ESTIMATION OF INEQUALITY INDICES[*]

by

Frank Cowell
London School of Economics and Political Science

Discussion Paper
No.DARP/25
October 1996

The Toyota Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020-7955 6678

---

[*] I am very grateful to Joanna Gomulka, Christian Schluter and Maria-Pia Victoria-Feser for helpful comments on an earlier draft. Research partially supported by the ESRC, Grant no. R000235725. Prepared for: *Income Inequality Measurement: From Theory to Practice*, edited by Jacques Silber.

Abstract

Inequality measures are powerful tools of applied welfare analysis. However, to use the tools effectively one has to take into account the characteristics of the data with which one usually has to work. These raise a number of common statistical problems which are addressed here for both micro-data and group data. The theoretical properties of inequality measures can often be used to simplify these problems and derive implementable algorithms.

Keywords: Inequality, asymptotic estimates, resampling, grouping, robustness.
JEL Nos.: C13, D63.

Contact address: Frank Cowell, STICERD, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, email: **f.cowell@lse.ac.uk**

Estimation of Inequality Indices

1 Introduction

Inequality indices appear both as central concepts in the formal analysis of welfare economics, and as empirical tools applied to micro-data on incomes or to grouped data published by statistical agencies. There is an important interface between theoretical abstraction that is appropriate to the ethical interpretation of income distribution and the practically-minded approach to estimation. The purpose of this chapter is to examine this interface and to point out some of the statistical pitfalls in implementing measures of economic inequality.

The sceptical reader might be wondering why the whole subject cannot just be delegated to a few page references in a standard statistical text. The answer to this is twofold: the special nature of income distribution data, and the special nature of inequality indices. Sections 2 and 3 deal with these special features. Then sections 4 and 5 focus upon the classical estimation problem using micro-data; section 6 extends the analysis to encompass the problem of estimation from grouped data, and section 7 deals with problems of data contamination

2 Data Issues

There are several reasons for claiming that data on income distribution deserve special treatment in the analysis of statistical distributions. For example some have drawn attention to the empirical curiosity of regularity of the shape of income distributions across a wide variety of historical, cultural and economic circumstances: this information about shape is

often useful in statistical modelling for the purposes of estimating income inequality indices. However the primary reason for treating the subject as special lies in the way in which personal income data are usually collected and the way in which they are interpreted in inequality analysis. This is apparent from a brief consideration of two fundamental topics: the meaning of "income" and of the "income receiver".

Applied welfare economics is usually individualistic so that a distribution of income by persons is what is particularly relevant in an economic application. Furthermore when we consider the economic question of what income is - or what it is supposed to represent - it is often that case that one wants it to be defined in a way that is sufficiently comprehensive for it to be a reasonable proxy for a person's economic welfare. For some narrowly-defined concepts of "income" the data may be collected on an individual basis: for example earnings of employees, personal wealth of testators. However, data on broadly-defined income concepts are usually collected at family or household level.

These considerations usually mean that two types of adjustment must be made in order to interpret distributional data in an economically coherent fashion. First, household income must be adjusted to allow for differences in needs between households - the *equivalisation* process. Second, each household must be weighted in the distribution to reflect the number of persons - the *reweighting* process. One can then analyse the distribution of equivalised incomes amongst individual persons.[1]

To express these ideas in a formal model assume that agreement has been reached on the definition of an income concept and on the appropriate way in which incomes are to be

[1] See Cowell (1984), Danziger and Taussig (1979) for a fuller discussion of this issue.

equivalised. For each household write w for the weight of the household in the distribution[2] and x the income to be imputed to each household member. In this treatment we will abstract from the problem of there being inequality within each family. Write W for the set of household weights: if the weights are automatically normalised then $W=[0,1]$, otherwise it is $\mathbb{R}_+$ the non-negative half-line. Also write X for the set of all incomes which we will usually assume to be $\mathbb{R}$ the real line: this allows us to address the important practical problem of zeros and negative values of the income variable. An income distribution can then be characterised by the (bivariate) distribution function $F:W\times X\rightarrow[0,1]$ with standard properties; $F(w,x)$ gives the proportion of the population with weight $\leq w$ and incomes $\leq x$. Write the space of all valid income distribution functions as $\mathcal{F}$. Key concepts in distributional analysis can then be expressed as functionals of the bivariate distribution F. For example the mean $\mu: \mathcal{F}\rightarrow X$ is defined as $\mu(F)=\int wx\ dF(w,x)/\int w\ dF(w,x)$.

3 Inequality

What is an inequality measure? In principle it is just a statistic defined on the space of income distributions $I: \mathcal{F}\rightarrow\mathbb{R}$, but it is given economic meaning by endowing it with properties derived from an appropriate axiom system or some social-welfare function. These properties are what make the statistic special. Chief amongst them is the *transfer principle* (Dalton, 1920), which may be expressed thus: if distribution G is derived from F by a mean-preserving spread[3] (or a sequence of such spreads) then $I(G)>I(F)$. The class of statistics defined by

[2] As we have seen the weights could be just the number of persons in the family. However note that weights play other roles as well - see section 4 below.

