}^K p_k \left[\frac{\bar{y}_k}{\bar{y}} \right]^\xi I_{\text{GE},k}^\xi - \frac{1}{\xi^2 - \xi} \left(\sum_{k=1}^K p_k \left[\frac{\bar{y}_k}{\bar{y}} \right]^\xi - 1 \right) \quad (111)$$

where \bar{y}_k is the mean income in group k , \bar{y} is the mean income of the population ($\bar{y} = K^{-1} \sum_{k=1}^K p_k \bar{y}_k$), and $I_{\text{GE},k}^\xi$ is the GE index in group k . For an income distribution estimated as a finite mixture of lognormal distributions,

$$F_m(y) = \sum_{k=1}^K \hat{\pi}_k \Lambda(y; \hat{\mu}_k, \hat{\sigma}_k), \quad (112)$$

the value of the GE index is then equal to (111) with $p_k = \hat{\pi}_k$ and $\bar{y}_k = \exp(\hat{\mu}_k + \hat{\sigma}_k^2/2)$. The Gini index is not additively decomposable, but a formula can be found in Young (2011) for a mixture of lognormal distribution. Bootstrap samples are generated from the mixture distribution $F_m(y)$ and a bootstrap confidence interval is computed as defined in (106), where $c_{0.025}^*$ and $c_{0.975}^*$ are the 2.5-th and 97.5-th percentiles of the EDF of the bootstrap t -statistics $t_b^* = (\hat{W}_b^* - \hat{W}_m) / \widehat{\text{var}}(\hat{W}_b^*)^{1/2}$.

Table 7 presents coverage of asymptotic and bootstrap confidence intervals at the 95% level for the Theil index, with $n = 500$. The first two columns correspond to asymptotic (asym) and standard bootstrap (boot) methods: they reproduce the results given in Table 6, given here as benchmarks. The other columns show the results for the alternative bootstrap methods presented above. Results obtained by the approach proposed by Schluter (2012) are presented in the third column (varstab), bootstrapping a variance stabilizing transform of the Theil index. In the fourth column, the semi-parametric bootstrap proposed by Davidson and Flachaire (2007) and Cowell and Flachaire (2007) is used to generate bootstrap samples (semip), with $k = n^{1/2}$ and $h = 0.6$. Finally, bootstrap samples generated from a mixture of lognormal distributions is considered in the last column (mixture). The simulation results show that, in the presence

⁶⁰See Cowell (2011).

of very heavy-tailed distributions ($\sigma = 1.5$, $q = 0.7$), significant improvements can be obtained with alternative methods over asymptotic and standard bootstrap methods. However, none of the alternative methods provides very good results overall.

4.5.4 Testing equality of inequality measures

Confidence intervals are often used to make comparisons between two or more samples. The values of an index computed from independent samples are statistically different if the confidence intervals do not intersect. We can thus test if inequality or poverty measures are different between several countries, or over different periods of time, by comparing their confidence intervals. However, the previous results suggest that this approach may be unreliable when comparing inequality measures if the underlying distributions are quite heavy-tailed.

Another principal way of performing inference is by carrying out hypothesis tests. Testing the equality of inequality measures with a t -statistic, Dufour et al. (2013) show that almost exact inference can be obtained with permutation tests, if the samples come from distributions not too far away from each other, even with very heavy-tailed distributions and very small samples. They also show that this method outperforms other methods when the distributions are far away from each other.

Let us consider two independent samples $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_m\}$ assumed to be two sets of n and m independent observations from distributions F_x and F_y . The null hypothesis that an inequality measure W is the same in the two distributions, $H_0 : W_x = W_y$, can be tested with a t -statistic,

$$\tau = \left[\hat{W}_x - \hat{W}_y \right] / \left[\widehat{\text{var}}(\hat{W}_x) + \widehat{\text{var}}(\hat{W}_y) \right]^{\frac{1}{2}}, \quad (113)$$

where τ follows asymptotically the standard Normal distribution. A standard bootstrap approach would be to generate bootstrap samples X^* and Y^* by re-sampling with replacement, respectively, n observations from X and m observations from Y . The bootstrap samples are drawings from X and Y , from which an inequality measure would provide different numerical results. The null hypothesis tested from the original sample is then not respected in the bootstrap data generating process. A modified t -statistic has to be computed to test a true null hypothesis from a bootstrap sample (X_b^*, Y_b^*) ,

$$\tau_b^* = \left[\hat{W}_{x_b^*} - \hat{W}_{y_b^*} - (\hat{W}_x - \hat{W}_y) \right] / \left[\widehat{\text{var}}(\hat{W}_{x_b^*}) + \widehat{\text{var}}(\hat{W}_{y_b^*}) \right]^{\frac{1}{2}} \quad (114)$$

The bootstrap distribution is the EDF of the B bootstrap statistics, τ_b^* for $b = 1, \dots, B$; it is used as an approximation of the true distribution of the t -statistic τ .⁶¹

⁶¹Davidson and Flachaire (2007) consider testing the difference of two inequality measures with independent samples, but no significant improvement of their semi-parametric bootstrap method over standard bootstrap method is found.

In the permutation approach we generate samples by permuting the combined sample of N observations,

$$(X, Y) = \{x_1, \dots, x_n, y_1, \dots, y_m\}, \quad (115)$$

where $N = n + m$. The permutation samples X^* and Y^* are composed, respectively, by the first n and the remaining m observations of the permuted combined sample. Note that the combined sample can be permuted by resampling *without* replacement N observations from (X, Y) .

The permutation samples are drawings from the same set of observations, the null hypothesis tested from the original sample is then respected in the data generating process. From a permutation sample (X_p^*, Y_p^*) , the permutation t -statistic is,

$$\tau_p^* = \left[\hat{W}_{x_p^*} - \hat{W}_{y_p^*} \right] / \left[\widehat{\text{var}} \left(\hat{W}_{x_p^*} \right) + \widehat{\text{var}} \left(\hat{W}_{y_p^*} \right) \right]^{\frac{1}{2}}. \quad (116)$$

The permutation distribution is the EDF of the P permutation statistics, τ_p^* for $p = 1, \dots, P$; again it is used as an approximation of the true distribution of the t -statistic τ . If the underlying distributions are identical, permutation tests are known to provide *exact* inference in finite samples.⁶² Chung and Romano (2013) show that, under weak assumptions for comparing estimators, permutation tests are asymptotically valid when $F_x \neq F_y$, while retaining the exact rejection probability in finite samples when $F_x = F_y$.⁶³ Testing the equality of inequality measures, Dufour et al. (2013) show that almost exact inference is obtained with permutation tests in very small samples with heavy-tailed distributions, if the samples come from distributions not too far away from each other.

To illustrate, let us consider a simulation experiment concerned with testing the equality of the Gini index, $H_0 : I_{\text{Gini}}(F_x) = I_{\text{Gini}}(F_y)$. Data are generated from Singh-Maddala distributions, $\text{SM}(y; a, b, q)$, with parameters chosen such that the Gini index is identical in all distributions. The upper-tail of a Singh-Maddala distribution behaves like a Pareto distribution with shape parameter $\alpha = aq$.⁶⁴ Then, the smaller is α , the heavier is the upper tail. The sample size is very small, $n = m = 50$, and the distributions can be very heavy-tailed to stress-test the methods employed in testing. The number of replications is equal to 10 000 and the number of bootstrap and permutation samples are $B = P = 999$. We compute the rejection probability, or rejection frequency, as the proportion of p -values less than a nominal level equal to 0.05.⁶⁵ Inference is exact if the rejection probability is equal to 0.05.

⁶²See Fisher (1935), Dwass (1957), Good (2000), Dufour (2006).

⁶³Estimators have to be asymptotically linear and the use of a studentized statistic, as defined in (113), is crucial for the asymptotic validity of the permutation approach (it would not be valid with the statistic $\tau = \hat{W}_x - \hat{W}_y$).

⁶⁴Singh-Maddala distributions with parameters (a, q) equal to (2.5, 2.640350), (2.6, 2.218091), (2.7, 1.920967), (2.8, 1.7), (3.0, 1.3921126), (3.2, 1.1866026), (3.4, 1.0388049), (3.8, 0.8387663), (4.8, 0.5784599) and (5.8, 0.4473111), share the same (scale-invariant) Gini index, equals to 0.2887138. The tail parameters are, respectively, equal to

α_y	$\alpha_x = \alpha_y (F_x = F_y)$			$\alpha_x = 4.76 (F_x \neq F_y)$		
	asym	boot	permut	asym	boot	permut
6.60	0.0704	0.0609	0.0515	0.0752	0.0624	0.0510
5.77	0.0727	0.0609	0.0504	0.0762	0.0634	0.0515
5.19	0.0751	0.0606	0.0512	0.0761	0.0622	0.0508
4.76	0.0770	0.0614	0.0500	0.0770	0.0614	0.0500
4.18	0.0828	0.0636	0.0499	0.0796	0.0612	0.0498
3.80	0.0871	0.0656	0.0497	0.0825	0.0639	0.0503
3.53	0.0938	0.0664	0.0494	0.0865	0.0668	0.0518
3.19	0.1015	0.0686	0.0508	0.0956	0.0719	0.0565
2.78	0.1154	0.0685	0.0515	0.1138	0.0824	0.0707
2.59	0.1277	0.0693	0.0513	0.1289	0.0911	0.0800

Table 8: Rejection frequencies of asymptotic, bootstrap and permutation tests for testing the equality of Gini indices between two samples, when the underlying Singh-Maddala distributions are identical or different, $n = 50$.

Table 8 shows the empirical rejection frequencies of asymptotic, bootstrap and permutation tests for testing the equality of Gini indices between two samples, when the underlying distributions are identical and when they are different. As expected, when the distributions are identical, $F_x = F_y$, permutation tests provide exact inference, even in very small samples with very heavy-tailed distributions (column 4: permut). When the distributions are different, $F_x \neq F_y$, permutation tests provide almost exact inference, except when F_y is much more heavy-tailed than F_x (column 7: permut, when $\alpha_x = 4.76$ and $\alpha_y \leq 2.78$). Overall, permutation tests outperform asymptotic and bootstrap tests.

Hill plots can be useful for studying the tail behaviour in empirical studies. The “tail index” of a heavy-tailed distribution can be estimated by the Hill estimator on the k greatest-order statistics of a sample of size n , for $k \leq n - \text{see}$ equation (109). However, estimation results can be sensitive to the choice of k . Hill plots show the Hill estimate of the tail index, as a function of the number k of the greatest-order statistics used to compute it. An estimate of the tail index can be selected when the plot becomes stable about a horizontal straight line. For instance, Figure 15 shows Hill plots obtained from two samples of 1000 observations drawn from the Singh-Maddala distributions $SM(y; 2.8, 0.193, 1.7)$ and $SM(y; 5.8, 0.193, .447)$, with tail parameters respectively equal to 4.76 and 2.59, over the range of 0.5% and 25% of the greatest-order statistics used to compute it, with 95% confidence intervals (in gray). It is clear from this Figure that the second sample (on the right) comes from a much more heavy-tailed

$\alpha = 6.6, 5.77, 5.19, 4.76, 4.18, 3.80, 3.53, 3.19, 2.78, 2.59$.

⁶⁵For a two-tailed test, an asymptotic p -value is computed as $p_{as} = 2 \min(\Phi(\tau); 1 - \Phi(\tau))$. A bootstrap p -value is similar, but bootstrap distribution replaces the asymptotic one, $p_{boot} = 2 \min(\frac{1}{B} \sum_{b=1}^B \mathbb{1}(\tau_b^* \leq \tau); \frac{1}{B} \sum_{b=1}^B \mathbb{1}(\tau_b^* > \tau))$. A permutation p -value is similar to the bootstrap p -value, p_{boot} , with B and τ_b^* replaced by P and τ_p^* . The null hypothesis is rejected if the p -value is smaller than the nominal level.

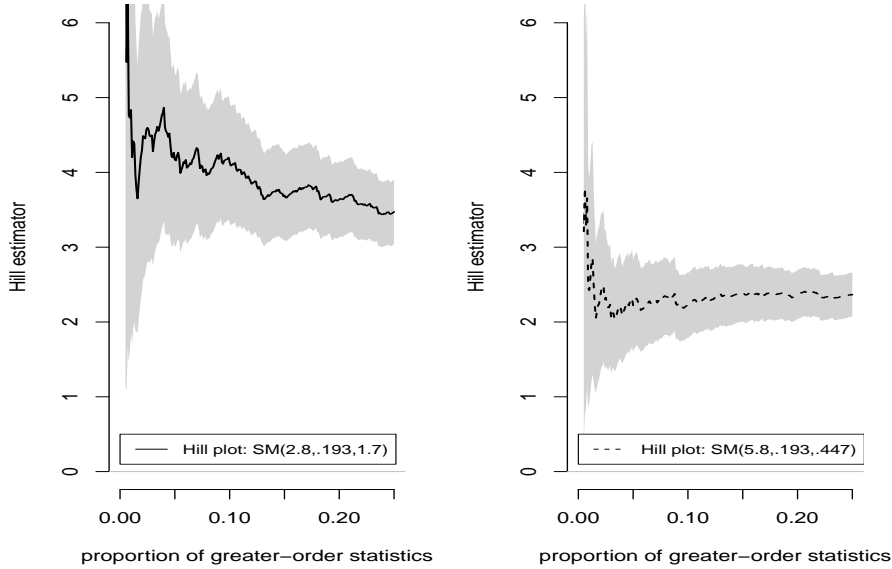


Figure 15: Hill plots: plot of the Hill estimate of the tail index (109) of two samples of 1000 observations drawn from Singh-Maddala distributions, as a function of the proportion of the greatest-order statistics used to compute it.

distribution than the first sample (on the left).⁶⁶

4.6 Parametric approaches

Sections 4.2 to 4.5 deal solely with distribution-free methods: in a sense we are working directly with the sample data. An alternative approach assumes that the distribution is known,⁶⁷ up to some parameters, and can be consistently estimated. A preliminary parametric estimation of the distribution is then obtained and the moments of the parametric distribution are estimated. When the distribution is parametric, inequality indices can be expressed as functions of the distribution parameters. Table 9 shows the formulas of the Theil, MLD and GE measures of inequality for the Lognormal and Pareto distributions. They are also given for the Generalized Beta distribution of the second kind (GB2), a four-parameter distribution defined in equation (5) when $c = 1$, $\Gamma(\cdot)$ is the gamma function and $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.⁶⁸

⁶⁶The Hill plot is not always revealing. The Hill estimator is designed for the Pareto distribution. But the Hill plot can be very volatile – and thus difficult to interpret – when the upper-tail of the underlying distribution is far from a Pareto; see Resnick (1997, 2007).

⁶⁷See section 3.1 for a discussion of the common functional forms that may be applied.

⁶⁸see Jenkins (2009)

The Singh-Maddala distribution is the special case of the GB2 distribution when $p = 1$; the Dagum distribution is the special case when $q = 1$ (see Figure 6). Then, to derive the expressions of the Theil, MLD and GE indices for the Singh-Maddala and Dagum distributions, set $p = 1$ and $q = 1$ in the equations given in the last column in Table 9. An inequality measure can be estimated by replacing the unknown parameters by consistent parameter estimates. Inequality measures are then expressed as non-linear functions of one or several consistent estimates. From the CLT, they are asymptotically Normal and their asymptotic variance can be derived using the delta method.⁶⁹

For several of the standard parametric distributions the Gini index can also be expressed fairly easily as a function of the unknown parameters of the underlying distribution. The Gini index for the Lognormal distribution, $\Lambda(y; \mu, \sigma^2)$, and for the Pareto distribution, $\Pi(y; \alpha)$, is equal to

$$I_{\text{Gini}}(\Lambda(y; \mu, \sigma^2)) = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1, \quad (117)$$

and

$$I_{\text{Gini}}(\Pi(y; \alpha)) = \frac{1}{2\alpha - 1} \quad (118)$$

respectively. For the Singh-Maddala distribution and for the Dagum distribution, defined in equation (5) when $c = 1$ and when, respectively, $p = 1$ and $q = 1$, the Gini index is equal to

$$I_{\text{Gini}}(\text{SM}(y; a, b, q)) = 1 - \frac{\Gamma(q)\Gamma(2q - \frac{1}{a})}{\Gamma(q - \frac{1}{a})\Gamma(2q)} \quad (119)$$

$$I_{\text{Gini}}(\text{D}(y; a, b, p)) = \frac{\Gamma(p)\Gamma(2p + \frac{1}{a})}{\Gamma(2p)\Gamma(p + \frac{1}{a})} - 1 \quad (120)$$

The Singh-Maddala and Dagum distributions are encompassed by the Generalized Beta distribution of the second kind (GB2), defined in equation (5) when $c = 1$, for which the formula of the Gini index can also be obtained. However, its expression is lengthy and involves the generalized hypergeometric function, see McDonald (1984) or Kleiber and Kotz (2003) for an explicit formula. Since the Gini index is defined as non-linear functions of one or several consistent estimates. From the CLT, it is asymptotically Normal and the asymptotic variance can be derived using the delta method.

5 Distributional comparisons

Apart from the simple welfare indices discussed in section 4, we also need to be able to implement *ranking tools*. These tools provide the researcher with intuitively appealing methods of making distributional comparisons and are associated with important results in the welfare economics of distributional analysis.

⁶⁹Software integrated commands can be used when calculations are cumbersome. Jenkins (2009) used the `ncom` command in STATA to compute standard errors of GE indices for the GB2 distribution.

