ISSN 1352-2469 **Distributional Analysis Research Programme**

Distributional Dominance with Dirty Data

by

Frank A. Cowell
London School of Economics

and

Maria-Pia Victoria-Feser
Université de Genève

Discussion Paper
No. DARP 51
August 2001

Distributional Analysis Research Programme
The Toyota Centre
Suntory and Toyota International
Centres for Economics and
Related Disciplines
London School of Economics
Houghton Street
London WC2A 2AE

Distributional Analysis Research Programme

The Distributional Analysis Research Programme was established in 1993 with funding from the Economic and Social Research Council. It is located within the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) at the London School of Economics and Political Science. The programme is directed by Frank Cowell. The Discussion Paper series is available free of charge and most papers are downloadable from the website. To subscribe to the DARP paper series, or for further information on the work of the Programme, please contact our Research Secretary, Sue Coles on:

Telephone:	UK+20 7955 6678
Fax:	UK+20 7955 6951
Email:	s.coles@lse.ac.uk
Web site:	http://sticerd.lse.ac.uk/DARP

Abstract

Distributional dominance criteria are commonly applied to draw welfare inferences about comparisons, but conclusions drawn from empirical implementations of dominance criteria may be influenced by data contamination. We examine a non-parametric approach to refining Lorenz-type comparisons and apply the technique to two important examples from the LIS data-base.

¤ Keywords: Distributional dominance; Lorenz curve; robustness

¤ JEL Classification: C13,D63

¤ Correspondence to: Professor F. A. Cowell. STICERD, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK

¤ Partially supported by ESRC grant #R000 23 5725. The second author is partially supported by the Swiss National Fund. We are grateful to the staff of the Luxembourg Income Study for cooperation in data analysis and to Julie Litchfield and Ceema Namazie for research assistance.

1 Introduction

This paper addresses the issue of how practical comparisons of income distributions can be founded on a sound statistical and economic base when there is good reason to believe that the data in at least one of the distributions are "dirty". Dirt includes the possibility of obvious gross errors in the data (such as arise from coding or transcribing mistakes) and also other more innocuous observations that in some sense do not really belong to the income-data set. The problem is often handled pragmatically: some empirical studies have concentrated upon a subset of the distribution delimited either by population subgroup (taking prime-age males only, for example) or by arbitrarily excluding some of the data in the tails (Gottschalk and Smeeding 2000). Although this research technique seems sensible the question of whether it is appropriate remains open – "appropriateness" here being understood in terms of the statistical properties of the underlying economic criteria. This question matters because the economic criteria are used explicitly or implicitly to make normative judgments and perhaps policy recommendations.

In this paper we combine consideration of practical ad hoc techniques (section 2.1) with an investigation of the relationship between economic ranking principles and statistical tools (sections 2.2 and 2.3). In section 3 we introduce considerations of data contamination and their likely impact on the estimates of statistics associated with distributional dominance; we also propose a method dealing with the contamination problem that uses a family of dominance comparisons based on the statistical concept of the trimmed mean. Finally section 4 illustrates the application of these methods in terms of Lorenz comparisons over time and between countries using the data-base of the Luxembourg Income Study.

2 Distributional Dominance

2.1 Informal methods

Empirical studies of income distribution use informal ranking criteria as a matter of routine. There is a variety of good reasons for doing so: they usually involve easy computations, and they have a direct intuitive appeal; more importantly, they are usually connected to deeper points that are particularly relevant to applied welfare economists. Some prominent examples of the informal approach are:

- Pragmatic indices involving quantiles. These include the semi-decile ratio (Wiles 1974, Wiles and Markowski 1971) and the comparative

function of Esberger and Malmquist (1972). An extreme example of the same type is the range – literally the maximum minus the minimum income, but sometimes implemented in practice as a difference between extreme quantiles.

- The "Parade of incomes" introduced by Pen (1971). This provides a persuasive picture of snapshot inequality and of the implications of an income distribution that is changing through time – see for example Jenkins and Cowell (1994).

- The use of distributive shares (sometimes known as quantile shares).

The quantile method can be explicitly linked to formal welfare criteria. For example in Rawls' work on a theory of justice there is a discussion of how to implement his famous "difference principle" which focuses upon the least advantaged: to do this Rawls himself suggests that it might be interpreted relative to the median of the distribution.[1] So too can the distributive shares approach: changes in the relative income shares of, say, the richest and the poorest 10% slices of the distribution can be directly interpreted in terms of the principle of transfers (Dalton 1920).