[3] The meaning of this is discussed further in Atkinson (1970).

this principle alone is however rather large and unwieldy; and there is some difficulty in agreeing on further restrictions that would generate a single generally applicable subclass. We shall find it expedient to examine a number of subclasses of inequality measures that together encompass most of the indices that are in common use.

Assume for the moment that the weights are normalised by definition so that $\int w \, dF(w,x) \equiv 1$.[4] Then we may define the *basic class*:

$$I(F) = \psi \left(\int w \; \Phi(x, \mu(F)) \; dF(w,x), \; \mu(F) \right) \tag{1}$$

where ψ is monotonic increasing its first argument. For this to make sense as an inequality measure the function Φ has to be convex in its first argument. An important sub-case of (1) is given by the *normalised basic* class ($\mathbf{I}_N$) which consists of measures of the form:

$$\psi \left(\int w \; \phi\left(\frac{x}{\mu(F)}\right) dF(w,x), \; \mu(F) \right) \tag{2}$$

where the function ϕ is defined on incomes that have been normalised by the mean. The class $\mathbf{I}_N$ defined by (2) overlaps with, but is distinct from, the class of *decomposable inequality indices* ($\mathbf{I}_D$) which can be characterised as measures of the form

$$\psi \left(\int w \; \phi(x) \; dF(w,x), \; \mu(F) \right) . \tag{3}$$

In both (2) and (3) ϕ: $X \rightarrow \mathbf{R}$ is an *evaluation function* which, in view of the transfer principle, must be convex; ψ: $\mathbf{R}^2 \rightarrow \mathbf{R}$ in (1)-(3) is a *cardinalisation function*. An example of the inequality index that belongs to the $\mathbf{I}_N$ class (2) but not the $\mathbf{I}_D$ class (3) is the relative mean deviation

[4] If this assumption is not satisfied then the w term in (1) to (10) below must be replaced by $w/\int w \, dF(w,x)$; likewise w' in (10).

$$\int w \left| \frac{x}{\mu(F)} - 1 \right| dF(w,x) \; ; \tag{4}$$

an example of the inequality index that belongs to $\mathbf{I}_D$ but not $\mathbf{I}_N$ is Kolm's index defined as

$$\frac{1}{\beta} \left[\int w \; e^{\beta [x-\mu(F)]} \, dF(w,x) - 1 \right], \tag{5}$$

where β is a positive parameter (Kolm, 1976a, 1976b). Measures that belong to both $\mathbf{I}_N$ and $\mathbf{I}_D$ are characterised by the $\mathbf{I}_{ND}$ class of indices that has as its typical member

$$\psi \left(I^{\varepsilon}_{GE}(F), \; \mu(F) \right), \tag{6}$$

where I^{α}_{GE} is one of the $\mathbf{I}_{GE}$ class of *generalised entropy* measures, given by:

$$I^{\alpha}_{GE}(F) := \frac{1}{\alpha^2 - \alpha} \left[\int w \left[\frac{x}{\mu(F)} \right]^{\alpha} dF(w,x) - 1 \right] \tag{7}$$

and where $\alpha \in \mathbb{R}$ is a parameter reflecting sensitivity to inequality in different parts of the distribution (Bourguignon, 1979; Cowell, 1980; Shorrocks, 1980, 1984). An important subclass of (6) is given by the *Atkinson indices*:

$$I^{\varepsilon}_{A}(F) := \left[\int w \left[\frac{x}{\mu(F)} \right]^{1-\varepsilon} dF(w,x) \right]^{\frac{1}{1-\varepsilon}} \tag{8}$$

where $\varepsilon \geq 0$ is an inequality aversion parameter (Atkinson, 1970). Although the measures (7) and (8) have different cardinalisation functions they are ordinally equivalent in that

$$I_A^{\varepsilon}(F) = \left[\left[\alpha^2-\alpha\right] I_{GE}^{\alpha}(F) + 1 \right]^{\frac{1}{\alpha}} \tag{9}$$

for $\alpha \leq 1$ where $\varepsilon = 1-\alpha$. [5]

Finally an important index that does not belong to the basic class at all is the Gini coefficient

$$\frac{1}{\mu(F)} \iint ww' \left| x - x' \right| dF(w,x)dF(w',x') ; \tag{10}$$

historically this index has occupied such an important place in the literature that it deserves separate special treatment.