	$\Lambda(y; \mu, \sigma^2)$	$\Pi(y; \alpha)$	$\text{GB2}(y; a, b, p, q)$
I_{GE}^1	$\frac{\sigma^2}{2}$	$\frac{1}{\alpha-1} - \log\left(\frac{\alpha}{\alpha-1}\right)$	$-\log\left(\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right) + \frac{\psi(p+\frac{1}{a})}{a} - \frac{\psi(q-\frac{1}{a})}{a}$
I_{GE}^0	$\frac{\sigma^2}{2}$	$\log\left(\frac{\alpha}{\alpha-1}\right) - \frac{1}{\alpha}$	$\log\left(\frac{\Gamma(p+\frac{1}{a})\Gamma(q-\frac{1}{a})}{\Gamma(p)\Gamma(q)}\right) - \frac{\psi(p)}{a} + \frac{\psi(q)}{a}$
I_{GE}^ξ	$\frac{e^{(\xi^2-\xi)\frac{\sigma^2}{2}}-1}{\xi^2-\xi}$	$\frac{\alpha}{\alpha-\xi}\left(\frac{\alpha}{\alpha-1}\right)^{-\xi}-1$	$\frac{1}{\xi^2-\xi}\left(\frac{\Gamma(p+\frac{\xi}{a})\Gamma(q-\frac{\xi}{a})\Gamma^{\xi-1}(p)\Gamma^{\xi-1}(q)}{\Gamma^\xi(p+\frac{1}{a})\Gamma^\xi(q-\frac{1}{a})}-1\right)$

Table 9: Parametric generalised entropy inequality measure for Lognormal (Λ), Pareto (Π) and Generalized Beta of the second kind (GB2) distributions.

5.1 Ranking and dominance: principles

The quantile and cumulation functionals Q and C defined in section 4.1 can be used to establish *dominance criteria* for income distribution comparisons in terms of welfare or inequality, and related concepts are available for comparisons in terms of poverty.

5.1.1 Dominance and welfare indices

First-order dominance

Using (23), for a given $F \in \mathbb{F}$, the graph $\{q, Q(F, q) : q \in \mathbb{Q}\}$ describes Pen's parade (Pen 1974). This is the basis for first-order distributional dominance (or first-order ranking) results. The concept of dominance can be explained as follows: consider two distributions $F, G \in \mathbb{F}$. Then F is said to *first-order dominate* G if the following pair of conditions hold:

$$\left. \begin{array}{l} \forall q \in \mathbb{Q} : Q(F, q) \geq Q(G, q), \\ \exists q \in \mathbb{Q} : Q(F, q) > Q(G, q). \end{array} \right\} \quad (121)$$

To see the importance of this concept suppose we consider the class of all indices expressible in the form⁷⁰ $W_{\text{AD}}(F)$ additive social-welfare functionals giving the aggregate of $\phi(y)$ where $\phi(\cdot)$ is some twice differentiable evaluation function of income. In particular take the important subclass where welfare respects the monotonicity principle – the evaluation of income is everywhere strictly increasing:

$$\mathbb{W}_1 := \left\{ W \mid W(F) = \int \phi(y) dF(y), \phi'(y) > 0 \right\}.$$

Then the statement “ $W(F) \geq W(G)$, for any $W \in \mathbb{W}_1$ ” is equivalent to the statement “ F first-order dominates G .” If the Parade graph of F lies somewhere above and nowhere below the Parade graph of G then welfare in F must be higher than in G , for any social-welfare function that respects monotonicity (Quirk and Saposnik 1962).

⁷⁰See equation (28)

Second-order dominance

The functional (25) can be used to characterise a number of standard concepts associated with second-order dominance.

- For a given $F \in \mathbb{F}$, the graph $\{q, C(F, q) : q \in \mathbb{Q}\}$ describes the *generalised Lorenz curve* (GLC) This is the basis for second-order distributional dominance results (Shorrocks 1983). The definition of second-order dominance can be derived from (121) just by replacing the quantile functional Q by the cumulant functional C . We also focus on a narrower subclass of welfare functions:

$$\mathbb{W}_2 := \left\{ W \mid W(F) = \int \phi(y) dF(y), \phi'(y) > 0, \phi''(y) \leq 0 \right\}.$$

The concavity restriction $\phi''(y) \leq 0$ implies that a transfer of income from a poorer to a richer individual can never increase social welfare; is a weak form of the transfer principle (Dalton 1920). Then the statement “ $W(F) \geq W(G)$, for any $W \in \mathbb{W}_2$ ” is equivalent to the statement “ F second-order dominates G .” If the GLC of F lies somewhere above and nowhere below the GLC of G then welfare in F must be higher than in G , for any social-welfare function that respects monotonicity and the transfer principle (Hadar and Russell 1969). However, in distributional analysis attention is focused not only on the basic principle of second order dominance, as just described, but also on restricted versions of this relationship that incorporate equivalence relationships on the members of \mathbb{F} .

- Suppose we want the second-order comparisons to be *scale independent*. This requires that, for any $F \in \mathbb{F}$ and any $\lambda > 0$ the distribution of y and of y/λ are regarded as equivalent for the the purposes of distributional comparison; this implies that, when comparing distributions we may divide incomes by an arbitrary positive constant. A natural choice for this constant is the mean of the distribution. The scale normalisation of the GLC by the mean (26) gives the (*relative*) Lorenz functional:⁷¹

$$L(F; q) := \frac{C(F; q)}{\mu(F)} \tag{122}$$

and the graph $\{q, L(F; q) : q \in \mathbb{Q}\}$ gives the *relative Lorenz curve* (RLC).

- As an alternative to scale independence, we might be interested in a form of origin independence for the distributional comparisons which would require that, for any $F \in \mathbb{F}$ and any $\delta \in \mathbb{R}$ the distribution of y and of $y + \delta$ are regarded as equivalent. Instead of the scale normalisation used in defining the relative Lorenz curve we impose a “translational”

⁷¹This is equivalent to the income share (27).

normalisation so as to define the absolute Lorenz curve. This is the graph $\{q, A(F; q) : q \in \mathbb{Q}\}$, where

$$A(F; q) := C(F; q) - q\mu(F).$$

5.1.2 Stochastic dominance

The first-order and second-order dominance previously defined can be encompassed in a unified method, stochastic dominance, and extended to higher-order dominance.⁷² Let us define dominance curves as follows:

$$D_F^s(y) := \frac{1}{(s-1)!} \int_0^y (y-t)^{s-1} dF(t). \quad (123)$$

Distribution F is said to dominate distribution G stochastically at order s if the following pair of conditions hold:

$$\left. \begin{array}{l} \forall y \in \mathbb{Y} : D_F^s(y) \leq D_G^s(y), \\ \exists y \in \mathbb{Y} : D_F^s(y) < D_G^s(y). \end{array} \right\} \quad (124)$$

The case $s = 1$ corresponds to first-order dominance based on Pen's parade, previously defined in (121). Indeed, first-order stochastic dominance of G by F implies that $F(y) \leq G(y)$ for all y and there exists y over some interval for which the inequality holds strictly. It is similar to (121) expressed in terms of the quantile functions rather than CDFs.

The case $s = 2$ corresponds to second-order dominance based on the generalised Lorenz curve (GLC), previously defined. Indeed, from (123) and (25), we have $D_F^2(y_q) - D_G^2(y_q) = C(G; q) - C(F; q)$. Then, the pair of conditions in (124) is similar to that in (121) where the quantile functional Q is replaced by the cumulant functional C .

There is a clear relation between dominance and poverty. From (86) and (123), we can see that $D_F^s(\zeta_0)$ is equal to the FGT poverty index, up to a scale factor. If, for all $[\zeta_0^-; \zeta_0^+]$, $D_F^s(\zeta_0) < D_G^s(\zeta_0)$, it follows that the FGT poverty index is lower in F than in G for all poverty lines in the interval $[\zeta_0^-; \zeta_0^+]$. The poverty measure can then be viewed as restricted stochastic dominance of G by F over that interval. The stochastic dominance criterion can also be viewed as a generalisation of poverty measures when we let the poverty line vary over the whole support of the distribution.⁷³

5.2 Ranking and dominance: implementation

In order to implement ranking criteria empirically a standard approach is as follows:⁷⁴

⁷²Fishburn (1980), O'Brien (1984), Stark and Yitzhaki (1988), Thistle (1989), O'Brien and Scarsini (1991), Fishburn and Lavallo (1995), Davidson (2008).

⁷³See Atkinson (1987) and Foster and Shorrocks (1988).

⁷⁴We consider distribution-free methods. For parametric Lorenz curve comparisons, see Sarabia (2008).

1. Choose a finite collection of population proportions $\Theta \subset \mathbb{Q}$.
2. For each $q \in \Theta$ compute the sample quantiles \hat{y}_q and income cumulations \hat{c}_q required for empirical implementation of first- and second-order rankings. To do this, we replace F in (23) and (25) by the EDF $F^{(n)}$ – see equation (8). Then we have

$$\hat{y}_q := Q\left(F^{(n)}; q\right) = y_{(\kappa(n,q))}, \quad (125)$$

where

$$\kappa(n, q) := \lfloor nq - q + 1 \rfloor \quad (126)$$

and $\lfloor x \rfloor$ denotes the largest integer no greater than x ; we also have

$$\hat{c}_q := C\left(F^{(n)}; q\right) = \frac{1}{n} \sum_{i=1}^{\kappa(n,q)} y_{(i)} \quad (127)$$

3. Compute the variances and covariances of the sample quantiles (first-order) or the income cumulations (second order)
4. Specify carefully the ranking hypothesis that is to be tested.

Step 1 – involves a choice of how many points to select on the Parade or on the Lorenz curve. Step 2 is easy. Step 3 is dealt with in section 5.2.1 and Step 4 in sections 5.2.3 and 5.2.4.

5.2.1 Asymptotic distributions

The main results follow from applying Lemmas in section 4.2.2. We also need to define one further functional analogous to (23) and (25):

$$S(F; q) := \int_{\underline{y}}^{y_q} y^2 dF(y) =: s_q. \quad (128)$$

and its sample counterpart:

$$\hat{s}_q := S\left(F^{(n)}; q\right) = \frac{1}{n} \sum_{i=1}^{\kappa(n,q)} y_{(i)}^2 \quad (129)$$

Then we have the following two theorems:

Theorem 1 *For any $q, q' \in \mathbb{Q}$, $\sqrt{n}\hat{y}_q$ and $\sqrt{n}\hat{y}_{q'}$ are asymptotically normally distributed with covariance.⁷⁵*

$$\frac{q[1-q']}{f(y_q)f(y_{q'})}. \quad (130)$$

⁷⁵See Lemma 1 of Beach and Davidson (1983).

Proof. Immediate from Lemmas 1 and 2 ■

Theorem 2 For any $q, q' \in \mathbb{Q}$, $\sqrt{n}\hat{c}_q$ and $\sqrt{n}\hat{c}_{q'}$ are asymptotically normally distributed with covariance:⁷⁶

$$\omega_{qq'} := s_q + [qy_q - c_q][y_{q'} - q'y_{q'} + c_{q'}] - y_q c_{q'} \quad \text{for } q \leq q'. \quad (131)$$

Proof. Using Lemmas 1 and 3 the asymptotic covariance of $\sqrt{n}C(F^{(n)}; q)$ and $\sqrt{n}C(F^{(n)}; q')$ is given by

$$\begin{aligned} \omega_{qq'} &= \int IF(z; C(F; q), F)IF(z; C(F; q'), F) dF(z) \\ &= \int [qy_q - c_q + \iota(y_q \geq z)[z - y_q]] [q'y_{q'} - c_{q'} + \iota(y_{q'} \geq z)[z - y_{q'}]] dF(z) \end{aligned} \quad (132)$$

Given that $\iota(x_{q'} \geq z) = 1$ whenever $\iota(x_q \geq z) = 1$ the right-hand side becomes

$$\begin{aligned} &[qy_q - c_q][q'y_{q'} - c_{q'}] + \int_{\underline{y}}^{y_{q'}} [qy_q - c_q][z - y_{q'}] dF(z) \\ &+ \int_{\underline{y}}^{y_q} [q'y_{q'} - c_{q'} + z - y_{q'}][z - y_q] dF(z) \end{aligned} \quad (133)$$

Using the definitions in (23), (25) and (128) we find that (133) becomes (131)

■

We can also rewrite the influence function as a random variable minus its expectation. From (42) in Lemma 3, we have

$$IF(z; C(F, q), F) = Z_q - E(Z_q) \quad \text{where } Z_q = [z - y_q]\iota(z \leq y_q). \quad (134)$$

From (132), we can see immediately that the asymptotic covariance of $\sqrt{n}\hat{c}_q$ and $\sqrt{n}\hat{c}_{q'}$ is equal to the covariance of Z_q and $Z_{q'}$

$$\omega_{qq'} = \text{cov}(Z_q, Z_{q'}). \quad (135)$$

From a sample y_i , for $i = 1, \dots, n$, we can then estimate the covariance of the generalised Lorenz curve ordinates \hat{c}_q and $\hat{c}_{q'}$ as the empirical covariance of Z_{iq} and $Z_{iq'}$ divided by n

$$\widehat{\text{cov}}(\hat{c}_q, \hat{c}_{q'}) = \frac{1}{n} \widehat{\omega}_{qq'} = \frac{1}{n^2} \sum_{i=1}^n (Z_{iq} - \bar{Z}_q)(Z_{iq'} - \bar{Z}_{q'}) \quad (136)$$

where

$$Z_{iq} = [y_i - \hat{y}_q]\iota(y_i \leq \hat{y}_q) \quad (137)$$

and $\bar{Z}_q = n^{-1} \sum_{i=1}^n Z_{iq}$, \hat{y}_q is given in (125).

⁷⁶See Theorem 1 of Beach and Davidson (1983).

In a practical implementation one replaces the individual components of the right-hand side of (131) by their sample counterparts⁷⁷ to obtain the following consistent estimate of $\omega_{qq'}$:

$$\widehat{\omega}_{qq'} := \widehat{s}_q + [q\widehat{y}_q - \widehat{c}_q][\widehat{y}_{q'} - q'\widehat{y}_{q'} + \widehat{c}_{q'}] - \widehat{y}_q\widehat{c}_{q'}. \quad (138)$$

These results can also be used for the ordinates of the (relative) Lorenz curve. Using the standard result on limiting distributions of differentiable functions of random variables (Rao 1973), or using the delta method in (60), the asymptotic covariances of $\sqrt{n}\widehat{c}_q/\widehat{\mu}$ and $\sqrt{n}\widehat{c}_{q'}/\widehat{\mu}$ are then given by

$$v_{qq'} = \frac{1}{\mu^4} [\mu^2\omega_{qq'} + c_q c_{q'}\omega_{11} - \mu c_q\omega_{q'1} - \mu c_{q'}\omega_{q1}] \quad \text{for } q \leq q'. \quad (139)$$

where $\omega_{q1} := s_q + [qy_q - c_q]\mu - y_q c_q$, $\omega_{11} := s_1 - \mu^2$ and $\mu = \mu(F)$. Again, for practical implementation, the components of the right-hand side of (139) are replaced by their sample counterparts. We can also use the influence function and express it as a random variable minus its expectation. The *IF* of the Lorenz curve ordinate (122) is given by⁷⁸

$$IF(z; L(F, q), F) = \frac{1}{\mu} \left[qy_q - \frac{zc_q}{\mu} + [z - y_q]\iota(z \leq y_q) \right] \quad (140)$$

We can rewrite the influence function as $IF(z; L(F, q), F) = Z_q - E(Z_q)$, where

$$Z_q = \frac{1}{\mu^2} \left[\mu[z - y_q]\iota(z \leq y_q) - c_q z \right]. \quad (141)$$

The asymptotic covariance of $\sqrt{n}\widehat{c}_q/\widehat{\mu}$ and $\sqrt{n}\widehat{c}_{q'}/\widehat{\mu}$ is equal to the covariance of Z_q and $Z_{q'}$. From a sample y_i , for $i = 1, \dots, n$, we can then estimate the covariance of the (relative) Lorenz curve ordinates $\widehat{c}_q/\widehat{\mu}$ and $\widehat{c}_{q'}/\widehat{\mu}$ as the empirical covariance of Z_{iq} and $Z_{iq'}$ divided by n ,

$$\widehat{\text{cov}} \left(\frac{\widehat{c}_q}{\widehat{\mu}}, \frac{\widehat{c}_{q'}}{\widehat{\mu}} \right) = \frac{1}{n} \widehat{v}_{qq'} = \frac{1}{n^2} \sum_{i=1}^n (Z_{iq} - \bar{Z}_q)(Z_{iq'} - \bar{Z}_{q'}) \quad (142)$$

where

$$Z_{iq} = \frac{1}{\widehat{\mu}^2} \left[\widehat{\mu}[y_i - \widehat{y}_q]\iota(y_i \leq \widehat{y}_q) - \widehat{c}_q y_i \right]. \quad (143)$$

and $\bar{Z}_q = n^{-1} \sum_{i=1}^n Z_{iq}$, \widehat{y}_q and \widehat{c}_q are given by (125) and (127) respectively.

The case of stochastic dominance can also be considered. For a given value z , a consistent estimator of $D_F^s(z)$ is

$$\widehat{D}_F^s(z) = \frac{1}{n(s-1)!} \sum_{i=1}^n (z - y_i)^{s-1} \iota(y_i \leq z) \quad (144)$$

⁷⁷ \widehat{y}_q , \widehat{c}_q and \widehat{s}_q are given by (125), (127) and (129) respectively.

⁷⁸See Cowell and Victoria-Feser (2002), Donald et al. (2012).

where $y_i, i = 1, \dots, n$ is a random sample of n independent observations. Since it is a sum of independent and identically distributed (IID) observations, this estimator is consistent and asymptotically normal. The asymptotic covariance is also easy to calculate.⁷⁹

When we compare two distributions, random samples can be obtained from two independent populations or from two correlated populations. The last case typically occurs when the two samples are independent paired drawings from the same population, as for instance with pre-tax and post-tax distributions. In both cases of independent and correlated samples, it can be shown that the difference $\hat{D}_F^s(z_q) - \hat{D}_G^s(z_{q'})$ is asymptotically normal, with asymptotic covariance equal to

$$\frac{1}{((s-1)!)^2} E[(z_q - y_F)_+^{s-1} (z_{q'} - y_G)_+^{s-1}] - D_F^s(z_q) D_G^s(z_{q'}) \quad (145)$$

where $(x)_+^{s-1} = x^{s-1} \iota(x \geq 0)$. This result comes from the central limit theorem, assuming that population moments of order $2s-2$ for each distribution exist. The asymptotic covariance can be estimated with sample counterparts, with the expectation in (145) replaced by

$$\frac{1}{n} \sum_{i=1}^n (z_q - y_{F^{(n)},i})_+^{s-1} (z_{q'} - y_{G^{(n)},i})_+^{s-1} \quad (146)$$

and $D^s(x)$ estimated as defined in (144). For $s = 2$, we find an estimate of the covariance matrix similar to that obtained in (136) and (137), for the generalised Lorenz curve ordinates. More details and explicit expressions for z being stochastic and for poverty measures can be found in the comprehensive approach to inference on stochastic dominance presented in Davidson and Duclos (2000).⁸⁰

5.2.2 Dominance: an intuitive application

Armed with theorems 1 and 2 an intuitive approach to dominance can be immediately applied. Using (127) we can plot an empirical Generalised Lorenz curve with confidence bands. Consistent estimates of the variance of the Generalised Lorenz curve ordinates can be calculated using (136) and (137) with $q = q'$. Therefore we can immediately construct an informative graphical presentation for distributional comparisons, (q, \hat{c}_q) , with 95% confidence bands computed as $[\hat{c}_q \pm 1.96 \times \widehat{\text{var}}(\hat{c}_q)]$. One could see whether it is reasonable to conclude that the GLC for distribution F lies above that for distribution G (second-order dominance). Clearly the same idea could be pursued with empirical quantiles and parade diagrams (first-order dominance).

⁷⁹It is equal to (145) with $F = G$.

⁸⁰On dominance with complex sample design see Beach and Kaliski (1986), Zheng (1999, 2002). For an alternative approach focusing on crossings in the tails of Lorenz curves see Schluter and Trede (2002) and for a Bayesian approach see Hasegawa and Kozumi (2003). On the extension to absolute dominance and deprivation dominance see Bishop et al. (1988), Xu and Osberg (1998) and on poverty dominance see also Chen and Duclos (2008), Thuysbaert (2008).

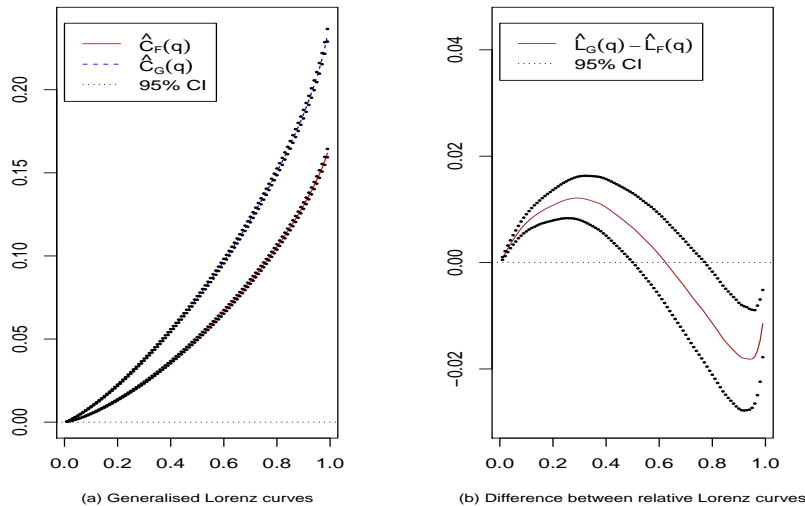


Figure 16: Generalised Lorenz curves and difference between Lorenz curves, $n = 5\,000$

Figure 16(a), to the left, shows two generalised Lorenz curves obtained from two independent samples of 5 000 observations drawn from Singh-Maddala distributions F and G respectively, with confidence bands at 95% evaluated at the percentiles, $q = 0.01, 0.02, \dots, 0.99$. F is the Singh-Maddala distribution with parameters $a = 2.8$, $b = 0.193$ and $q = 1.7$, used in the introduction, and G is the distribution with parameters $a = 3.8$, 0.193 and 0.839 ; the means are, respectively, 0.169 and 0.240 . This figure shows that distribution G second-order dominates distribution F . It suggests that poverty measures based on poverty gaps will exhibit more poverty in F than in G (Jenkins and Lambert 1997). Table 10 shows poverty measures computed from the two samples, with 95% confidence intervals (see section 4.4). As expected, poverty indices are significantly greater in F than in G .

We can also plot an empirical (relative) Lorenz curve with confidence bands. Consistent estimates of the variance of the (relative) Lorenz curve ordinates, $\widehat{\text{var}}(\hat{c}_q/\hat{\mu})$, can be calculated using (142) and (143) with $q = q'$. Therefore, we can construct a graphical representation of Lorenz curves, $(q, \hat{c}_q/\hat{\mu})$, with 95% confidence bands, $[(\hat{c}_q/\hat{\mu}) \pm 1.96 \times \widehat{\text{var}}(\hat{c}_q/\hat{\mu})]$. It is often difficult to distinguish between two relative Lorenz curves by eye: for this reason a plot of the difference between two relative Lorenz curves is often useful. When the samples are independent, the variance of the difference between the ordinates is the sum of the variances from each sample. In practice, (relative) Lorenz curves often intersect and in such cases no unambiguous ranking can be obtained. Nevertheless, useful information on inequality can be drawn from Lorenz curve comparisons.

	Distribution F		Distribution G	
	Index	CI _{95%}	Index	CI _{95%}
Poverty measures*				
P_{FGT}^0	0.1140	[0.1052;0.1228]	0.0180	[0.0143;0.0217]
P_{FGT}^1	0.0329	[0.0297;0.0360]	0.0038	[0.0028;0.0048]
P_{Sen}	0.0460	[0.0417;0.0503]	0.0055	[0.0041;0.0070]
P_{SST}	0.0635	[0.0575;0.0695]	0.0075	[0.0055;0.0096]
Generalised Entropy measures				
I_{GE}^{-1}	0.1998	[0.1858;0.2137]	0.1551	[0.1455;0.1647]
I_{GE}^0	0.1520	[0.1448;0.1591]	0.1418	[0.1331;0.1506]
I_{GE}^1	0.1458	[0.1376;0.1539]	0.1564	[0.1435;0.1693]
I_{GE}^2	0.1693	[0.1538;0.1847]	0.2164	[0.1863;0.2465]
I_{Gini}	0.2937	[0.2869;0.3005]	0.2920	[0.2833;0.3007]

* The poverty line is half the median of the sample drawn from distribution F : $\zeta_0 = 0.07517397$.

Table 10: Inequality and poverty measures, with confidence intervals at 95%, computed from two samples of 5 000 observations drawn independently from F and G

Figure 16(b), to the right, shows the difference between two relative Lorenz curves obtained from the two samples used in Figure 16(a). The two curves intersect in the upper part. This figure also shows that the empirical Lorenz curve of G lies significantly above the empirical Lorenz curve of F in the bottom part, whereas the reverse is true in the upper part. It suggests that inequality measures sensitive to the bottom part of income distributions would be smaller in G than in F , while inequality measures more sensitive to the upper part of income distributions would be smaller in F than in G . Table 10 shows inequality measures computed from the two samples, with 95% confidence intervals (see sections 4.3) Noting that Generalised Entropy (GE) measures, I_{GE}^ξ , are more sensitive to the bottom (top) of income distributions with smaller (higher) parameter ξ , we find the results suggested by the previous Lorenz curve comparisons. Indeed, I_{GE}^{-1} is significantly smaller in G than in F , while I_{GE}^2 is significantly smaller in F than in G .

This approach is clearly ad hoc and we need to examine the issues involved more carefully; we do this in sections 5.2.3 and 5.2.4. Graphical representation of two empirical Lorenz curves, with confidence intervals, allows us to make *individual* comparisons. We can test if each individual Lorenz curve ordinates are significantly different between two curves. To be able to make conclusions on dominance or non-dominance, we need to test simultaneously that all ordinates from one curve are significantly greater or not smaller than the ordinates from the other curve. Appropriate test statistics need to be used to make *multiple* comparisons and to test simultaneously that several inequalities hold. Moreover, Lorenz curve ordinates are typically strongly positively correlated and, thus, test

statistics need to take into account the covariance structure between the Lorenz curve ordinates.

5.2.3 The null hypothesis: dominance or non-dominance

Performing inference on stochastic dominance is more complex than on a single welfare index. The hypotheses tested are usually based on a set of inequalities. For instance, first-order stochastic dominance requires that,

$$F(y) \leq G(y) \quad \text{for all } y \geq 0, \quad (147)$$

in order to say that distribution F dominates distribution G stochastically at order one. The theoretical literature also include the condition that $F(y) < G(y)$ for some y , as defined in (124). However, no statistical test can distinguish between these two forms of weak and strict dominance.⁸¹ Since we are interested on statistical issues hereafter, we make no distinction between weak and strict dominance and we can write all inequalities as weak.

Inference on dominance in the population would be drawn from the corresponding sample properties. From a given sample, we can consistently estimate the two distributions by their EDF counterparts, $F^{(n)}(x)$ and $G^{(n)}(x)$. Sample dominance is then defined as $F^{(n)}(y) \leq G^{(n)}(y)$ for all y . It is clear that dominance in the population cannot be rejected if there is dominance in the sample. It is rejected if sample non-dominance is statistically significant only. A similar reasoning applies for non-dominance in the population. It follows that, to infer dominance, we should test the null hypothesis of non-dominance, and, to infer non-dominance, we should test the null of dominance.

It can be illustrated with a simple example of two distributions with the same support and three points, $y_1 < y_2 < y_3$.⁸² Since $F(y_3) = G(y_3) = 1$, we will say that distribution F dominates distribution G in the population if $d_i = G(y_i) - F(y_i) \geq 0$ for $i = 1, 2$. Figure 17 shows two bi-dimensional plots of \hat{d}_1 and \hat{d}_2 , where the null hypothesis is, respectively, dominance and non-dominance. Distribution F dominates G in the sample when $\hat{d}_i \geq 0$, for $i = 1, 2$. Then, the first quadrant, denoted I (grey area), corresponds to dominance in the sample, while the quadrants II, III and IV correspond to non-dominance in the sample.

First, let us consider the null hypothesis of dominance, as shown in Figure 17a, to the left. To reject dominance in the population, the non-dominance in the sample must be statistically significant, that is, the rejection zone has to be far enough from the dominance area, for example, in the shading lines area. The rejection zone is exclusively in the area of non-dominance, while the (remaining) non-rejection zone corresponds to the dominance area plus the white L-shaped band within the non-dominance area. Then, rejecting the null hypoth-

⁸¹Under the null of an inequality in one direction, a test cannot reject equality. Indeed, equality is on the frontier of the inequality hypothesis and, a test cannot distinguish statistically between being on the frontier and being very close to the frontier.

⁸²see Davidson and Duclos (2013).

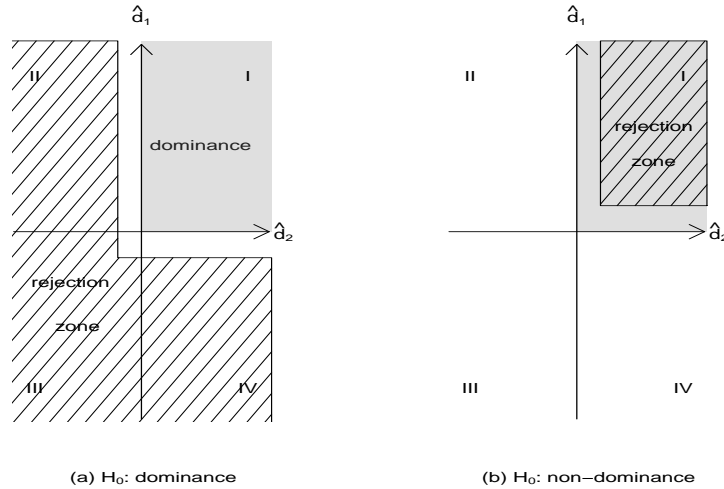


Figure 17: Tests of dominance and non-dominance. The first quadrant, I, corresponds to dominance of G by F in the sample (grey area). The quadrants II, III and IV correspond to non-dominance.

esis of dominance corresponds to the case of non-dominance, while non-rejecting it is inconclusive.

Second, let us consider the null hypothesis of non-dominance, as shown in Figure 17b, to the right. With a similar reasoning, we can see that the rejection zone is exclusively in the area of dominance, while the non-rejection zone is composed of both non-dominance and dominance situations (grey L-shape band). Then, rejecting the null hypothesis of non-dominance corresponds to the case of dominance, while non-rejecting it is inconclusive.

The previous example illustrates that, positing the null of non-dominance is the only way to draw strong conclusion of dominance. However, it comes at cost: dominance will be inferred if there is strong evidence in its favour only. From Figure 17b, we can see that rejecting the null of non-dominance is quite demanding, since it requires that both statistics \hat{d}_1 and \hat{d}_2 are statistically significant. It may be too demanding, especially in the tails where both distributions tend to the same values and where we usually have sparse data and little information. Davidson and Duclos (2013) show that, with distributions continuous in the tails, it is impossible to reject the null of non-dominance over the full support of the distributions. It leads them to develop restricted stochastic dominance, limiting attention to some interval in the middle of the distribution.

The most common approach in the literature has developed tests of stochas-

tic dominance positing the null of dominance.⁸³ The previous example illustrates the standard feature in statistics that, non-rejecting the null does not imply that the null is true, and so selecting the null hypothesis remains allows for the possibility of being wrong at some level.⁸⁴ The level at which we may be wrong by accepting the null is unknown (L-shape bands in Figures 17a and 17b), but it would be reduced by using statistical tests with greater power properties in finite sample.

Finally, both approaches can be seen as complementary. Rejecting the null of dominance or non-dominance allows us to infer, respectively, non-dominance and dominance when comparing two distributions.

5.2.4 Hypothesis testing

Test statistics have been developed in the literature under the null hypothesis of dominance and non-dominance. We distinguish both cases, for which we can interpret them, respectively, as union-intersection and intersection-union tests.

Under the null of dominance

Statistical test can be constructed to test the null hypothesis of dominance, against the alternative of non-dominance. Under the null hypothesis that F dominates G , we have

$$\begin{aligned} H_0 : & D_F^s(y) \leq D_G^s(y), \quad \text{for all } y \in \mathbb{Y}, \\ H_1 : & D_F^s(y) > D_G^s(y), \quad \text{for some } y \in \mathbb{Y}. \end{aligned} \tag{148}$$

where \mathbb{Y} denotes a given set contained in the union of the support of the two distributions. An appropriate test statistic could be interpreted as a *union-intersection* test, since the null hypothesis is expressed as an intersection of individual hypotheses, and the alternative as an union (Roy 1953). A natural test is based on the supremum of individual differences,

$$\tau = \sup_{y \in \mathbb{Y}} \left(\hat{D}_F^s(y) - \hat{D}_G^s(y) \right). \tag{149}$$

It is clear that the null hypothesis is rejected if τ is significant and positive. McFadden (1989) proposed a test based on (149) for two independent samples of IID observations. For $s = 1$, it is a variant of the Kolmogorov-Smirnov statistic, with known properties. For $s = 2$, the asymptotic distribution under the null is not tractable. Barrett and Donald (2003) proposed simulation-based methods for estimating critical values, taking into account comparisons at all points of the support (functional approach) rather than at a fixed number of

⁸³See Beach and Richmond (1985), McFadden (1989), Bishop et al. (1992), Anderson (1996), Schmid and Trede (1996), Davidson and Duclos (2000), Barrett and Donald (2003), Linton et al. (2005), Maasoumi and Heshmati (2008).

⁸⁴It is usually done when a coefficient is not significant in estimation results and, when the analysis that follows is based on the selected regression model without the associated covariate.

arbitrarily chosen points. Linton et al. (2005) proposed to use subsampling methods, permitting to estimate critical values in general settings, with arbitrary order s , dependent observations, continuous and discrete supports. For multiple comparisons restricted to a fixed number of points (y_1, \dots, y_T) , a Wald test of inequality restrictions can also be used. Let us note the covariance matrix estimates of \hat{D}_F^s and \hat{D}_G^s , respectively, as $\hat{\Omega}_F$ and $\hat{\Omega}_G$, the Wald test statistic is computed by solving

$$\min_{\delta \geq 0} (\hat{D}_F^s - \hat{D}_G^s - \delta)^\top (\hat{\Omega}_F + \hat{\Omega}_G) (\hat{D}_F^s - \hat{D}_G^s - \delta). \quad (150)$$

The statistic is obtained by using an algorithm to solve quadratic programming problems. The distribution of the statistic is a mixture of chi-square with weights that require simulation methods to be consistently estimated (Dardanoni and Forcina 1999).