4 Estimation from micro data

Now consider how inequality measures would be implemented in practice. The problem is simplified considerably if we restrict attention to particular special classes of inequality measures such as $\mathbf{I}_N$ and $\mathbf{I}_D$ in section 3. First let us extend the notation of section 2 by introducing the following family of weighted moments about zero

$$\mu_{q\alpha}(F) := \int w^q x^{\alpha} \, dF(w,x). \tag{11}$$

for any $q \in \{0,1,2\}$, $\alpha \in \mathbb{R}$. The moment $\mu_{10}(F)$ can be interpreted as the "effective population

[5] The formulas (7) and (8) have appropriate limiting forms for some parameter values. The evaluation function becomes - log (x) in the case $\alpha=0$ ($\varepsilon=1$), and for the formula (7) the evaluation function becomes x log (x) in the case $\alpha=1$, which correspond's to Theil's first index (Foster, 1983; Theil, 1967).

size",[6] and $\mu_{11}(F)$ is effective total income; mean income is given by $\mu(F)=\mu_{11}(F)/\mu_{10}(F)$. The inequality measures in the $\mathbf{I}_{GE}$ class can then be written

$$\frac{1}{\alpha^2-\alpha}\left[\frac{\mu_{1\alpha}\,\mu_{10}^{\alpha-1}}{\mu_{11}^{\alpha}}-1\right] \tag{12}$$

for $\alpha\neq 0,1$[7] (if the weights are normalised by definition then $\mu_{10}(F)=1$).

Now let us examine the empirical counterparts to these concepts. Assume that a *simple random sample*[8] of n observations is drawn from the distribution F; each observation is a pair (w_i,x_i), where w_i is the weight value of observation i in the sample, and x_i the corresponding income value of observation i, $i=1,2,...n$. Let S_n be the distribution consisting of n point-masses, one at each observation or data point. The counterpart of the moments (11) for the sample are given by

$$m_{q\alpha} := \mu_{q\alpha}(S_n) \tag{13}$$

for any $q \in \{0,1,2\}$, $\alpha\in\mathbf{R}$. Equation (13) may be rewritten:

$$m_{q\alpha} = \frac{1}{n}\sum_{i=1}^{n} w_i^q\, x_i^{\alpha}\ . \tag{14}$$

An unbiased estimator of (12) (a member of the $\mathbf{I}_{GE}$ class) is then given by

---

[6] If income-receivers are households and if the weight on each observation corresponds to the number of persons in each household, then μ_{10} is exactly the number of persons in the population.

[7] The analysis is easily extended to these cases also by introducing modified moments - see Cowell (1989a) for a full treatment.

[8] On the general issues of sampling see Hansen *et al.* (1953), Kish (1965), Levy and Lemeshaw (1991), Scheaffer *et al.* (1990).

$$\frac{1}{\alpha^2-\alpha}\left[\frac{m_{1\alpha}\, m_{10}^{\alpha-1}}{m_{11}^{\alpha}} - 1\right] \tag{15}$$

An important practical point to note here is that in the empirical distribution the weights now play two roles. Apart from their use in reweighting the distribution by households to get the individual income distribution they may also incorporate sample weights: the weight for observation i is then $w_i=w_i^s w_i^h$ where w_i^s is the ith observation's sampling weight and w_i^h is the household-to-individual weighting factor.

The extension of this analysis to the $\mathbf{I}_D$ class is straightforward and requires the moments

$$\mu_{q\phi}(F) := \int w^q\ \phi(x)\ dF(w,x), \tag{16}$$

and the sample counterpart of (16):

$$m_{q\phi} := \frac{1}{n}\sum_{i=1}^{n} w_i^q\ \phi(x_i), \tag{17}$$

The $\mathbf{I}_N$ class and the Gini coefficient - equations (2) and (10) - can be expressed in terms of the sample in the same sort of way. However for examining its statistical behaviour it is helpful to rewrite the expression for the sample Gini as follows. Let $(w_{[i]},x_{[i]})$ be the observation with the ith smallest x-value in the sample. Then the Gini coefficient can be expressed as a weighted sum of the $x_{[i]}$:

$$I_{Gini}(S_n) = \sum_{i=1}^{n} \kappa_{[i]}\ x_{[i]} \tag{18}$$

where the weights are given by:

$$\kappa_{[i]} := \frac{w_{[i]}}{m_{10}m_{11}}\left[2\sum_{j=1}^{i} w_{[j]} - w_{[i]} - 1\right], \quad i=1,2,...n \ . \tag{19}$$

The form (18) is particularly convenient for use in computational algorithms.[9]

5 Inference from micro data

Let us begin with the most convenient subcase, $\mathbf{I}_{ND}$, the intersection of the normalised basic class and the class of decomposable inequality measures.

The basic result can be seen by means of an example based on a simplified version of $\mathbf{I}_{GE}$. Suppose we take a case where $w\equiv 1$ for everyone and a non-normalised form of (12), namely $\mu_{1\alpha}$: this is itself a valid inequality measure for $\alpha > 1$ and $\alpha < -1$, and the sample estimate is of course $m_{1\alpha}$. The variance of the random variable $m_{1\alpha}$ in this special case is

$$\mathscr{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^{\alpha}x_j^{\alpha}\right) - \mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right)^2 \tag{20}$$

where $\mathscr{E}$ is the expectations operator. In view of the facts that $\mathscr{E}(x_ix_j)= \mathscr{E}x_i\mathscr{E}x_j$ if $i\neq j$ and that $\mathscr{E}x_j^{\alpha}= \mu_{1\alpha}$ we find that (20) becomes

$$\frac{1}{n}\left[\ \mu_{2\,2\alpha} - \mu_{1\,\alpha}^2\ \right] \tag{21}$$

and an unbiased estimate of this from the sample is provided by

[9] See Cowell (1989b).

$$\frac{1}{n-1}\left[\, m_{22\alpha} - m_{1\alpha}^2 \,\right] \tag{22}$$

The main results for the $\mathbf{I}_{ND}$ class follow from this. If we express the population and sample moments as vectors:

$$\boldsymbol{\mu} := (\mu_{10}, \mu_{11}, \mu_{1\alpha})^T \tag{23}$$

$$\mathbf{m} := (m_{10}, m_{11}, m_{1\alpha})^T \tag{24}$$

then in the light of (6) and (7) we can see that the population and sample values for a member of $\mathbf{I}_{ND}$ can be written in the form $\Psi(\boldsymbol{\mu})$ and $\Psi(\mathbf{m})$. The following holds asymptotically (Rao, 1973, page 387):

$$\sqrt{n}\,[\mathbf{m} - \boldsymbol{\mu}] \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{25}$$

where $\Sigma:=[\sigma_{ij}]$, $\sigma_{ij} = n\ \mathrm{cov}(m_{1i}, m_{1j})$, $i,j=0,1,\alpha$ and N denotes the normal distribution. From this we obtain as an asymptotic result:

$$\sqrt{n}\,[\,\Psi(\mathbf{m}) - \Psi(\boldsymbol{\mu})] \sim N(0, nV), \tag{26}$$

where

$$V := \frac{1}{n}\,\Psi_{\boldsymbol{\mu}}{}^T \boldsymbol{\Sigma}\, \Psi_{\boldsymbol{\mu}}\ . \tag{27}$$

and $\Psi_{\boldsymbol{\mu}}$ denotes the vector differential:

$$\Psi_{\boldsymbol{\mu}} := \left[\ \frac{\partial \Psi(\boldsymbol{\mu})}{\partial \mu_{10}}, \frac{\partial \Psi(\boldsymbol{\mu})}{\partial \mu_{11}}, \frac{\partial \Psi(\boldsymbol{\mu})}{\partial \mu_{1\alpha}}\ \right]^T . \tag{28}$$

The quadratic form V in (27) is the asymptotic sampling variance of the inequality statistic, Σ is the variance-covariance matrix of the sample moments and Ψ_{μ} encapsulates the role of

the cardinalisation in the sampling variance. In applying this result to the case of the class $\mathbf{I}_{ND}$ we have:

$$\Sigma = \begin{bmatrix} \mu_{20} - \mu_{10}^2 & \mu_{21} - \mu_{11}\mu_{10} & \mu_{2\alpha} - \mu_{1\alpha}\mu_{10} \\ \mu_{21} - \mu_{11}\mu_{10} & \mu_{22} - \mu_{11}^2 & \mu_{2\alpha+1} - \mu_{1\alpha}\mu_{11} \\ \mu_{2\alpha} - \mu_{1\alpha}\mu_{10} & \mu_{2\alpha+1} - \mu_{1\alpha}\mu_{11} & \mu_{22\alpha} - \mu_{1\alpha}^2 \end{bmatrix}, \tag{29}$$

whatever the cardinalisation of the inequality measure. For the specific cardinalisation represented by the $\mathbf{I}_{GE}$ class we would have:

$$\Psi_{\mu} = \frac{\mu_{10}^{\alpha-1}}{\mu_{11}^{\alpha}} \left[[\alpha-1]\frac{\mu_{1\alpha}}{\mu_{10}}, \ \alpha\frac{\mu_{1\alpha}}{\mu_{11}}, \ 1 \right], \tag{30}$$

The bottom right-hand term in (29) corresponds to the expression obtained for the variance in the elementary case (21): it would be the only relevant term if the data were unweighted and one had independent information about the true mean of the distribution. The neighbouring off-diagonal terms show that if the mean is to be estimated from the sample, then its covariance with the income-evaluation function ϕ must be accounted for; likewise the remaining off-diagonal terms in Σ illustrate the way individual weights are correlated with income (terms involving μ_{10}) and with the income-evaluation function ϕ (bottom-left and top-right in the matrix): this correlation depends upon sample design and inherent population heterogeneities which are inherent in inequality measurement. [10] The variance of the inequality estimate in the case of weighted data could be larger or smaller than the corresponding variance in the unweighted case.