Lorenz dominance can be tested using similar methods. Bishop et al. (1989) and Davidson and Duclos (1997) proposed a test for a fixed number of points,⁸⁵ while Donald and Barrett (2004) and Bhattacharya (2007) have considered versions of Lorenz dominance tests in a functional approach, taking into account comparisons at all points of the supports.

Under the null of non-dominance

Other statistical tests have been developed to test the null hypothesis of non-dominance, against the alternative of dominance. Under the null that F does not dominate G , we have

$$\begin{aligned} H_0 : & D_F^s(y) \geq D_G^s(y), \quad \text{for some } y \in \mathbb{Y}, \\ H_1 : & D_F^s(y) < D_G^s(y), \quad \text{for all } y \in \mathbb{Y}. \end{aligned} \quad (151)$$

An appropriate test could be interpreted as an *intersection-union* test, since the null hypothesis is expressed as an union and the alternative as an intersection of individual hypotheses (Gleser 1973). The idea behind the intersection-union method is that the null is rejected only if each of the individual hypotheses can be rejected. A natural test is based on the infimum of individual differences,

$$\tau' = \inf_{y \in \mathbb{Y}^r} (\hat{D}_G^s(y) - \hat{D}_F^s(y)). \quad (152)$$

It is clear that the null hypothesis is rejected if τ' is significant and positive. The statistic τ' has to be defined over \mathbb{Y}^r , some closed interval contained in the interior of the joint support of the two distributions \mathbb{Y} . The main reason is that the null hypothesis would never be rejected if we consider the tails of the distributions, where data are sparse and where the differences between the two distributions tend to zero. Specifically, \mathbb{Y}^r should be a restricted interval in \mathbb{Y} that removes the tails of the distributions. Kaur et al. (1994) proposed a test based on (152) for $s = 2$ with independent samples and continuous distributions

⁸⁵see also Bishop et al. (1991a, 1991b, 1992)

F and G . Critical values can be taken from the Normal distribution, making the test easy to implement. However, it can have low power properties (Dardanoni and Forcina 1999). Davidson and Duclos (2013) and Davidson (2009b) proposed a test for higher order, for correlated samples as well as uncorrelated samples, and for continuous and discrete distributions. They also show that appropriate bootstrap methods permit to obtain much better finite sample properties.

6 Other estimation problems

In sections 4 and 5 we assumed that data are always drawn from a representative sample of the whole population. For some researchers this state of affairs is something of a luxury. In this section we discuss a number of common problems that need to be taken into account in practical application and the statistical methods of dealing with them.

6.1 Contamination

By data contamination we mean a set of observations that do not “belong” to the sample – see section 2.3. The essentials of the formal approach can be explained using a simple model based on the distribution function (in fact we have already seen the elements of this model in a different context – see section 4.2). The idea is that, instead of observing a distribution $F \in \mathbb{F}$ directly, one sees it after it has been mixed with another distribution that represents contamination. The elementary model of this is presented in equation (33) where one observes a distribution G given by

$$G = [1 - \delta]F + \delta H^{(z)} \quad (153)$$

where δ represents the proportion of contamination in the mixture that we observe and $H^{(z)}$ is the elementary “contamination distribution” (32) a single point mass at $z \in \mathbb{Y}$. From this minimalist structure one can easily develop more interesting specifications of the model of contamination using a mixture of F with a distribution that is richer than $H^{(z)}$. A number of questions immediately arise: Does contamination matter in analysing income distributions? How does contamination affect distributional comparisons? How may one appropriately estimate models of income distribution if there is reason to believe that contamination is an important issue?

6.1.1 The concept of robustness

To address the question “how important”, we can use the tool introduced in the discussion of asymptotic inference (section 4.2). The Influence Function is precisely designed to gauge the sensitivity of a statistic to contamination. Consider some statistic T (for example an inequality measure, a poverty index or a Lorenz ordinate): then IF quantifies the impact of an infinitesimal amount of contamination on the statistic T , namely $\frac{\partial}{\partial \delta} T(G)|_{\delta \rightarrow 0}$ (assuming that T is

differentiable) – see Hampel (1971, 1974), Hampel et al. (1986). Clearly the size of this differential will depend on the exact specification of the contamination distribution: in the context of the elementary model (153) this would mean that it will depend on the exact location in \mathbb{Y} of the point z (where the contamination is concentrated). Of particular interest are cases where this IF is unbounded for some value of z : the interpretation of this is that the statistic T is highly sensitive to an infinitesimal amount of contamination at point z . In the present context this is precisely what we mean by saying that a statistic is *non-robust*; obviously if the IF for the statistic T is bounded for all values of z then it makes sense to describe T as a *robust* statistic: we will come to some examples of robust and non-robust statistics in a moment. However, first it is worth making the common-sense point that even if we are only using robust statistics in our analysis, this does not mean that we can ignore the possibility of data contamination: in practice it may be that the assumption that δ is vanishingly small is just unreasonable.

6.1.2 Robustness, welfare indices and distributional comparisons

Does contamination matter for the tools that we discussed in sections 4 and 5?

Basic cases. First, take two statistics whose properties can be easily deduced, the mean and median. Using the definition of the mixture distribution (153) with point-contamination (32) and the linearity of the mean functional, we can write the mean of the observed mixture distribution as

$$\mu(G) = \mu\left([1 - \delta]F + \delta H^{(z)}\right) = [1 - \delta]\mu(F) + \delta\mu\left(H^{(z)}\right) \quad (154)$$

Evaluating (154) for the elementary point-contamination distribution (32) we obtain:

$$\mu(G) = [1 - \delta]\mu(F) + \delta z. \quad (155)$$

The observed mean is a simple weighted sum (with weights $1 - \delta, \delta$) of the true mean $\mu(F)$ and the value of z where the contamination is concentrated. Now differentiate (155) with respect to δ and we find the IF for the functional μ as follows:

$$IF(z; \mu, F) = z - \mu(F). \quad (156)$$

It is easy to see from (156) that $IF(z; \mu, F)$ is unbounded as z tends to $-\infty$ or $+\infty$: the mean is a non-robust statistic. So if you want to use the mean as a welfare index then the introduction of a very small amount of contamination sufficiently far out in one of the tails of the distribution will cause the observed value of the mean to be pulled away from the true value.⁸⁶ Now consider the

⁸⁶Using (42) the same type of reasoning can be used to show that Lorenz ordinates are also non-robust (Cowell and Victoria-Feser 2002).

median, as a particular case of the quantile functional (23); using the basic result Lemma 2 and setting $q = 0.5$ to obtain the median we have

$$IF(z; Q(\cdot, 0.5), F) = \frac{q - \iota(q \geq F(z))}{f(Q(F, 0.5))} = \frac{q - \iota(y_{0.5} \geq z)}{f(y_{0.5})}. \quad (157)$$

It is clear that, as long as there is positive density at the median $y_{0.5}$, the IF in (157) is bounded (Cowell and Victoria-Feser 2002). So, in contrast to the mean, the median is robust. The intuition is clear: if you throw a single alien observation into the formula for the mean then, if that observation is large enough, it can have a huge effect when averaged in with the other sample values. But the median simply marks the “half-way” point in the distribution: if you introduce a single alien observation to the right of the median, then the size of that observation (how far it is to the right of the median) has no effect on the observed half-way point.

Inequality. It turns out that most commonly-used inequality indices behave in a way that is similar to the mean: they are non-robust (Cowell and Victoria-Feser 1996b). To see why, let us check the properties of the W_{QAD} class of welfare indices (30) on which many standard inequality measures are based. The influence function for a typical member of this class is

$$\varphi(z, \mu(F)) - W_{\text{QAD}}(F) + [z - \mu(F)] \int \varphi_{\mu}(z, \mu(F)) dF(z) \quad (158)$$

where $\varphi(y, \mu(F))$ is the evaluation of each individual income y used in the formula (30). It is clear that contamination could have an impact through more than one route – there is the direct effect from the evaluation of z , the first term in (158); there is also an indirect route through the effect on the mean, the third term in (158). Notice that this indirect route contains the expression, $z - \mu(F)$, as the right-hand side of (156): from this we can see that, if $\varphi_{\mu}(z, \mu(F))$ is not everywhere zero, contamination will cause quasi-additive welfare indices to be non-robust. Now consider the direct route: clearly if $\varphi(z, \mu(F))$ is unbounded as z approaches infinity or as z approaches zero the particular index in the W_{QAD} class will be non-robust; this is precisely what happens with nearly all commonly used inequality measures.⁸⁷ Why does this happen? Inequality measures are usually designed to be sensitive to extreme values at one or other end of the distribution, so placing a tiny amount of contamination sufficiently far out in one of the tails is going to have a big impact on measured inequality, because of its built-in sensitivity. As an example take the generalised entropy measures. From equations (49)-(51) we see that

$$\varphi(z, \mu(F)) = \frac{[z/\mu(F)]^{\xi} - 1}{\xi^2 - \xi}.$$

⁸⁷The implication of this is that, even with a richer model of contamination than the elementary (32) leaves the mean unchanged, quasi-additively decomposable inequality indices will be non-robust.(Cowell and Victoria-Feser 1996b).

Clearly this is unbounded for $\xi \geq 0$ as $z \rightarrow \infty$ and is unbounded for $\xi \leq 0$ as $z \rightarrow 0$: so the inequality indices are non-robust for contamination amongst very high incomes in the case of top-sensitive members of the GE family and for contamination near zero in the case of bottom-sensitive members of the GE family.⁸⁸

Poverty. By contrast conventional poverty indices such as the FGT class (86) and the Sen index (93) are robust if the poverty line is exogenous or is a function of a robust statistic such as the median (Cowell and Victoria-Feser 1996a). Again the intuition is straightforward. From (79) the influence function for an additively decomposable poverty measure with a fixed poverty line ζ_0 is

$$IF(z; P, F) = p(z, \zeta_0) - P(F)$$

where $p(\cdot)$ is the poverty evaluation function. From (86) we can see that for the FGT class

$$p(z, \zeta_0) = [\max(1 - z/\zeta_0, 0)]^\xi$$

so that $p(0, \zeta_0) = 1$, $p(z, \zeta_0)$ is non-increasing in z for $z < \zeta_0$ and $p(z, \zeta_0) = 0$ for $z \geq \zeta_0$. In plain language contamination at the very bottom of the distribution (below the poverty line) has an impact that it is bounded below; but a very high observation has no effect on poverty, whether that observation is a genuine high income or contamination; poverty measures such as the FGT class are robust under contamination.

6.1.3 Model estimation

If inequality measures are typically non-robust, what is to be done about the possibility of contamination? A potentially useful approach is to use a parametric functional form $f(y; \theta)$ to model all or part of the income distribution and then compute inequality from the modelled distribution. But of course the robustness property of the inequality index based on the modelled distribution will depend on the parameter vector $\theta \in \mathbb{R}^p$ is estimated. If one consider using Maximum Likelihood Estimators (MLE), for example, the robustness problem remains: although MLE are attractive in terms of their efficiency properties, they are usually non-robust. If we consider the wider class of *M-estimators* characterised by⁸⁹

$$\sum_{i=1}^n \psi(y_i; \theta) = 0 \tag{159}$$

where ψ is a function $\mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ one can find estimators with suitable robustness properties: these are the bounded-*IF* *M-estimators* with minimal asymptotic covariance matrix, known as *Optimal Bias-Robust Estimators* (OBRE) –

⁸⁸The logarithmic variance and the Gini coefficient are also non-robust – see (71) above and Cowell and Victoria-Feser (1996b), Cowell and Flachaire (2007).

⁸⁹The MLE belong to the class (159): in this case ψ is equal to the score function.

see Huber (1981), Hampel et al. (1986). One can see OBRE as the solution to a trade-off between efficiency and robustness.

A standard way of specifying the OBRE is as follows. Fix a bound $c \geq \sqrt{p}$ on the IF ; then the OBRE are defined as the solution in θ of

$$\sum_{i=1}^n \psi(x_i; \theta) = \sum_{i=1}^n [s(x_i; \theta) - a(\theta)] \cdot W_c(x_i; \theta) = 0 \quad (160)$$

where $s(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta)$, the score function, and

$$W_c(x; \theta) = \min \left\{ 1; \frac{c}{\|A(\theta)[s(x; \theta) - a(\theta)]\|} \right\} \quad (161)$$

and $W_c(x_i; \theta)$ is a weight imputed to each observation according to its influence on the estimator. The $p \times p$ matrix $A(\theta)$ and $a(\theta) \in \mathbb{R}^p$ are defined by

$$E [\psi(x; \theta)\psi(x; \theta)^T] = [A(\theta)^T A(\theta)]^{-1} \quad (162)$$

$$E [\psi(x; \theta)] = 0 \quad (163)$$

The constant c acts as a regulator between efficiency (high values of c) and robustness (low values of c). The solution of (160) must usually be found iteratively.⁹⁰

6.2 Incomplete data

We now turn to the problems of estimation and inference in a situation where, in part of the sample, some information is unavailable. As we noted in section 2.3, this situation is sometimes imposed by data providers, sometimes created by researchers who are attempting to deal with problems of data contamination as discussed in section 6.1.

6.2.1 Censored and truncated data

Here we are dealing with the cases summarised in the first row of Table 2 in section 2.3 in which we take \underline{z} and \bar{z} as fixed boundaries.

Truncated data. For data represented by case A in Table 2 inference can be approached as for inference in the complete-information case with a redefined population: the limits of the support of the distribution (\underline{y}, \bar{y}) are replaced by the narrower truncation limits (\underline{z}, \bar{z}) . If we wish to say more it may be possible to use a parametric method to estimate the truncated part of the distribution.

⁹⁰See Victoria-Feser and Ronchetti (1994), Cowell and Victoria-Feser (1996b) and for grouped data see Victoria-Feser and Ronchetti (1997).

Censoring with minimal information. Now consider Case B in Table 2. Clearly if we do not use the observed point masses at \underline{z} and \bar{z} , this could be just treated as case A. However, if we want to do more first-order comparisons can be carried out. We need the following statistics: n (the full sample size), \underline{n} (the number of observations equal to \underline{z}) and \bar{n} (the number of observations equal to \bar{z}).

Censoring with rich information. Clearly it is possible to do more in case C than in the previous two cases: more welfare indices (for the whole population) can be handled. Depending on the richness of information in the censored parts it may be possible to carry out inference on Lorenz-curve ordinates and some welfare indices. First, if, in addition to the information described in the discussion of case B, the means of the censored parts of the sample are given,⁹¹ then second order rankings and the Gini coefficient can be estimated. Then it makes sense to define the following:

$$\hat{c}_{\text{low}} := \frac{1}{n} \sum_{i=1}^{\underline{n}} y_{(i)},$$

$$\hat{c}_{\text{high}} := \frac{1}{n} \sum_{i=n-\bar{n}+1}^n y_{(i)}.$$

Inference may also be possible using the same methodology as for the complete data case. To do this we would additionally need the following information

$$\hat{s}_{\text{low}} := \frac{1}{n} \sum_{i=1}^{\underline{n}} y_{(i)}^2$$

$$\hat{s}_{\text{high}} := \frac{1}{n} \sum_{i=n-\bar{n}+1}^n y_{(i)}^2$$

If these variance terms from the excluded portion of the sample are also made available then the asymptotic variances and covariances for the income cumulations (GLC ordinates) for $q, q' \in (\underline{\beta}, \bar{\beta})$ are found as follows. Replace (127) and (129) by the following

$$\hat{c}_q := \hat{c}_{\text{low}} + \frac{1}{n} \sum_{i=\kappa(n,\underline{\beta})+1}^{\kappa(n,q)} y_{(i)} \quad (164)$$

$$\hat{s}_q := \hat{s}_{\text{low}} + \frac{1}{n} \sum_{i=\kappa(n,\underline{\beta})+1}^{\kappa(n,q)} y_{(i)}^2 \quad (165)$$

⁹¹In some cases means will be available from data-providers.

and plug into (138). To compute asymptotic variance for the (relative) Lorenz curve and the Gini coefficient we also need the following: $\hat{\mu} = \hat{c}_{\bar{\beta}} + \hat{c}_{\text{high}}$, $\hat{s}_1 := \hat{s}_{\bar{\beta}} + \hat{s}_{\text{high}}$, $\hat{\omega}_{q1} := \hat{s}_q + [q\hat{y}_q - \hat{c}_q]\hat{\mu} - y_q\hat{c}_q$, $\hat{\omega}_{11} := \hat{s}_1 - \hat{\mu}^2$.

Comparing the outcome from these computations with the full-information case in section 5.2.1 we can draw two important conclusions (Cowell and Victoria-Feser 2003). First, if the necessary information about the censored part of the distribution is used, the standard errors are the same as in the full information case. Second, when the information about the censored part is not available the standard errors are smaller.

6.2.2 Trimmed data

In the case of trimmed data a fixed proportion of the sample is discarded – see the second row of Table 2. The trimmed samples for computing welfare indices and making distributional comparisons is usually based on robustness arguments (Cowell and Victoria-Feser 2006): outliers may seriously bias the point estimates as well as the variances of the distributional statistics that are of interest – see the discussion in section 6.1.