This methodology can be extended to inequality indices that do not belong to the $\mathbf{I}_{GE}$ class such as the relative mean deviation (Gastwirth, 1974) although the formulas for the

[10] For further discussion of these issues see Coulter *et al.* (1996).

standard errors are not so neat. It can also be applied to order statistics which form the basis for empirical implementation of Lorenz curves and so also to the Gini coefficient.[11]

The normality of the sampling distribution (26) means that it is straightforward to apply standard statistical tests to problems involving distributional comparisons. For example a straightforward "difference-of-means" test could be applied to test whether inequality in one year was higher than that in another.[12]

As I have emphasised, these results are valid asymptotically. The assumption is sometimes made that sample data on income distribution will, of their nature, have a large n so that in practice the issue of sampling error can be neglected as being of secondary importance. However this should not be assumed to be true in general. First, for some particularly sensitive indices (for example the coefficient of variation) the standard error of the estimate of the inequality index may be large even for apparently large samples. Second it is often the case that the particular problem of economic interest requires a subsample that is of fairly modest size. Furthermore the sample or subsample may be so small that the asymptotic results which are commonly invoked are invalid.[13] Under these circumstances it may be appropriate to use statistical methods which involve resampling with ("the bootstrap") or without ("the jackknife") replacement using S_n as raw materials.[14] Bootstrap estimates have

---

[11] The underlying theory of the sampling distribution of order statistics was developed by Hoeffding (1948) - see also Sillitto (1969). On the Gini coefficient see also Nygård and Sandström (1981), Cowell (1989a), Glasser (1962), Sandström *et al.* (1985, 1988).

[12] Cf Bishop *et al.* (1991), page 463 in an analogous application using Lorenz ordinates

[13] See Maasoumi (1994).

[14] For a discussion of the bootstrap and jackknife approaches see Bhattacharya and Qumsiyeh (1989), Efron (1979, 1982), Hall, (1992), Rubin (1981), Shao and Tu (1995); see also Kish and Frankel (1970) for a discussion of cases where samples are complex. For an application of the bootstrap and jackknife to inequality statistics see Mills and Zandvakili (1995), Yitzhaki (1991).

the advantage of a smaller sampling variance than their jackknife counterparts, but are usually much more time-consuming computationally.

6 Problems with grouped data

Most of the early studies on income distribution and inequality had to be done using grouped data. Even today, when micro-data sets on income are commonly available, estimation from grouped data has an important role to play: some data are available only in grouped form for reasons of confidentiality or political sensitivity; some countries just do not release micro-data. "Grouped data" usually means a data set presented in the following form. X, the set of all incomes, is partitioned into a set of k intervals $\{X_1, X_2, ..., X_k\}$ where $X_i=[a_i, a_{i+1})$, $i=1,2,...,k$; in interval i we may have only few specific items of information such as the total population frequency in the interval and the mean of the interval.

Apart from extending the results on sampling errors,[15] two other issues arise. First, given that some information grouping will have been lost in the process of grouping what bounds can be put on estimates of the inequality measures, and second what central value between those bounds is appropriate?

The issues are conveniently expressed using the basic class of inequality measures; in the case of grouped data a typical member of this class can be expressed:

$$I(F) = \psi\left(\sum_{i=1}^{k+1}\int_{a_i}^{a_{i+1}} w\ \Phi(x,\mu(F))\ \mathrm{d}F(w,x)\ ,\ \mu(F)\right) \qquad (31)$$

We do not know what the detail of the distribution F is within the interval X_i, but we could

[15] On this see Gastwirth *et al.* (1986).

make some alternative extreme assumptions that are consistent with the known partial information. In particular we can find distributions F_L and F_U such that

$$I(F_L) \leq I(F) \leq I(F_U) \tag{32}$$

for a large class of inequality measures, and for a minimal number of prior restrictions on F. Assume that the inequality index I satisfies the transfer principle. If we know μ_i, the mean of interval i, then F_L (providing a lower bound on I) can be found by assuming that the distribution within X_i is just a pointmass at μ_i, and F_U (giving a least upper bound on I) is found by assuming that there are two point masses within X_i, - one at a_i and the other arbitrarily close to a_{i+1}, - in proportions chosen so that the implied mean equals μ_i.[16]

The second issue raises the question of what method of interpolation one should use. The idea is to estimate F by a collection $\{\hat{F}_i: i=1,2,...,k\}$ of within-interval distribution functions. Clearly there is a large number of candidate functions for possible use: what constitutes a "good" interpolation? We could think of simplicity, ease of interpretation or flexibility of form as appropriate criteria, but the touchstone for an interpolation method is surely the performance of inequality estimates using that interpolation. Perhaps the most straightforward interpolation method is the simple histogram where each $\hat{F}_i$ is rectangular over X_i: but this will be in general be inconsistent with grouped data where interval means are known. However the only slightly more complicated *split*-histogram interpolation provides very satisfactory results: the interpolated distribution is rectangular over each subinterval $[a_i, \mu_i)$ and $[\mu_i, a_{i+1})$. Although the method is unsophisticated in that it just produces a discontinuous "stepped" density function, it performs as well as more complex interpolation algorithms in terms of the inequality estimates that it provides. Remarkably one finds that