Here we assume that a given proportion $\underline{\beta}$ has been removed from the bottom of the distribution and $1 - \bar{\beta}$ has been removed from the top. If $(y_{\underline{\beta}}, y_{\bar{\beta}})$ denotes the range of the trimmed sample values, then $y_{\underline{\beta}}$ and $y_{\bar{\beta}}$ are *random*. Because of this, and in contrast to the discussion of truncated and censored data (section 6.2.1), case D in Table 2 requires more extensive reworking of the full-information analysis in section 5.2.

Inference is carried out on the full distribution⁹² conditional on the fact that known proportions have been trimmed from the tails. The *trimmed distribution* \tilde{F}_{β} is defined as:

$$\tilde{F}_{\beta}(y) := \begin{cases} 0 & \text{if } y < Q(F, \underline{\beta}) \\ b [F(y) - \underline{\beta}] & \text{if } Q(F, \underline{\beta}) \leq y < Q(F, \bar{\beta}) \\ 1 & \text{if } y \geq Q(F, \bar{\beta}) \end{cases} . \quad (166)$$

where $b := 1/[\bar{\beta} - \underline{\beta}]$. Using (166) the β -trimmed counterparts to (25) and (128) the income cumulations are given by

$$c_{\beta,q} := C(\tilde{F}_{\beta}; q) = b \int_{y_{\underline{\beta}}}^{y_q} y \, dF(y), \quad (167)$$

$$s_{\beta,q} := S(\tilde{F}_{\beta}; q) := b \int_{y_{\underline{\beta}}}^{y_q} y^2 \, dF(y) \quad (168)$$

⁹²Given that the integration of $IF \cdot IF^T$ is required over the full distribution to derive the asymptotic covariance matrix, this might appear to invalidate the applicability of nonparametric techniques because of the lack of information on the structure of the trimmed data. Cowell and Victoria-Feser (2003) show that this supposition is groundless.

and the counterpart of (26) is given by $\mu_\beta := \mu(\tilde{F}_\beta)$. Once again, the sample analogues of (166)-(168) are obtained by replacing F by the empirical distribution $F^{(n)}$. For example $c_{\beta,q}$ is estimated by⁹³

$$\hat{c}_{\beta,q} := C\left(\tilde{F}_\beta^{(n)}; q\right) = \frac{b}{n} \sum_{i=1}^{\kappa(n,q)} y_{(i)} \iota(i > \kappa(n, \underline{\beta}) + 1), \quad (169)$$

where $\{y_{(i)}, i = 1, \dots, n\}$ is the ordered sample, and μ_β is estimated by the mean of the trimmed sample

$$\hat{\mu}_\beta := \mu(\tilde{F}_\beta^{(n)}) = \frac{b}{n} \sum_{i=1}^n y_{(i)} \iota(\kappa(n, \underline{\beta}) + 1 < i < \kappa(n, \bar{\beta})). \quad (170)$$

Lorenz criteria In order to apply second-order dominance criteria we need to know the properties of the income cumulation for the trimmed distribution \tilde{F}_β and its empirical counterpart $\tilde{F}_\beta^{(n)}$. The income cumulations based on the ordinary and trimmed distributions are related as follows:

$$C(\tilde{F}_\beta; q) = b [C(F; q) - C(F; \underline{\beta})], \quad (171)$$

from which it is clear that plotting Lorenz curves, generalised Lorenz curves and so on is straightforward.

The estimation of the asymptotic covariance between $\sqrt{n}\hat{c}_{\beta,q}$ and $\sqrt{n}\hat{c}_{\beta,q'}$ follows as before, from an application of the *IF*. We need to evaluate

$$\int IF(z; C(\cdot; q), \tilde{F}_\beta) IF(z; C(\cdot; q'), \tilde{F}_\beta) dF(z)$$

and then we may compare the results with those in the complete-information case.⁹⁴ Using the definition of the Influence Function then (171) implies that *IF* for the cumulative income functional with trimmed data is:⁹⁵

$$\begin{aligned} IF(z; C(\cdot; q), \tilde{F}_\beta) &= -c_{\beta,q} + b \left[qy_q - \underline{\beta}y_{\underline{\beta}} + \iota(y_q \geq z)[z - y_q] - \iota(y_{\underline{\beta}} \geq z)[z - y_{\underline{\beta}}] \right] \\ &= -c_{\beta,q} + b \left[qy_q - \underline{\beta}y_{\underline{\beta}} - \iota(y_q \geq z)y_q + \iota(y_{\underline{\beta}} \geq z)y_{\underline{\beta}} \right] + b \left[\iota(y_q \geq z) - \iota(y_{\underline{\beta}} \geq z) \right] z. \end{aligned} \quad (172)$$

Taking the mean of $IF(z; C(\cdot; q), \tilde{F}_\beta)IF(z; C(\cdot; q'), \tilde{F}_\beta)$ for each $z = y_i$ it is clear that no value of $z = y_i < y_{\underline{\beta}}$ or $z = y_i > y_{\bar{\beta}}$ will contribute to the value of (172).

Assume that the set of population proportions satisfies $\Theta \subset [\underline{\beta}, \bar{\beta}]$. Then equation (172) yields the following result (Cowell and Victoria-Feser 2003):

⁹³Note that at $q = \bar{\beta}$ for $\underline{\beta} = 1 - \bar{\beta}$ one gets the traditional trimmed mean which generalises the median as a robust estimator of location.

⁹⁴In Lemma 1 $F(z)$ is estimated by $F^{(n)}$ so that the integral reduces to the mean over the sample.

⁹⁵As before $c_{\beta,q}$, $c_{\beta,q'}$, y_q , $y_{q'}$ and $y_{\underline{\beta}}$ can be estimated by their sample counterparts.

Theorem 3 Given an original untrimmed sample of size n and lower and upper trimming proportions $\underline{\beta}, 1 - \bar{\beta} \in \mathbb{Q}$, for any $q, q' \in \Theta$ such that $q \leq q'$ the asymptotic covariance of $\sqrt{n}\hat{c}_{\beta,q}$ and $\sqrt{n}\hat{c}_{\beta,q'}$ is given by

$$\varpi_{qq'} = b^2 \left[\omega_{qq'} + \omega_{\underline{\beta}\underline{\beta}} - \omega_{\underline{\beta}q} - \omega_{q'\underline{\beta}} \right]$$

where $\omega_{qq'}$ is defined in (131).

If we take the set of proportions $\Theta = \{q_i = \underline{\beta} + \frac{i}{n} : i = 1, \dots, n/b\}$, then $\varpi_{qq'}$ can be estimated by

$$\begin{aligned} \hat{\varpi}_{q_i q_j} &= \left[q_i y_{(i)} - \underline{\beta} y_{(1)} - \sum_{k=1}^i \frac{y_{(k)}}{bn_{\underline{\beta}}} \right] \times \\ &\quad \left[[1 - q_j] y_{(j)} - [1 - \underline{\beta}] y_{(1)} + \sum_{k=1}^j \frac{y_{(k)}}{bn_{\underline{\beta}}} \right] - \\ &\quad \sum_{k=1}^i \frac{y_{(i)} y_{(k)} - y_{(k)}^2}{bn_{\underline{\beta}}} + y_{(1)} \left[q_i y_{(i)} - \underline{\beta} y_{(i)} - \sum_{k=1}^i \frac{y_{(i)}}{bn_{\underline{\beta}}} \right] \end{aligned} \quad (173)$$

In the case of the Lorenz curve ordinates the asymptotic covariance of $\sqrt{n}\hat{c}_{\beta,q}/\hat{\mu}_{\beta}$ and $\sqrt{n}\hat{c}_{\beta,q'}/\hat{\mu}_{\beta}$ is given by

$$v_{qq',\beta} = \frac{b^2}{\mu_{\underline{\beta}}^4} \left[\mu_{\underline{\beta}}^2 \varpi_{qq'} + c_{\beta,q} c_{\beta,q'} \varpi_{\underline{\beta}\underline{\beta}} - \mu_{\beta} c_{\beta,q} \varpi_{q'\underline{\beta}} - \mu_{\beta} c_{\beta,q'} \varpi_{q\underline{\beta}} \right]. \quad (174)$$

Compare this with (139).

QAD Welfare indices To evaluate inequality and poverty indices we can again follow the method of section 4, but perform the computations on the trimmed distribution \tilde{F}_{β} defined in (166) – once again ignoring the information on the excluded part of the sample. This means that the trimmed version of (30) becomes

$$W_{\text{QAD}}(\tilde{F}_{\beta}) = b \int \varphi(x, \mu(\tilde{F}_{\beta})) dF(x) \quad (175)$$

The sample analogues of $W_{\text{QAD}}(\tilde{F}_{\beta})$ in (175) are then given by

$$\hat{w}_{\text{QAD},\beta} := W_{\text{QAD}}(\tilde{F}_{\beta}^{(n)}) := \frac{b}{n} \sum_{i=1}^n \varphi(y_{(i)}, \hat{\mu}_{\beta}) \iota(\kappa(n, \underline{\beta}) + 1 < i < \kappa(n, \bar{\beta})) \quad (176)$$

which is the counterpart of (43) but applied to the trimmed sample. Evaluating the IF we have⁹⁶

$$\begin{aligned} IF(z; W_{\text{QAD}}, \tilde{F}_\beta) &= b\varphi\left(\max\left(y_\beta, \min(z, y_{\bar{\beta}})\right), \mu(\tilde{F}_\beta)\right) - W_{\text{QAD}}(\tilde{F}_\beta) \\ &+ bIF(z, C(\cdot; \bar{\beta}), \tilde{F}_\beta) \int_{Q(F, \underline{\beta})}^{Q(F, \bar{\beta})} \varphi_\mu(x, \mu(\tilde{F}_\beta)) dF(x) \end{aligned} \quad (177)$$

Once again, an estimate of the asymptotic variance of $\sqrt{n}W_{\text{QAD}}(\tilde{F}_\beta^{(n)})$ can be easily obtained by computing the mean of squares of $IF(z; W_{\text{QAD}}, \tilde{F}_\beta)$, $z = y_i$, $i = 1, \dots, n$.⁹⁷ Define the following distribution (corresponding to case E in Table 2):

$$F_\beta^*(y) := \begin{cases} 0 & \text{if } y < Q(F, \underline{\beta}) \\ F(y) & \text{if } Q(F, \underline{\beta}) \leq y < Q(F, \bar{\beta}) \\ 1 & \text{if } y \geq Q(F, \bar{\beta}) \end{cases} . \quad (178)$$

We can then state the following (Cowell and Victoria-Feser 2003):

Theorem 4 *The asymptotic variance of $\sqrt{n}W_{\text{QAD}}(\tilde{F}_\beta^{(n)})$ for the trimmed distribution \tilde{F}_β is*

$$\begin{aligned} &b^2 \text{var}(\varphi(x, \mu(\tilde{F}_\beta)); F_\beta^*) \\ &+ 2b^3 \text{cov}(x, \varphi(x, \mu(\tilde{F}_\beta)); F_\beta^*) \int_{Q(F, \underline{\beta})}^{Q(F, \bar{\beta})} \varphi_\mu(x, \mu(\tilde{F}_\beta)) dF(x) \\ &+ b^4 \text{var}(x; F_\beta^*) \left[\int_{Q(F, \underline{\beta})}^{Q(F, \bar{\beta})} \varphi_\mu(x, \mu(\tilde{F}_\beta)) dF(x) \right]^2 \end{aligned} \quad (179)$$

Note that in (179) the variance and covariance terms for the linear functionals are defined on the distribution F_β^* as opposed to the trimmed distribution (166). All the components of (179) can be estimated from the trimmed sample.

⁹⁶To see this evaluate the mixture distribution and apply (34) to get

$$\begin{aligned} &-W_{\text{QAD}}(\tilde{F}_\beta) + b\varphi(z, \mu(\tilde{F}_\beta)) \iota(z \leq y_\beta) \iota(z \geq y_{\bar{\beta}}) - b\varphi(y_{\bar{\beta}}, \mu(\tilde{F}_\beta)) \iota(z \leq y_{\bar{\beta}}) \\ &+ b\varphi(y_\beta, \mu(\tilde{F}_\beta)) \iota(z \leq y_\beta) + bIF(z, C(\cdot; \bar{\beta}), \tilde{F}_\beta) \int_{Q(F, \underline{\beta})}^{Q(F, \bar{\beta})} \varphi_\mu(x, \mu(\tilde{F}_\beta)) dF(x) \\ &+ b\bar{\beta}\varphi(y_{\bar{\beta}}, \mu(\tilde{F}_\beta)) - b\underline{\beta}\varphi(y_\beta, \mu(\tilde{F}_\beta)) \end{aligned}$$

where the first two lines follow by analogy with (45). The third line is found by considering the way the mixture distribution affects the limits of integration in (175) using Lemma 2. Rearranging gives (177).

⁹⁷Notice that the contribution of $z = y_i < y_\beta$ or $z = y_i > y_{\bar{\beta}}$ to (177) is nil.

The Gini coefficient With trimmed data, the Gini coefficient can be expressed as

$$I_{\text{Gini}}(\tilde{F}_\beta) = 1 - 2 \int_{\underline{\beta}}^{\bar{\beta}} \frac{C(\tilde{F}_\beta, q)}{C(\tilde{F}_\beta, \beta)} dq. \quad (180)$$

Using the same procedure as before we first evaluate the IF for the Gini coefficient with trimmed data as:

$$\begin{aligned} IF(z; I_{\text{Gini}}, \tilde{F}_\beta) &= \frac{2}{\mu_\beta} \int_{\underline{\beta}}^{\bar{\beta}} c_{\beta, q} dq - \frac{2b}{\mu_\beta} \left[\int_{\underline{\beta}}^{\bar{\beta}} q y_q dq + \int_{\underline{\beta}}^{\bar{\beta}} \iota(y_q \geq z)[z - y_q] dq \right] \\ &+ \frac{2}{\mu_\beta} \left[\iota(y_{\underline{\beta}} \geq z)[z - y_{\underline{\beta}}] + \underline{\beta} y_{\underline{\beta}} \right] + \frac{2}{\mu_\beta^2} \int_{\underline{\beta}}^{\bar{\beta}} c_{\beta, q} dq (-\mu_\beta + \\ &b \left[\bar{\beta} y_{\bar{\beta}} - \underline{\beta} y_{\underline{\beta}} + \iota(y_{\bar{\beta}} \geq z)[z - y_{\bar{\beta}}] - \iota(y_{\underline{\beta}} \geq z)[z - y_{\underline{\beta}}] \right]). \end{aligned}$$

Using this or the results of Theorem 3, we can obtain⁹⁸

Theorem 5 *The asymptotic variance of $\sqrt{n}I_{\text{Gini}}(\tilde{F}_\beta^{(n)})$ is $4b^2\vartheta_\beta/\mu_\beta^4$ where*

$$\begin{aligned} \vartheta_\beta &= \mu_\beta^2 \int_{\underline{\beta}}^{\bar{\beta}} \int_{\underline{\beta}}^q \varpi_{q'q} dq' dq + \mu_\beta^2 \int_{\underline{\beta}}^{\bar{\beta}} \int_q^{\bar{\beta}} \varpi_{qq'} dq dq + \\ &\varpi_{\bar{\beta}\bar{\beta}} \left[\int_{\underline{\beta}}^{\bar{\beta}} c_{\beta, q} dq \right]^2 - 2\mu_\beta \int_{\underline{\beta}}^{\bar{\beta}} c_{\beta, q} dq \int_{\underline{\beta}}^{\bar{\beta}} \varpi_{q\bar{\beta}} dq \end{aligned} \quad (181)$$

The estimates of ϑ_β are found by making use of (173), with $\hat{\mu}_\beta$ being the trimmed sample mean (170).

6.3 Semi-parametric methods

The problems that we address here may have arisen from situations where the researcher has concerns about data contamination and robustness (see section 6.1) where the data provider has truncated or censored the data (see section 6.2).⁹⁹

The type of problem to be analysed can be simplified if we restrict attention to one leading case. If the support of the income distribution is bounded below then the problems with contaminated data are going to occur only in the upper tail of the distribution (Cowell and Victoria-Feser 2002). It may be reasonable to use a parametric model for the upper tail of the distribution (modelled on a proportion $\beta \in \mathbb{Q}$ of upper incomes) and to use the empirical distribution function directly for the rest of the distribution (the remaining proportion the $1 - \beta$ of lower incomes. There are four main issues

⁹⁸For proof of $IF(z; I_{\text{Gini}}, \tilde{F}_\beta)$ and Theorem 5 see Cowell and Victoria-Feser (2003).

⁹⁹This section draws on Cowell and Victoria-Feser (2007).

- What parametric model should be used for the tail?
- How should the model be estimated?
- How should the proportion β be chosen?
- What are the implications for welfare indices and dominance criteria?

6.3.1 The model

The parametric model most commonly used for the upper tail is the Pareto distribution (2) – see the discussion in section 3.1.1. In principle the Pareto model has two parameters: we suppose here that the parameter y_0 is determined by the $1 - \beta$ quantile $Q(F; 1 - \beta)$ defined in (23) ; the dispersion parameter α is of special of interest and is to be estimated from the data.¹⁰⁰

The *semi-parametric distribution* is then

$$\tilde{F}(y) = \begin{cases} F(y) & y \leq Q(F; 1 - \beta) \\ 1 - \beta \left(\frac{y}{Q(F; 1 - \beta)} \right)^{-\alpha} & y > Q(F; 1 - \beta) \end{cases} . \quad (182)$$

For $y > Q(F; 1 - \beta)$, the density \tilde{f} is

$$\tilde{f}(y; \alpha) = \beta \alpha Q(F; 1 - \beta)^\alpha y^{-\alpha-1} .$$

In particular

$$\tilde{f}(y_{1-\beta}; \alpha) = \frac{\beta \alpha}{y_{1-\beta}} . \quad (183)$$

6.3.2 Model estimation

To estimate the Pareto model for the upper tail of the distribution, one could of course use the MLE but the MLE for the Pareto model is known to be sensitive to data contamination (Victoria-Feser and Ronchetti 1994). Alternatively one could use OBRE as discussed in section 6.1.3, with $p = 1$. Given a sample $\{y_i, i = 1, \dots, n\}$ and a bound $c \geq 1$ on the *IF*, the OBRE are defined implicitly by the solution $\hat{\alpha}(\tilde{F})$ in

$$\int_{Q(F; 1 - \beta)}^{\infty} \psi(y; \hat{\alpha}(\tilde{F}), Q(F; 1 - \beta)) d\tilde{F}(y) = 0.$$

When ψ is the score function $s(y; \alpha, Q(F; 1 - \beta)) = \frac{1}{\alpha} - \log(y) + \log(Q(F; 1 - \beta))$ we get the MLE. We get the OBRE when

$$\psi(y; \alpha) = [s(y; \alpha) - a(\alpha)] W_c(y; \alpha)$$

¹⁰⁰For the results which follow α is assumed to be greater than 2 for the variance to exist.

with

$$W_c(y; \alpha) = \min \left\{ 1; \frac{c}{\|A(\alpha)[s(y; \alpha) - a(\alpha)]\|} \right\} \quad (184)$$

$A(\alpha)$ and vector $a(\alpha)$ are defined implicitly by

$$\begin{aligned} E[\psi(y; \alpha)\psi'(y; \alpha)] &= [A(\alpha)'A(\alpha)]^{-1} \\ E[\psi(y; \alpha)] &= 0. \end{aligned}$$

As explained in section 6.1.3, the constant c parameterises the efficiency-robustness tradeoff. A common method for choosing c is to choose an efficiency level (relative to that of MLE) and derive the corresponding value for c : for the Pareto model, a value of $c = 2$ leads to an OBRE achieving approximately 85% efficiency.

6.3.3 Choice of β

Clearly one could adopt a heuristic approach selecting by eye the amount β of the upper tail to be replaced.

Alternatively one could use the robust approach in Dupuis and Victoria-Feser (2006) who develop a robust prediction error criterion by viewing the Pareto model as a regression model. Rearranging (2) or (4) one can represent the linear relationship between the log of the y and the log of the inverse cumulative distribution function

$$\log\left(\frac{y}{y_0}\right) = -\frac{1}{\alpha} \log(1 - F(y; \alpha)).$$

Given a sample of ordered data $y_{(i)}$, the Pareto regression plot of $\log(y_{(i)})$ versus $-\log\left(\frac{n+1-i}{n+1}\right)$, $i = 1, \dots, n$ can be used to detect graphically the point above which the plot yields a straight line.

6.3.4 Inequality and dominance

The effect on inequality of semi-parametric modelling is easy to see. For example if we wish to see how the Generalised Entropy indices are affected one substitutes \tilde{F} – defined in (182) – into (49)-(51) to obtain $I_{\text{GE}}^{\xi}(\tilde{F})$. For first-order and second-order dominance results we need to look once more at the quantile and cumulative-income functionals.

The quantile functional obtained using (182) is given by

$$Q(\tilde{F}, q) = \begin{cases} Q(F, q) & q \leq 1 - \beta \\ Q(F; 1 - \beta) \left(\frac{1-q}{\beta}\right)^{-1/\hat{\alpha}(\tilde{F})} & q > 1 - \beta \end{cases}. \quad (185)$$

The cumulative-income functional becomes

$$C(\tilde{F}; q) = \begin{cases} \int_{\underline{z}}^{Q(F, q)} y \, dF(y) & q \leq 1 - \beta \\ \int_{\underline{z}}^{Q(F, 1-\beta)} y \, dF(y) \\ + \beta \frac{\hat{\alpha}(\tilde{F})}{1 - \hat{\alpha}(\tilde{F})} Q(F; 1 - \beta) \left[\left(\frac{1-q}{\beta} \right)^{\frac{\hat{\alpha}(\tilde{F})-1}{\hat{\alpha}(\tilde{F})}} - 1 \right] & q > 1 - \beta \end{cases} . \quad (186)$$

The graph of (186) gives the semi-parametric generalised Lorenz curve. The mean of the semi-parametric distribution is given by (186) with $q = 1$, namely

$$\mu(\tilde{F}) = \int_{\underline{z}}^{Q(F, 1-\beta)} y \, dF(y) - \beta Q(F; 1 - \beta) \frac{\hat{\alpha}(\tilde{F})}{1 - \hat{\alpha}(\tilde{F})} = c_{1-\beta} - \beta y_{1-\beta} \frac{\hat{\alpha}}{1 - \hat{\alpha}} . \quad (187)$$

So, using (186) and (187), the semi-parametric Lorenz curve is just the graph of

$$L(\tilde{F}; q) = \frac{C(\tilde{F}; q)}{\mu(\tilde{F})} . \quad (188)$$

Estimates of the GLC and the Lorenz curve for the semi-parametric model can be found by replacing F with $F^{(n)}$ in (182) to obtain

$$\hat{F}_\beta(y) := \begin{cases} F^{(n)}(y) & y \leq Q(F; 1 - \beta) \\ 1 - \beta \left(\frac{y}{Q(F; 1 - \beta)} \right)^{-\alpha} & y > Q(F; 1 - \beta) \end{cases} . \quad (189)$$

and replacing \tilde{F} by \hat{F}_β in (186)-(188).

To illustrate consider the problem of comparing wealth distributions across countries where we are concerned that the upper tail of the distribution may be suffering from some sort of contamination. Table 11 shows the results of estimating a Pareto tail for the net-worth distributions of the UK, Sweden and Canada around the time of the millennium: it employs two different methods of estimating α (OLS and robust) and three different values for the modelled proportion β (top 10%, top 5% and top 1%).¹⁰¹ It is clear that the difference between the two methods of estimation in computing $\hat{\alpha}$ can easily be as large as the differences in $\hat{\alpha}$ between countries: in the case where $\beta = 0.10$ compare the OLS and robust estimates for the UK with the OLS estimate for Sweden.

Panel (a) of Figure 18 shows the two regression methods in detail for the case of the top 10% in the UK using a Pareto plot. It is clear that there are some high net-worth observations which “pull down” the OLS regression line so to speak; if one down-weights these observations, as in the robust regression, one finds a much flatter regression line, corresponding to a lower value of $\hat{\alpha}$

¹⁰¹The data are from the Luxembourg Wealth Study a harmonised data base that facilitates international comparisons – see <http://www.lisdatacenter.org/our-data/lws-database/> . See Cowell (2013) for more detail of this example.

	$\beta = 0.10$	$\beta = 0.05$	$\beta = 0.01$
OLS estimation			
UK 2000	2.55	2.90	3.52
Sweden 2002	1.78	1.76	1.52
Canada 1999	1.37	1.53	1.94
Robust estimation			
UK 2000	1.71	2.08	3.07
Sweden 2002	2.10	2.18	1.61
Canada 1999	1.89	2.15	2.58

Table 11: Estimates of α for net worth using different specifications of β . Source: Luxembourg Wealth Study.

and, consequently, a higher estimate of inequality within the top 10% group. The results of OLS and robust methods used in semi-parametric modelling are further illustrated for this case in Figure 18 (b) which shows the Lorenz curves for the raw data and for the semi-parametric distributions produced by OLS and robust regression. Notice that if one considers the robust method appropriate then the Lorenz curve for the whole distribution will lie well outside the Lorenz curve for the raw data and for the OLS semi-parametric distribution.

Notice that the contrast between the OLS and robust estimates can differ dramatically between countries. This is evident from a comparison of the UK with Sweden or with Canada in Table 11. In the case of Sweden and Canada the outliers pull the regression line in the opposite direction from that seen in the case of the UK: the robustly computed $\hat{\alpha}$ is accordingly higher than that found under OLS. The consequence for the Lorenz curves is shown in Figure 18 (c) and (d): it is clear that for Sweden and Canada the robustly estimated semi-parametric Lorenz curve is close to the Lorenz curve for the raw data; but the OLS-estimated Lorenz curve is quite far away.

The effect of the different estimation methods on inequality *within* the top $100\beta\%$ is obvious – remember that the Gini coefficient for a Pareto distribution with parameter α is just $1/[2\alpha - 1]$. The resulting effect on Gini in the whole distribution is shown in Table 12. Although the effect can be quite large for $\beta = 0.10$ in none of the cases modelled here does it change the conclusion about the ranking by inequality of the three countries.

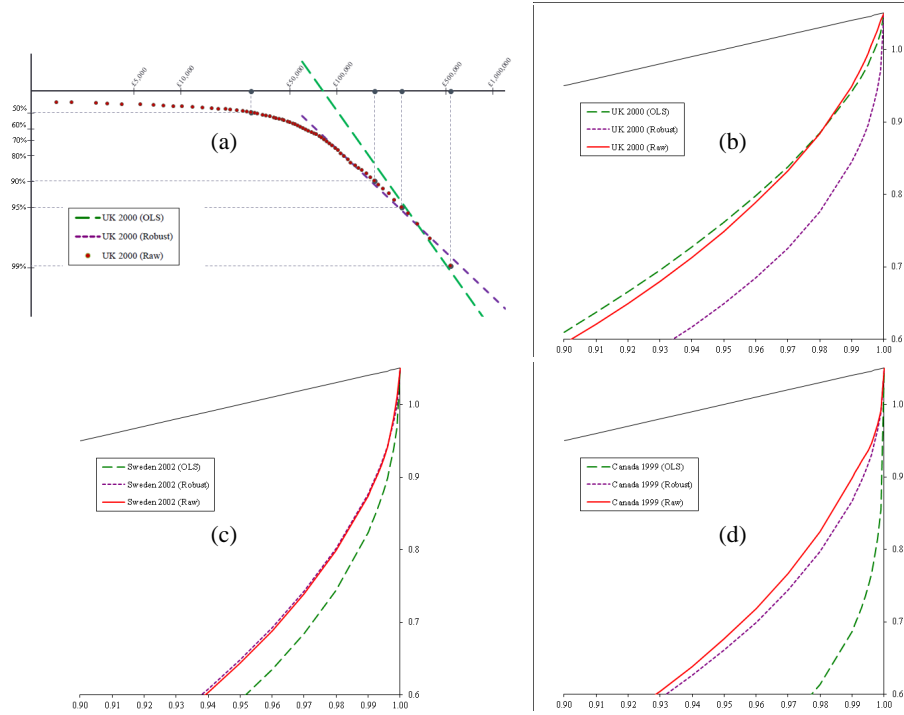


Figure 18: Semi-parametric modelled Lorenz curves of net worth

	\hat{I}_{Gini}	$I_{Gini}(\hat{F}_{0.10})$	$I_{Gini}(\hat{F}_{0.05})$	$I_{Gini}(\hat{F}_{0.01})$
OLS estimation				
UK 2000	0.665	0.657	0.660	0.665
Sweden 2002	0.893	0.901	0.901	0.902
Canada 1999	0.747	0.820	0.788	0.754
Robust estimation				
UK 2000	0.665	0.711	0.683	0.667
Sweden 2002	0.893	0.893	0.892	0.900
Canada 1999	0.747	0.752	0.747	0.745

Table 12: Estimates of the Gini coefficient for net worth from raw data and from semi-parametric distributions. Source: Luxembourg Wealth Study.

7 Conclusions

On reaching the end of a lengthy and technical chapter the authors should confess to an uneasy feeling: a proportion of our potential readership might not have the stamina to work their way through every equation and every footnote. So, we would like to offer time-poor readers three things that may capture the essence of this chapter’s contribution:

- a summary of lessons learned that we hope will be useful for practitioners and for other researchers;
- a little worked example that includes an application of many of the tools that we have discussed;
- a quick-reference table of the main formulas that should be useful to data-providers as well as to the users of data.

7.1 Important lessons: a round-up

Density estimation, parametric (section 3.1):

(1) The Generalised Beta distribution encompass all the standard parametric distribution for income distribution. (2) A “good” goodness-of-fit criterion is important: do use the Anderson-Darling statistic, the Cramér-von-Mises statistic or the Cowell-Davidson-Flachaire (2011) measure; do not use the χ^2 statistic.

Density estimation, semi- and non-parametric (sections 3.2-3.4):

Standard kernel-density methods are very sensitive to the choice of the bandwidth. If the concentration of the data is markedly heterogeneous in the sample then the standard approach (the Silverman rule-of-thumb) is known to often oversmooth in parts of the distribution where the data are dense and under-smooth where the data are sparse, although in other cases it works well. However, this standard approach may not be suitable for income distributions, which are typically heavy-tailed: here the use of the adaptive kernel method or mixture model may be more appropriate.

Welfare measures (section 4):

(1) We propose a global approach to the derivation of variance expressions for all inequality measures. The method uses the Influence Function (see subsection 4.2.1) to provide a shortcut to the formulas we need. (2) It is necessary to analyse the tails (plot of Hill estimators) and use appropriate methods with heavy-tailed distributions (see subsection 4.5.3).

Distributional comparisons (section 5):

(1) As with the welfare measures we propose an approach to the variance and covariance formulas that again makes use of the Influence Function. (2) A plot of Lorenz curve differences can provide useful information, even where Lorenz curves cross.

Data problems (section 6):

(1) Careful modelling is essential to understanding what can be done in the case of possible data-contamination or incomplete data; again the Influence Function is a valuable tool. (2) If one tries to “patch” an empirical distribution with a parametric model for the upper tail then special attention needs to be given to the way the parameters of the model are to be estimated.

7.2 A worked example

To illustrate these lessons, let us consider an empirical analysis of inequality measurement on the income distribution in the United Kingdom in 1992 and 1999.¹⁰²

1. As noted in section 7.1, income distributions are usually very skewed and heavy-tailed: so fixed-bandwidth kernel density estimation, selected by Silverman’s rule-of-thumb, may not be ideal (see section 3.2). Figure 19(a) shows the application of one of the recommended methods adaptive kernel density estimation (where the bandwidth varies with the degree of concentration of the data) of income distributions in 1992 and 1999.¹⁰³ The distribution in 1999 has a smaller mode and is shifted to the right, compared to 1992.
2. Statistical inference on inequality measures may be unreliable, in particular when the underlying distribution is quite heavy-tailed (see subsections 4.5.3 and 4.5.4). A Hill plot is a useful tool for studying the tail behaviour in empirical studies: it represents the Hill estimator of the tail parameter, against the number of k -greatest order statistics used to compute it. An estimate of the tail parameter can be selected when the plot becomes stable about a horizontal straight line. Figure 19(b) shows Hill plots of income distribution in 1992 and 1999, over the range of 0.25% and 25% of order statistics used to compute it, with 95% confidence intervals (in gray). In 1992, the Hill estimate appears to be slightly more than 3,

¹⁰²The data are from the family expenditure survey (FES), a continuous survey of samples of the UK population living in households. We take disposable household income before housing costs, divide household income by an adult-equivalence scale defined by McClements, and exclude the self-employed. The number of observations in 1992 and 1999 are, respectively, equal to 6597 and 5491.

¹⁰³We obtain a very similar Figure with an estimation based on a mixture of lognormal distributions. A kernel density estimation with a fixed bandwidth gives a slightly different picture, the difference being quite similar to that obtained in Figure 9.

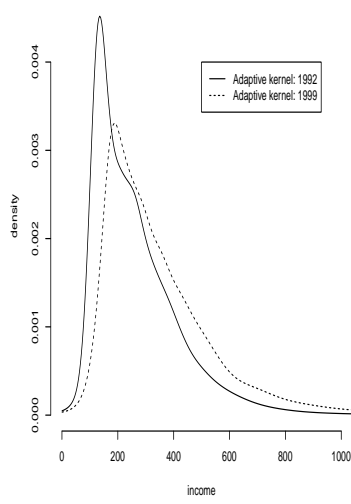
while it is very close to 3 in 1999. It suggests that the distribution in 1999 is slightly more heavy-tailed than those in 1992, both being quite heavy-tailed.¹⁰⁴

3. Strong results on inequality ranking can be drawn from the comparison of relative Lorenz curves, if the curves do not intersect (see section 5). However, in empirical studies intersecting relative Lorenz curves are not unusual and we find that this is the case in our example, with the difference between the two Lorenz curves plotted in Figure 19(c). The Lorenz curve for 1999 is above that for 1992 at the bottom of the distribution; the situation is reversed at the top of the distribution. It suggests that inequality measures more sensitive to transfers in the top (bottom) of the distribution would be larger (smaller) in 1999 than in 1992. However, the 95% confidence intervals shows that, at each point, Lorenz curve differences are not clearly statistically significant, and, thus inequality measures may not be statistically different in 1992 and 1999.
4. Several inequality measures are computed in Figure 19(d): the Gini index and the Generalised Entropy (GE) measures with a sensitivity parameter equals to $-0.5, 0, 1, 2$. GE inequality measures are known to be more sensitive to transfers in the top (bottom) of the distribution as its parameter increases (decreases). Moreover, GE indices with parameters 0, 1 and 2 are, respectively, the mean logarithmic deviation, the Theil and half the square of the coefficient of variation indices. Standard bootstrap confidence intervals are given in brackets. The two distributions are quite heavy-tailed, but the tail parameters are not very different. Reliable inference for testing equality of coefficients can then be obtained with permutation tests (see section 4.5.4): the p -values are given in the last column. The results show that the values of inequality measures that are more sensitive to the top (bottom) of the distribution are larger (smaller) in 1999 than in 1992. However, taking into account statistical inference leads us not to reject the hypothesis that the inequality measures are similar in 1992 and in 1999. These results are consistent with the previous analysis drawn from the Lorenz curves comparison.