[16] For other cases with finer or coarser information about the intervals see Cowell (1991) and Gastwirth (1975), Gastwirth and Glauberman (1976).

unsophisticated and sophisticated interpolation rules alike yield results that are approximated by a very simple rule for basic-class inequality measures: using the bounding methods discussed (32) you take ⅔ of the lower-bound estimate of the inequality index and add it to ⅓ of the upper-bound value.[17] Of course the quality of inequality estimates from grouped data will depend on the way in which X has been partitioned into component intervals and the information about the distribution within each interval. What is crucial is the availability of data on interval means: k intervals with information about frequencies and interval means gives much more accuracy than $2k$ intervals with information about frequencies alone.[18]

Finally let us consider some problems related to intervals 1 and k which have been glossed over in the above discussion. If X is $\mathbf{R}_+$ then $a_1=0$; otherwise a_1 could be unbounded below, that is the bottom interval might just be infinitely wide so as to accommodate indefinitely large losses; a_{k+1} is usually seen as inherently more problematic in that it is commonly assumed that X is unbounded above. These extreme cases usually require special treatment: typically one models the distribution in the open interval with a functional form that has suitable asymptotic properties.[19]

---

[17] See Cowell and Mehta (1982). In the case of the Gini coefficient instead of a (⅔,⅓)-rule you have a corresponding (⅓,⅔)-rule - see Cowell (1995, page 116). For further discussion of the implementation of the interpolation methods see the Appendix in Cowell (1995).

[18] For discussion of how the interval boundaries should be chosen see Aghevli and Mehran (1981), and Davies and Shorrocks (1989).

[19] For example in many cases the *Pareto distribution* is used, which is given by

$$F(x) = 1 - \left[\frac{x}{x_0}\right]^{\alpha}$$

where x_0 and α are parameters to be estimated from the top one or two intervals. For the method estimation see Cowell (1995, pages 160,161) and Needleman (1978).

7 Robustness

The final problem area in the field of inequality estimation concerns the effect of data contamination on estimates of inequality indices. In a sense all data should be regarded as potentially contaminated. However carefully the sample may have been designed, however carefully interviewing and coding procedures are carried out, in practice tiresome errors will creep in. The errors could be the result of human fallibility or of wilful misreporting; sometimes they are actually the result of misunderstanding (weekly for monthly income) or of data coders trying to be helpful (recoding negative incomes as small positive values).

To see the implications of this for the problem of inequality measurement let us express data contamination in an analytically tractable form. Suppose F is the true (but unobservable) income distribution so that $I(F)$ is the true amount of inequality in the population. Now let $H^{(y)}$ be an elementary perturbation distribution which consists of a point mass at income y; this can be represented as the probability distribution:[20]

$$H^{(y)}(w, x) = \begin{cases} 1 & \text{if } w = 1 \text{ and } x \geq y \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

and from (33) we may define the following mixture distribution:

$$G_{\delta}^{(y)} = [1-\delta] F + \delta H^{(y)} \tag{34}$$

where the parameter δ captures the importance of the perturbation. It is, of course, the mixture distribution $G_{\delta}^{(y)}$ that is actually observed. To quantify the impact of the contamination upon the statistic under consideration we may use the concept of the *influence*

[20] This specification ensures that, while we continue to use the same notation as in earlier sections, the weights w play no role hear. For the standard case see Cowell and Victoria-Feser (1996a).

function.[21] In the present case this is given by

$$\mathrm{IF}(y;\, I,\, F) = \lim_{\delta \to 0} \frac{I(G_\delta^{(y)}) - I(F)}{\delta} \tag{35}$$

or, where the derivative exists, by

$$\mathrm{IF}(y;\, I,\, F) = \frac{\partial\, I(G_\delta^{(y)})}{\partial \delta}\bigg|_{\delta=0} \tag{36}$$

The influence function is a tool that characterises the sensitivity of a statistic to "dirt" in the data: it quantifies the importance of an infinitesimal amount of contamination upon the value of the statistic. Here the statistic is the estimator of the inequality measure and (36) indicates to what extent estimated inequality is stable in the presence of a small proportion of arbitrary extreme observations. The influence function captures information about the bias of the estimate of the inequality measure.[22] It is important to know how the IF will behave for various types of data contamination for a wide class of inequality measures and, in particular, to know whether it can be unbounded for contamination at some point. If it were unbounded this would imply that a single observation - if sufficiently extreme - could drive the inequality measure by itself.

Let us compute the IF for the class $\mathbf{I}_{GE}$. Substituting (7) into (36) we obtain which becomes

[21] See Hampel (1974), and Hampel *et al.* (1986).