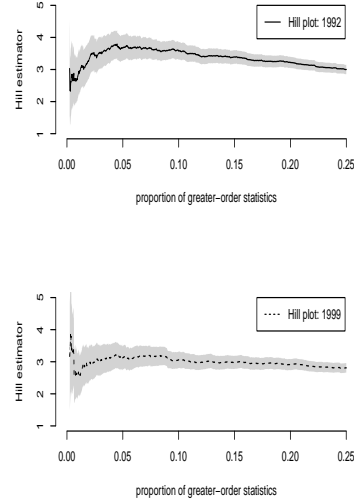
7.3 A cribsheet

Finally we offer something for those who are really short of time or patience. In this chapter we have proposed a unified approach for computing variances and covariances for many inequality and poverty measures, as well as Lorenz curve ordinates. This unified approach involves some quite simple – or at least not very complicated – formulas. Table 13 provides a one-page summary of the key formulas for the principal statistical tasks in distributional analysis.

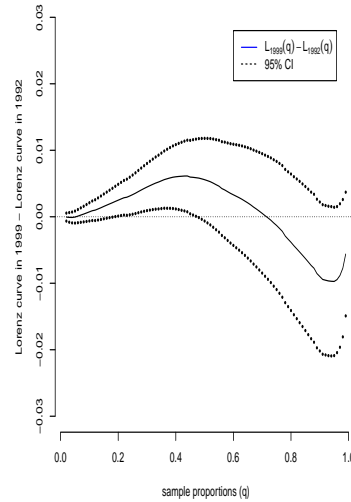
¹⁰⁴Note that the variance of a Pareto distribution exists if the Pareto index is greater than 2.



(a) Density estimation



(b) Upper-tail analysis



(c) Lorenz curve comparison

Index ^a	1992	1999	<i>p</i> -value ^b
I_{Gini}	0.3214	0.3200	0.7988
$I_{\text{GE}}^{-0.5}$	0.2279	0.2175	0.8288
I_{GE}^0	0.1742	0.1740	0.9529
I_{GE}^1	0.1794	0.1853	0.6006
I_{GE}^2	0.2415	0.2651	0.4244

^a I_{GE}^0 , I_{GE}^1 and I_{GE}^2 are, respectively, the MLD, the Theil index and half the square of the coefficient of variation.

^bPermutation *p*-value for testing the equality of coefficients.

(d) Inequality measures

Figure 19: Inequality analysis on household income in 1992 and 1999 in United Kingdom: (a) Adaptive kernel density estimation; (b) Hill estimator of the tail index (Hill plots); (c) Difference of Lorenz curves; and, (d) Inequality measures, with bootstrap confidence intervals and permutation *p*-values for testing equality.

Coefficient	Variance: $\widehat{\text{var}}(\text{Coef}) = \frac{1}{n} \text{var}(Z) = \frac{1}{n^2} \sum_{i=1}^n (Z_i - \bar{Z})^2$, where Z_i is equal to
<i>Inequality measures</i>	
$\hat{I}_{\text{GE}}^\xi = \frac{1}{n(\xi^2 - \xi)} \sum_{i=1}^n \left[\left(\frac{y_i}{\hat{\mu}} \right)^\xi - 1 \right]$	$Z_i = \frac{1}{\xi^2 - \xi} \left(\frac{y_i}{\hat{\mu}} \right)^\xi - \xi \frac{y_i}{\hat{\mu}} \left[\hat{I}_{\text{GE}}^\xi + \frac{1}{\xi^2 - \xi} \right]$
$\hat{I}_{\text{GE}}^0 = -\frac{1}{n} \sum_{i=1}^n \log\left(\frac{y_i}{\hat{\mu}}\right)$	$Z_i = \frac{y_i}{\hat{\mu}} - \log y_i$
$\hat{I}_{\text{GE}}^1 = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i}{\hat{\mu}} \right) \log\left(\frac{y_i}{\hat{\mu}}\right)$	$Z_i = \frac{y_i}{\hat{\mu}} \left[\log\left(\frac{y_i}{\hat{\mu}}\right) - \hat{I}_{\text{GE}}^1 - 1 \right]$
$\hat{I}_{\text{MD}} = \frac{1}{n} \sum_{i=1}^n y_i - \hat{\mu} $	$Z_i = 2(\hat{q} - 1)y_i + 2(y_i - \hat{\mu}) \iota(y_i \geq \hat{\mu})$
$\hat{I}_{\text{Gini}} = \frac{2}{\hat{\mu}n(n-1)} \sum_{i=1}^n i y_{(i)} - \frac{n+1}{n-1}$	$Z_i = \frac{1}{\hat{\mu}} \left[-(\hat{I}_{\text{Gini}} + 1)y_{(i)} + \frac{2i-1}{n} y_{(i)} - \frac{2}{n} \sum_{j=1}^i y_{(j)} \right]$
<i>Poverty measures</i>	
$\hat{P}_{\text{FGT}}^\xi = \frac{1}{n} \sum_{i=1}^n Z_i = \bar{Z}$	$Z_i = \left 1 - \frac{y}{\zeta_0} \right ^\xi \iota(y \leq \zeta_0)$
$\hat{P}_{\text{Sen}} = \frac{2}{nn_p \zeta_0} \sum_{i=1}^{n_p} (\zeta_0 - y_{(i)}) (n_p - i + \frac{1}{2})$	$Z_i = \frac{2n}{\zeta_0 n_p} \left[\frac{\zeta_0}{2} \left(\frac{2n_p}{n} - \hat{P}_{\text{Sen}} \right) - \frac{2n_p - 2i + 1}{2n} y_{(i)} - \frac{1}{n} \sum_{j=1}^i y_{(j)} \right] \iota(y_{(i)} \leq \zeta_0)$
$\hat{P}_{\text{SST}} = \frac{2}{\zeta_0 n(n-1)} \sum_{i=1}^{n_p} (\zeta_0 - y_{(i)}) (n - i)$	$Z_i = \frac{2n}{\zeta_0(n-1)} \left[\zeta_0 \left(1 - \frac{n_p}{n} \right) - \frac{2n - 2i + 1}{2n} y_{(i)} + \frac{1}{n} \sum_{j=1}^{n_p} y_{(j)} - \frac{1}{n} \sum_{j=1}^i y_{(j)} \right] \iota(y_{(i)} \leq \zeta_0)$
<i>Lorenz curves</i>	
$\hat{c}_q = \frac{1}{n} \sum_{i=1}^{\kappa(n,q)} y_{(i)}$	$Z_{iq} = [y_i - \hat{y}_q] \iota(y_i \leq \hat{y}_q)$
$\hat{c}_q / \hat{\mu} = \frac{1}{\hat{\mu}n} \sum_{i=1}^{\kappa(n,q)} y_{(i)}$	$Z_{iq} = \frac{1}{\hat{\mu}^2} \left[\hat{\mu}(y_i - \hat{y}_q) \iota(y_i \leq \hat{y}_q) - \hat{c}_q y_i \right]$

$Z = \{Z_1, \dots, Z_n\}$ and $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$; $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$ is the sample mean; $y_{(i)}$ is the i th order statistic of the sample; $\hat{q} = \frac{1}{n} \sum_{i=1}^n \iota(y_i \leq \hat{\mu})$; ζ_0 is the poverty line; $n_p = \sum_{i=1}^n \iota(y_i \leq \zeta_0)$ is the number of poor; q is a sample proportion; $\kappa(n, q) = \lfloor nq - q + 1 \rfloor$ is the largest integer no greater than $nq - q + 1$; and $\hat{y}_q = y_{(\kappa(n,q))}$ is a sample quantile. \hat{I}_{GE}^0 and \hat{I}_{GE}^1 are, respectively, the Mean Logarithmic Deviation and the Theil inequality indices.

Table 13: Formulas for computing coefficient estimates and variances for inequality measures, poverty measures and (general or relative) Lorenz curve ordinates.

References

- Ahamada, I. and E. Flachaire (2011). *Non-Parametric Econometrics*. Oxford University Press.
- Aitchison, J. and J. A. C. Brown (1957). *The Lognormal Distribution*. London: Cambridge University Press.
- Alvaredo, F. and E. Saez (2009). Income and wealth concentration in Spain from a historical and fiscal perspective. *Journal of the European Economic Association* 7(5), 1140–1167.
- Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica* 64, 1183–1193.
- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory* 2, 244–263.
- Atkinson, A. B. (1987). On the measurement of poverty. *Econometrica* 55, 749–764.
- Autor, D. H., L. F. Katz, and M. Kearney (2008). Trends in U.S. wage inequality: revising the revisionists. *Review of Economics and Statistics* 90(2), 300–323.
- Bandourian, R., J. B. McDonald, and R. S. Turley (2003). A comparison of parametric models of income distribution across countries and over time. *Estadística* 55, 135–152.
- Bantilan, M. C. S., N. B. Bernal, M. M. de Castro, and J. M. Pattugalan (1995). Income distribution in the Philippines, 1957-1988: An application of the Dagum model to the family income and expenditure survey (FIES) data. In C. Dagum and A. Lemmi (Eds.), *Research on Economic Inequality, Volume 6: Income Distribution, Social Welfare, Inequality and Poverty*, pp. 11–43. Greenwich: JAI Press.
- Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica* 71, 71–104.
- Barrett, G. F. and S. G. Donald (2009). Statistical inference with generalized Gini indices of inequality, poverty, and welfare. *Journal of Business and Economic Statistics* 27, 1–17.
- Beach, C. M. and R. Davidson (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies* 50, 723–735.
- Beach, C. M. and S. F. Kaliski (1986). Lorenz Curve inference with sample weights: an application to the distribution of unemployment experience. *Applied Statistics* 35(1), 38–45.
- Beach, C. M. and J. Richmond (1985). Joint confidence intervals for income shares and Lorenz curves. *International Economic Review* 26(6), 439–450.
- Beran, R. (1988). Pivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* 83, 687–697.

- Berger, Y. G. and C. J. Skinner (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society: series C (applied statistics)* 52(4), 457–468.
- Berndt, E. R. (1990). *The Practice of Econometrics: Classic and Contemporary*. Addison Wesley.
- Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics* 137, 674–707.
- Biewen, M. and S. P. Jenkins (2006). Variance estimation for generalized entropy and atkinson inequality indices: the complex survey data case. *Oxford Bulletin of Economics and Statistics* 68, 371–383.
- Binder, D. A. and M. S. Kovacevic (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology* 21,2, 137–145.
- Bishop, J. A., S. Chakraborti, and P. D. Thistle (1988). Large sample tests for absolute Lorenz dominance. *Economics Letters* 26, 291–294.
- Bishop, J. A., J. B. Formby, and B. Zheng (1997). Statistical inference and the Sen index of poverty. *International Economic Review* 38, 381–387.
- Bishop, J. A., J. P. Formby, and W. J. Smith (1991a). International comparisons of income inequality: Tests for Lorenz dominance across nine countries. *Economica* 58, 461–477.
- Bishop, J. A., J. P. Formby, and W. J. Smith (1991b). Lorenz dominance and welfare: Changes in the u.s. distribution of income, 1967-1986. *The Review of Economics and Statistics* 73, 134–139.
- Bishop, J. A., J. P. Formby, and P. Thistle (1992). Convergence of the south and non-south income distributions, 1969-1979. *American Economic Review* 82, 262–272.
- Bishop, J. A., J. P. Formby, and P. D. Thistle (1989). Statistical inference, income distributions and social welfare. In D. J. Slottje (Ed.), *Research on Economic Inequality I*. JAI Press.
- Boos, D. D. and R. J. Serfling (1980). A note on differential and the clt and lil for statistical functions, with application to m -estimates. *Annals of Statistics* 8, 618–624.
- Bordley, R. F., J. B. McDonald, and A. Mantrala (1997). Something new, something old: Parametric models for the size distribution of income. *Journal of Income Distribution* 6, 91–103.
- Bourguignon, F. and C. Morrisson (2002). Inequality among world citizens: 1820-1992. *American Economic Review* 92, 727–744.
- Bowman, A. (1984). An alternative method of cross-validation for the smoothing of kernel density estimates. *Biometrika* 71, 353–360.
- Brachmann, K., A. Stich, and M. Trede (1996). Evaluating parametric income distribution models. *Allgemeines Statistisches Archiv* 80, 285–298.

- Burkhauser, R. V., A. D. Crews Cutts, M. C. Daly, and S. P. Jenkins (1999). Testing the significance of income distribution changes over the 1980s businesscycle: a cross-national comparison. *Journal of Applied Econometrics* 14, 253–272.
- Burkhauser, R. V., S. Feng, and S. P. Jenkins (2009). Using the p90/p10 index to measure U.S. inequality trends with Current Population Survey data: a view from inside the Census Bureau vaults. *Review of Income and Wealth* 55(1), 166–185.
- Burkhauser, R. V., S. Feng, and J. Larrimore (2010). Improving imputations of top incomes in the public-use current population survey by using both cell-means and variances. *Economics Letters* 108(1), 69–72.
- Butler, R. J. and J. B. McDonald (1989). Using incomplete moments to measure inequality. *Journal of Econometrics* 42, 109–19.
- Chen, W. and J. Duclos (2008). Testing for poverty dominance: An application to Canada. Discussion Paper 3829, IZA.
- Chernozhukov, V., I. Fernandez-Val, and B. Melly (2009). Inference on counterfactual distributions. Working Paper.
- Chesher, A. and C. Schluter (2002). Measurement error and inequality measurement. *Review of Economic Studies* 69, 357–378.
- Chotikapanich, D. and W. Griffiths (2008). Estimating income distributions using a mixture of gamma densities. In D. Chotikapanich (Ed.), *Modelling Income Distributions and Lorenz Curves*, Chapter 16, pp. 285–302. New-York: Springer.
- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28, 591–605.
- Chung, E. and J. P. Romano (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics* 41, 484–507.
- Cowell, F. A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics* 42, 27–41.
- Cowell, F. A. (1991). Grouping bounds for inequality measures under alternative informational assumptions. *Journal of Econometrics* 48, 1–14.
- Cowell, F. A. (1999). Estimation of inequality indices. In J. Silber (Ed.), *Handbook on Income Inequality Measurement*. Denter: Kluwer.
- Cowell, F. A. (2000). Measurement of inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Chapter 2, pp. 87–166. New York: Elsevier Science B. V.
- Cowell, F. A. (2011). *Measuring Inequality* (Third ed.). Oxford: Oxford University Press.
- Cowell, F. A. (2013). UK wealth inequality in international context. In J. R. Hills (Ed.), *Wealth in the UK*, Chapter 3. Oxford: Oxford University Press.

- Cowell, F. A., R. Davidson, and E. Flachaire (2011). Goodness of fit: an axiomatic approach. Working Paper 2011/50, Greqam.
- Cowell, F. A. and C. Fiorio (2011). Inequality decompositions – a reconciliation. *Journal of Economic Inequality* 9(99), 509–528.
- Cowell, F. A. and E. Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics* 141, 1044–1072.
- Cowell, F. A., E. Flachaire, and S. Bandyopadhyay (2013). Reference distributions and inequality measurement. *Journal of Economic Inequality* 11, 421–437.
- Cowell, F. A. and S. P. Jenkins (2003). Estimating welfare indices: household sample design. *Research on Economic Inequality* 9, 147–172.
- Cowell, F. A., S. P. Jenkins, and J. A. Litchfield (1996). The changing shape of the UK income distribution: Kernel density estimates. In J. R. Hills (Ed.), *New Inequalities: the Changing Distribution of Income and Wealth in the United Kingdom*, Chapter 3, pp. 49–75. Cambridge: Cambridge University Press.
- Cowell, F. A. and M.-P. Victoria-Feser (1996a). Poverty measurement with contaminated data: A robust approach. *European Economic Review* 40, 1761–1771.
- Cowell, F. A. and M.-P. Victoria-Feser (1996b). Robustness properties of inequality measures. *Econometrica* 64, 77–101.
- Cowell, F. A. and M.-P. Victoria-Feser (2002). Welfare rankings in the presence of contaminated data. *Econometrica* 70, 1221–1233.
- Cowell, F. A. and M.-P. Victoria-Feser (2003). Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* 1, 191–219.
- Cowell, F. A. and M.-P. Victoria-Feser (2006). Distributional dominance with trimmed data. *Journal of Business and Economics Statistics* 24, 291–300.
- Cowell, F. A. and M.-P. Victoria-Feser (2007). Robust stochastic dominance: A semi-parametric approach. *Journal of Economic Inequality* 5, 21–37.
- Dagsvik, J., Z. Jia, B. H. Vatne, and W. Zhu (2013). Is the Pareto-Lévy law a good representation of income distributions? *Empirical Economics* 44, 719–737.
- Dagum, C. (1977). A new model of personal income distribution: Specification and estimation. *Economie Appliquée* 30, 413–436.
- Dagum, C. (1980). The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée* 33, 327–367.
- Dagum, C. (1983). Income distribution models. In D. L. Banks, C. B. Read, and S. Kotz (Eds.), *Encyclopedia of Statistical Sciences*, Volume 4, pp. 27–34.