[22] It is the first-order term in the linear expansion of the asymptotic bias of the estimator - see Hampel *et al.* (1986).

$$\mathrm{IF}(y;\, I^{\alpha}_{\mathrm{GE}},\, F) \;=\; \frac{\partial\, I^{\alpha}_{\mathrm{GE}}(G^{(y)}_{\delta})}{\partial\delta}\bigg|_{\delta=0}$$

$$= \frac{1}{\alpha^2-\alpha}\,\frac{\partial}{\partial\delta}\,\frac{\int w\, x^{\alpha}\,[\,[1-\delta]\,\mathrm{d}F(w,x) \;+\; \delta\,\mathrm{d}\,H^{(y)}(w,x)]}{\mu\left(G^{(y)}_{\delta}\right)^{\alpha}}\bigg|_{\delta=0} \tag{37}$$

$$\mathrm{IF}(y;\, I^{\alpha}_{\mathrm{GE}},\, F) \;=\; \frac{y^{\alpha} \;+\; \int w\, x^{\alpha}\,\mathrm{d}F(w,x)\left[\;\alpha-1 \;-\; \dfrac{\alpha y}{\mu(F)}\;\right]}{[\alpha^2-\alpha]\,\mu\,(F)^{\alpha}} \tag{38}$$

Equation (38) neatly illustrates two aspects of the contamination problem. First, the evaluation function itself may be unbounded for contamination at some point on the income line. A typical example of this is I_{GE} with $\alpha<0$ (ordinally equivalent to the Atkinson index with inequality aversion parameter $\varepsilon>1$): if the point of contamination y is close to zero then y^{α}, the first term in the numerator of (38), becomes very large, which means that a false observation of a very low income may have an overwhelming impact upon the estimate of the inequality index. The other aspect is that the influence of contamination on the mean may play a rôle. Again in (38) we can see that the term in parentheses in the numerator will be unbounded if y is unbounded: this implies that a false observation of a very large income can have such an impact on the mean that it seriously distorts estimates of the inequality index.

There are several strategies for dealing with this problem. In many cases researchers use informal screening methods to weed out what appear to be alien observations by eye: this has the obvious disadvantage of arbitrariness. Alternatively one could model the distribution - or part of it - using an appropriate functional form (such as the Pareto distribution, lognormal or gamma) and estimating the parameters of the fitted distribution using robust

techniques (Cowell and Victoria-Feser, 1996a). Finally one could carry out a sensitivity analysis of inequality estimates on systematically "trimmed" data from which a proportion α of extreme values have been removed (Cowell and Victoria-Feser, 1996b).

Two instructive lessons which may be drawn from this analysis may appear surprising in the light of conventional wisdom on inequality measures and their estimation. The first is that the cardinalisation function is important. This is evident from the specification of ψ in (1)-(3): the explicit dependence on the mean μ and the fact that the mean is a non-robust statistic will mean that different cardinalisations of the same inequality ordering have fundamentally different statistical properties. To illustrate this contrast the standard definition of the Kolm index (5) (which is non-robust because of the mean) with the ordinally equivalent form

$$\frac{1}{\beta}\left[\int w\, e^{\beta x}\, dF(w,x) - 1\right], \tag{39}$$

which is robust. The second lesson is that the problem will not go away with increasing sample size, in contrast to the issues discussed in sections 4 and 5.

8 Concluding remarks

Inequality analysis involves constructing a practical bridge between ethical or mathematical principles that are used to give meaning to the concept of inequality, and the nitty-gritty of data-handling and processing that are the stuff of competent applied economists and statisticians. The basic questions which the practical analyst must address can be summarised as: (1) where have the data come from and what may have happened to it along the way? and (2) how do these practical issues affect the choice of tool for distributional analysis?

By addressing these questions intelligent implementation of inequality concepts can perhaps assist in clarifying the picture of developments in income distribution, and in influencing national economic policies.

9 References

Aghevli, B.B., and Mehran, F. (1981), "Optimal grouping of income distribution data", *Journal of the American Statistical Association*, **76**.

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

Bhattacharya, R. N. and Qumsiyeh, M. (1989), "Second Order L^p-Comparisons Between the Bootstrap and Empirical Edgeworth Expansion Methodologies", *Annals of Statistics*, 17, 160-169.

Bishop, J. A., Formby, J. P. and Smith, W. P. (1991) "International comparisons of income inequality: Tests fot Lorenz dominance across nine countries", *Economica*, **58**, 461-477.

Bourguignon, F. (1979), "Decomposable income inequality measures", *Econometrica*, **47**, 901-920.

Coulter, F.A.E., Cowell, F.A., and Jenkins, S.P. (1996), "Inequality Estimation with Weighted Data" , *Distributional Analysis Research Programme Discussion Paper* 26, STICERD, LSE.

Cowell, F.A. (1980), "On the structure of additive inequality measures", *Review of Economic Studies*, **47**, 521-531

Cowell, F.A. (1984), "The structure of American income inequality", *Review of Income and Wealth;* **30**; 351-375

Cowell, F.A. (1989a), "Sampling variance and decomposable inequality measures", *Journal of Econometrics*, **42**, 27-41.

Cowell, F.A. (1989b), "Analysis of income distributions using microcomputer technology", *Research on Economic Inequality*, **1**, 249-267.

Cowell, F.A. (1991), "Bounds on inequality measures under alternative informational assumptions", *Journal of Econometrics*, **48**, 1-14.

Cowell, F.A. (1995), *Measuring Inequality*, Harvester Wheatsheaf Publishers. (First edition 1977)

Cowell, F.A. and Mehta, F. (1982) "The estimation and interpolation of inequality measures", *Review of Economic Studies*, **49**, 273-290.

Cowell, F.A. and Victoria-Feser, M.-P. (1996a), "Robustness properties of inequality measure" *Econometrica*, **64**, 77-101.

Cowell, F.A. and Victoria-Feser, M.-P. (1996b), "Welfare judgements in the presence of contaminated data", *Distributional Analysis Research Programme Discussion Paper* 13, STICERD, LSE.

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

Danziger, S. and Taussig, M.K. (1979), "The income unit and the anatomy of income distribution", *Review of Income and Wealth*, **25**, 365-375.

Davies, J.B., and Shorrocks A.F. (1989), "Optimal grouping income and wealth data", *Journal of Econometrics*, **42**, 97-108.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jacknife", *Annals of Statistics*,**7**, 1-26.

Efron, B. (1982), *The Jacknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Foster, J.E. (1983), "An axiomatic characterization of the Theil measure of income Inequality", *Journal of Economic Theory*, **31**, 105-121.

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. and Glauberman, M. (1976), "The interpolation of the Lorenz curve and Gini index from grouped data", *Econometrica*, **44**, 479-483.

Gastwirth, J.L., Nayak, T.K and Krieger,A.N.(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.

Glasser, G.J. (1962) "Variance formulas for the mean difference and the coefficient of concentration", *Journal of the American Statistical Association*, **57**, 648-654.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hansen, M. H., Hurwitz, W.N. and Madow, W.G. (1953): *Sample Survey Methods and Theory, Volume II*, John Wiley and Sons, New York.

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., Ronchetti, E., Rousseeuw, P.J. and Stahel, W.A.(1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley: New York.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution", *The Annals of Mathematical Statistics*, **19**, 293-325.

Kish, L. (1965), *Survey Sampling*, John Wiley and Sons, New York.

Kish, L. and M. Frankel (1970), "Balanced Repeated Replications for Standard Errors", *Journal of the American Statistical Association*, **65**(331), 1071-1093.

Kolm, S.-Ch. (1976a), "Unequal inequalities I", *Journal of Economic Theory*, **12**, 416-442.

Kolm, S.-Ch. (1976b), "Unequal inequalities II", *Journal of Economic Theory*, **13**, 82-111.

Levy, P. and Lemeshaw, S. (1991), *Sampling of Populations: Methods and Applications*, John Wiley and Sons, New York.

Maasoumi, E., (1994), "Empirical Analysis of Inequality and Welfare", in Schmidt, P. and Pesaran, H.(eds.) *Handbook of Applied Microeconomics.*

Mills, J.A. and Zandvakili, S. (1995) "Statistical inference via bootstrapping for measures of inequality",mimeo, University of Cincinnati.

Needleman, L. (1978), "On the approximation of the Gini coefficient of concentration", *Manchester School*, 105-122.

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

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, Wiley, New York

Rubin, D. B. (1981): "The Bayesian Bootstrap", *Annals of Statistics*, 9, 130-134.

Sandström, A., Wretman, J.H. and Walden, B. (1985), "Variance estimators of the Gini coefficient: simple random sampling", *Metron*, **43**, 41-70.

Sandström, A., Wretman, J.H. and Walden, B. (1988), "Variance estimators of the Gini coefficient: probability sampling", *Journal of Business and Economic Statistics*, **6**,

113-120.

Scheaffer, R., Mendenhall, W. and Ott, L. (1990), *Elementary Survey Sampling, Fourth Edition*, Duxbury Press, California.

Shao, J. and Tu, D. (1995), *The Jackknife and the Bootstrap*, Springer-Verlag, New York.

Sillitto, G.P. (1969), "Derivation of approximations to inverse distribution function of a continuous univariate population from the order statistics of a sample", *Biometrika*, **56**, 641-650.

Shorrocks, A.F. (1980), "The class of additively decomposable inequality measures", *Econometrica*, **48**, 613-625.

Shorrocks, A.F. (1984), "Inequality decomposition by population subgroup", *Econometrica*, **52**, 1369-13.

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

Yitzhaki, S. (1991), "Calculating jackknife variance estimators for parameters of the Gini method", *Journal of Business and Economic Statistics*, **9**, 235-238