- Dalton, H. (1920). Measurement of the inequality of incomes. *The Economic Journal* 30, 348–361.
- Daly, M. C., A. D. Crews, and R. V. Burkhauser (1997). A new look at the distributional effects of economic growth during the 1980s: A comparative study of the United States and Germany. *Economic Review* 2, 18–31.
- Dardanoni, V. and A. Forcina (1999). Inference for Lorenz curve orderings. *The Econometrics Journal* 2, 49–75.
- Davidson, R. (2008). Stochastic dominance. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics (2nd edition)*. Palgrave Macmillan.
- Davidson, R. (2009a). Reliable inference for the Gini index. *Journal of Econometrics* 150(1), 30–40.
- Davidson, R. (2009b). Testing for restricted stochastic dominance: Some further results. *Review of Economic Analysis* 1, 34–59.
- Davidson, R. (2010). Innis lecture: Inference on income distributions. *Canadian Journal of Economics* 43, 1122–1148.
- Davidson, R. (2012). Statistical inference in the presence of heavy tails. *The Econometrics Journal* 15, C31–C53.
- Davidson, R. and J.-Y. Duclos (1997). Statistical inference for the measurement of the incidence of taxes and transfers. *Econometrica* 65, 1453–1466.
- Davidson, R. and J.-Y. Duclos (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* 68, 1435–1464.
- Davidson, R. and J.-Y. Duclos (2013). Testing for restricted stochastic dominance. *Econometric Reviews* 32, 84–125.
- Davidson, R. and E. Flachaire (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* 141, 141 – 166.
- Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews* 19, 55–68.
- Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods*. Cambridge: Cambridge University Press.
- Deaton, A. S. (1997). *The Analysis of Household Surveys*. Baltimore, Maryland: Johns Hopkins Press for the World Bank.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1–38.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Statistics Canada: Survey Methodology* 25, 193–203.

- DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica* 64, 1001–1044.
- Donald, S. G. and G. F. Barrett (2004). Consistent nonparametric tests for Lorenz dominance. Econometric Society 2004 Australasian Meetings 321, Econometric Society.
- Donald, S. G., D. A. Green, and H. J. Paarsch (2000). Differences in wage distributions between Canada and the United States: An application of a flexible estimator of distribution functions in the presence of covariates. *Review of Economic Studies* 67, 609–633.
- Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012). Incorporating covariates in the measurement of welfare and inequality: methods and applications. *Econometrics Journal* 15, C1–C30.
- Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters : A general approach to finite-sample inference and nonstandard asymptotics in econometrics. *Journal of Econometrics* 133, 443–477.
- Dufour, J.-M., E. Flachaire, and L. Khalaf (2013). Permutation tests for comparing inequality measures with heavy-tailed distributions. Technical report. Paper presented at the 53th conference of the Société Canadienne de Science Économique in Québec.
- Dupuis, D. J. and M.-P. Victoria-Feser (2006). A robust prediction error criterion for Pareto modeling of upper tails. *Canadian Journal of Statistics*. To appear.
- Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28, 181–187.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Volume 38. Philadelphia: Society for Industrial and Applied Mathematics.
- Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Its Applications* 14, 153–158.
- Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90, 577–588.
- Fan, J. and T. H. Yim (2004). A cross-validation method for estimating conditional densities. *Biometrika* 91, 819–834.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, 287–302.
- Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics 19. New York: Springer.
- Fishburn, P. C. (1980). Stochastic dominance and moments of distributions. *Mathematics of Operations Research* 5, 94–100.

- Fishburn, P. C. and I. H. Lavalley (1995). Stochastic dominance on unidimensional grids. *Mathematics of Operations Research* 20, 513–525.
- Fisher, R. (1935). *The Design of Experiments*. London: Oliver & Boyd.
- Flachaire, E. and O. Nuñez (2007). Estimation of income distribution and detection of subpopulations: an explanatory model. *Computational Statistics & Data Analysis* 51, 3368–3380.
- Foster, J. E., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. *Econometrica* 52, 761–776.
- Foster, J. E. and A. F. Shorrocks (1988). Poverty orderings. *Econometrica* 56(1), 173–177.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.
- Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica* 39, 1037–1039.
- Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* 54, 306–316.
- Gastwirth, J. L. (1974). Large-sample theory of some measures of inequality. *Econometrica* 42, 191–196.
- Gastwirth, J. L. (1975). The estimation of a family of measures of economic inequality. *Journal of Econometrics* 3, 61–70.
- Gastwirth, J. L., T. K. Nayak, and A. N. Krieger (1986). Large sample theory for the bounds on the Gini and related indices from grouped data. *Journal of Business and Economic Statistics* 4, 269–273.
- Ghosal, S. and A. W. van der Vaart (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics* 29(5), 1233–1263.
- Gibrat, R. (1931). *Les Inégalités Economiques*. Paris: Sirey.
- Giles, D. E. A. (2004). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics* 66, 425–433.
- Gleser, L. J. (1973). On a theory of intersection-union tests. *Institute of Mathematical Statistics Bulletin* 2, 2330.
- Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics.
- Guerrero, V. M. (1987). A note on the estimation of Atkinson’s index of inequality. *Economics Letters* 25, 379–384.
- Hadar, J. and W. R. Russell (1969). Rules for ordering uncertain prospects. *American Economic Review* 79, 25–34.
- Hajargasht, P., W. E. Griffiths, J. Brice, D. S. P. Rao, and D. Chotikapanich (2012). Inference for income distributions using grouped data. *Journal of Business and Economic Statistics* 30, 563–575.

- Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99, 1015–1026.
- Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematics and Statistics* 42, 1887–1896.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383–393.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.
- Hasegawa, H. and H. Kozumi (2003). Estimation of Lorenz curves: a Bayesian nonparametric approach. *Journal of Econometrics* 115, 277–291.
- Howes, S. R. and J. O. Lanjouw (1998). Poverty comparisons and household survey design. *Review of Income and Wealth* 44, 99–108.
- Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.
- Jenkins, S. P. (1995). Did the middle class shrink during the 1980s? UK evidence from kernel density estimates. *Economics Letters* 49, 407–413.
- Jenkins, S. P. (2009). Distributionally-sensitive inequality indices and the GB2 income distribution. *Review of Income and Wealth* 55, 392–298.
- Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore (2011, March). Measuring inequality using censored data: A multiple imputation approach. *Journal of the Royal Statistical Society, Series A* 174(866), 63–81.
- Jenkins, S. P. and P. J. Lambert (1997). Three ‘I’s of poverty curves, with an analysis of UK poverty trends. *Oxford Economic Papers* 49, 317–327.
- Jenkins, S. P. and P. Van Kerm (2005). Accounting for income distribution trends: A density function decomposition approach. *Journal of Economic Inequality* 3, 43–61.
- Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions* (Second ed.), Volume 1 of *John Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley.
- Kakwani, N. (1993). Statistical inference in the measurement of poverty. *Review of Economics and Statistics* 75, 632–639.
- Kaur, A., B. L. S. Prakasa Rao, and H. Singh (1994). Testing for second order stochastic dominance of two distributions. *Econometric Theory* 10, 849–866.
- Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken, N.J.: John Wiley.
- Kovacevic, M. S. and D. A. Binder (1997). Variance estimation for measures of income inequality and polarization. *Journal of Official Statistics* 13, 41–58.

- Langel, M. and Y. Tillé (2013). Variance estimation of the Gini index: revisiting a result several times published. *Journal of the Royal Statistical Society A* 176, 521–540.
- Lemieux, T. (2006). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review* 96, 461–498.
- Li, J. and J. S. Racine (2006). *Nonparametric Econometrics*. Princeton University Press.
- Linton, O., E. Maasoumi, and Y.-J. Whang (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies* 72, 735–765.
- Lubrano, M. and A. A. J. Ndoye (2011). Inequality decomposition using the Gibbs output of a mixture of lognormal distributions. Working paper, Greqam 2011-19.
- Maasoumi, E. and A. Heshmati (2008). Evaluating dominance ranking of PSID incomes by various household attributes. In G. Betti and A. Lemmi (Eds.), *Advances on Income Inequality and Concentration Measures*, Chapter 4, pp. 47–69. Routledge Frontiers of Political Economy.
- Majumder, A. and S. R. Chakravarty (1990). Distribution of personal income: Development of a new model and its application to U.S. income data. *Journal of Applied Econometrics* 5, 189–196.
- Mandelbrot, B. (1960). The Pareto-Lévy law and the distribution of income. *International Economic Review* 1(2), 79–106.
- Marron, J. S. and H. P. Schmitz (1992). Simultaneous density estimation of several income distributions. *Econometric Theory* 8, 476–488.
- Marron, J. S. and M. P. Wand (1992). Exact mean integrated squared error. *Annals of Statistics* 20, 712–736.
- McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica* 52, 647–664.
- McDonald, J. B. and Y. J. Xu (1995). A generalization of the beta distribution with applications. *Journal of Econometrics* 66, 133–152.
- McFadden, D. (1989). Testing for stochastic dominance. In T. B. Fomby and T. K. Seo (Eds.), *Studies in the Economics of Uncertainty*, pp. 113–134. New-York: Springer-Verlag.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.
- Mills, J. A. and S. Zandvakili (1997). Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* 12, 133–150.
- Modarres, R. and J. L. Gastwirth (2006). A cautionary note on estimating the standard error of the Gini index of inequality. *Oxford Bulletin of Economics And Statistics* 68, 385–390.

- Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. *Statistica* 51, 561–577.
- Nygård, F. and A. Sandström (1985). Estimating Gini and entropy inequality parameters. *Journal of Official Statistics* 1, 399–412.
- O’Brien, G. L. (1984). Stochastic dominance and moment inequalities. *Mathematics of Operations Research* 9, 475–477.
- O’Brien, G. L. and M. Scarsini (1991). Multivariate stochastic dominance and moments. *Mathematics of Operations Research* 16, 382–389.
- Ogwang, T. (2000). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics* 62, 123–129.
- Osberg, L. and K. Xu (2000). International comparisons of poverty intensity: index decomposition and bootstrap inference. *Journal of Human Resources* 35, 51–81. Errata. *Journal of Human Resources*, 35(3), inside front cover page, 2000.
- Paap, R. and H. K. Van Dijk (1998). Distribution and mobility of wealth of nations. *European Economic Review* 42, 1269–1293.
- Pareto, V. (1895). La legge della domanda. *Giornale degli Economisti* (10), 59–68.
- Paul, S. (1999). The population sub-group income effects on inequality: Analytical framework and an empirical illustration. *Economic Record* 75(229), 149–155. 7.
- Pen, J. (1974). *Income Distribution* (second ed.), Chapter 3, pp. 48–59. London: Allen Lane, The Penguin Press.
- Pittau, M. G. and R. Zelli (2004). Testing for changes in the shape of income distribution: Italian evidence in the 1990s from kernel density estimates. *Empirical Economics* 29, 415–430.
- Pittau, M. G. and R. Zelli (2006). Empirical evidence of income dynamics across EU regions. *Journal of Applied Econometrics* 21, 605–628.
- Pittau, M. G., R. Zelli, and P. A. Johnson (2010). Mixture models, convergence clubs and polarization. *Review of Income and Wealth* 56, 102–122.
- Polivka, A. E. (1998). Using earnings data for the Current Population Survey after the redesign. Working Paper 306, US Bureau of Labour Statistics, Washington, DC.
- Quah, D. T. (1997). Empirics for growth and distribution: Stratification, polarization, and convergence clubs. *Journal of Economic Growth* 2, 27–59.
- Quirk, J. D. and R. Saposnik (1962). Admissibility and measurable utility functions. *Review of Economic Studies* 29, 140–146.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (Second ed.). New York: John Wiley.

- Ray, S. and B. G. Lindsay (2005). The topography of multivariate normal mixtures. *Annals of Statistics* 33, 2042–2065.
- Reeds, J. A. (1976). On the definition of von Mises functionals. Research Report S 44, Department of Statistics, Harvard University, Cambridge, Mass.
- Resnick, S. I. (1997). Heavy tail modeling and teletraffic data. *Annals of Statistics* 25, 1805–1849.
- Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering.
- Robert, C. P. and G. Casella (2005). *Monte Carlo Statistical Methods*. Springer.
- Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics* 155, 56–70.
- Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* 24, 220–238.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* 9, 65–78.
- Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90, 1257–69.
- Sala-i-Martin, X. (2006). The world distribution of income: Falling poverty and ... convergence, period. *Quarterly Journal of Economics* 121, 351–397.
- Salem, A. B. Z. and T. D. Mount (1974). A convenient descriptive model of income distribution: The Gamma density. *Econometrica* 42, 1115–1127.
- Sandström, A., J. H. Wretman, and B. Walden (1988). Variance estimators of the Gini coefficient: probability sampling. *Journal of Business and Economic Statistics* 6, 113–120.
- Sarabia, J. M. (2008). Parametric Lorenz curves: Models and applications. In D. Chotikapanich (Ed.), *Modeling Income Distributions and Lorenz Curves*, Chapter 9, pp. 167–190. Springer.
- Schechtman, A. (1991). On estimating the asymptotic variance of a function of U statistics. *The American Statistician* 45, 103–106.
- Schluter, C. (2012). On the problem of inference for inequality measures for heavy-tailed distributions. *The Econometrics Journal* 15, 125–153.
- Schluter, C. and M. Trede (2002). Tails of Lorenz curves. *Journal of Econometrics* 109, 151–166.
- Schluter, C. and K. van Garderen (2009). Edgeworth expansions and normalizing transforms for inequality measures. *Journal of Econometrics* 150, 16–29.

- Schmid, F. and M. Trede (1996). Testing for first-order stochastic dominance: A new distribution-free test. *Statistician* 45, 371–380.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New-York: John Wiley & Sons.
- Sen, A. K. (1976). Poverty: An ordinal approach to measurement. *Econometrica* 44, 219–231.
- Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society* 53, 683–690.
- Shorrocks, A. F. (1983). Ranking income distributions. *Economica* 50, 3–17.
- Shorrocks, A. F. (1995). Revisiting the Sen poverty index. *Econometrica* 63, 1225–1230.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Singh, S. K. and G. S. Maddala (1976). A function for the size distribution of income. *Econometrica* 44, 963–970.
- Stacy, E. W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics* 33, 1187–1192.
- Stark, O. and S. Yitzhaki (1988). Merging populations, stochastic dominance and lorenz curves. *Journal of Population Economics* 1, 157–161.
- Stephens, M. A. (1986). Tests based on EDF statistics. In R. B. d’Agostino and M. A. Stephens (Eds.), *Goodness-of-fit techniques*, pp. 97–193. New York: Marcel Dekker.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B* 36, 111–47.
- Summers, R. and A. W. Heston (1991). The Penn world table (mark 5): an expanded set of international comparisons 1950-1988. *Quarterly Journal of Economics* 106, 327–368.
- Tachibanaki, T., T. Suruga, and N. Atoda (1997). Estimations of income distribution parameters for individual observations by maximum likelihood method. *Journal of the Japan Statistical Society* 27, 191–203.
- Thistle, P. D. (1989). Ranking distributions with generalized Lorenz curves. *Southern Economic Journal* 56, 1–12.
- Thuysbaert, B. (2008). Inference for the measurement of poverty in the presence of a stochastic weighting variable. *Journal of Economic Inequality* 6, 33 – 55.
- Titterton, D. M., U. E. Makov, and A. F. M. Smith (1985). *Statistical Analysis of Finite Mixture Distributions*. New-York: Wiley.

- Verma, V. and G. Betti (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics* 38, 1549–1576.
- Victoria-Feser, M.-P. (1995). Robust methods for personal income distribution models with application to Dagum’s model. In C. Dagum and A. Lemmi (Eds.), *Research on Economic Inequality, Volume 6: Income Distribution, Social Welfare, Inequality and Poverty*, pp. 225–239. Greenwich: JAI Press.
- Victoria-Feser, M.-P. (2000). Robust methods for the analysis of income distribution, inequality and poverty. *International Statistical Review* 68, 277–293.
- Victoria-Feser, M.-P. and E. Ronchetti (1994). Robust methods for personal income distribution models. *Canadian Journal of Statistics* 22, 247–258.
- Victoria-Feser, M.-P. and E. Ronchetti (1997). Robust estimation for grouped data. *Journal of the American Statistical Association* 92, 333–340.
- Xu, K. and L. Osberg (1998). A distribution-free test for deprivation dominance. *Econometrics Reviews* 17(4), 415–429.
- Xu, K. and L. Osberg (2002). The social welfare implications, decomposability, and geometry of the Sen family of poverty indices. *Canadian Journal of Economics* 35, 138–152.
- Yitzhaki, S. and E. Schechtman (2013). *The Gini Methodology: A Statistical Primer*. Springer Series in Statistics. Springer.
- Young, A. (2011). The Gini coefficient for a mixture of lognormal population. Working Paper.
- Zheng, B. (1999). Statistical inferences for testing marginal rank and (generalized) Lorenz dominances. *Southern Economic Journal* 65, 557–570.
- Zheng, B. (2001). Statistical inference for poverty measures with relative poverty lines. *Journal of Econometrics* 101, 337–356.
- Zheng, B. (2002). Testing Lorenz curves with non-simple random samples. *Econometrica* 70, 1235–1243.
- Zheng, B. and B. J. Cushing (2001). Statistical inference for testing inequality indices with dependent samples. *Journal of Econometrics* 101, 315–335.