2.2 A formal framework

Assume that the concept of income and of income receiver have been well defined. An individual's income is a number $x \in \mathsf{X}$, where $\mathsf{X} \subseteq \mathsf{R}$ and R is the real line. Let F be the set of probability distributions (distribution functions) with support X. An income distribution is one particular member $F \in \mathsf{F}$.

In this approach a statistic of any distribution $F \in \mathsf{F}$ is a functional $T(F)$, for example the mean $\mu : \mathsf{F} \mapsto \mathsf{R}$ given by

$$\mu(F) := \int x dF(x). \tag{1}$$

The properties of any functional T may play a role in both economic and statistical interpretations. Of particular interest here is the case where the range of T is a profile of values rather than a single number as in the example of (1); T is then a family of statistics. Individual family members may be of interest in their own right; the behaviour of the whole family when applied to a pair of distributions F and G will provide important information about

[1] See Rawls (1972) page 98.

distributional comparisons that is richer than that provided by a single real-valued functional.

The basic distributional concept employed here is a ranking which amounts to a partial ordering on the space of distributions F. Use the symbol $\succeq_T$ to denote the ranking induced on F by a statistic T, from which a number of other concepts are derived:

Definition 1 For all $F, G \in \mathsf{F}$:
(a) (strict dominance) $G \succ_T F \Rightarrow G \succeq_T F$ and $F \not\succeq_T G$
(b) (equivalence) $G \sim_T F \Rightarrow G \succeq_T F$ and $F \succeq_T G$
(c) (non-comparability) $G \perp_T F \Rightarrow G \not\succeq_T F$ and $F \not\succeq_T G$.

For example, if T were the Lorenz criterion then (a) would read in plain language "Distribution G strictly Lorenz-dominates distribution F if G weakly dominates F and F does not weakly dominate G". We use the T-ranking concept to motivate a discussion of welfare economic issues in distributional analysis and their statistical implementation.

2.3 Statistics and ranking criteria

As noted in section 2.1 quantiles and incomplete moments provide convenient tools for judgments about income distributions. To give economic meaning to a class of distributional rankings we introduce standard welfare criteria expressed in terms of a social-welfare function (SWF). The SWF embodies the ethical judgments of a normative analyst or policy maker; in statistical terms the SWF is just a statistic of the distribution. To get specific results it is useful to focus upon a particular additively separable class of SWF:

Definition 2

$$\mathsf{W} := \left\{ W : \mathsf{F} \mapsto \mathsf{R} \,\middle|\, W(F) = \Psi\left(\int u(x) dF(x) \right) \right\}. \tag{2}$$

where $u : \mathsf{X} \mapsto \mathsf{R}$ is an evaluation function of individual incomes, and $\Psi : \mathsf{R} \mapsto \mathsf{R}$ is monotonic.

Let W_1 be the subclass of W for which the evaluation function is everywhere increasing, and W_2 be the subclass of W_1 for which the evaluation function is also concave. The SWF subclasses W_1 and W_2 play a crucial role in interpreting two fundamental ranking principles – first- and second-order distributional dominance – and have a close relationship with the informal quantiles and shares criteria introduced in Section 2.1.

First-order dominance criteria are based on the quantiles of the distribution:

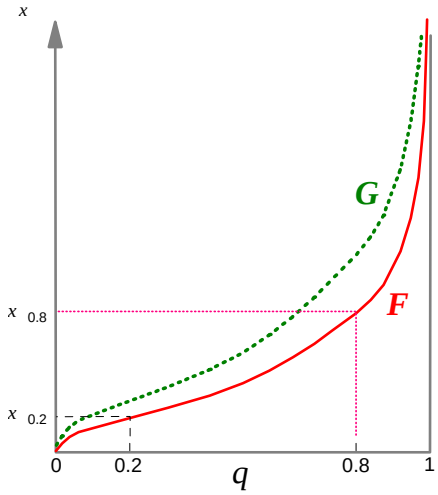

Figure 1: G first-order dominates F

Definition 3 For all $F \in \mathsf{F}$ and for all $0 \leq q \leq 1$[2]

$$Q(F;q) = \inf\{x | F(x) \geq q\} = x_q \tag{3}$$

For example $\{Q(F;0.1), Q(F;0.2), ..., Q(F;0.9)\}$ are the deciles of the distribution F. For any distribution of income F, the graph of Q formalises the concept of Pen's Parade (see section 2.1) and has a simple welfare interpretation: if every quantile in distribution G is greater than the corresponding quantile in distribution F – if some persons "grow" (and nobody shrinks) as in the $F \rightarrow G$ transformation depicted in Figure 1 – then distribution G will be assigned a higher welfare level by every SWF in class W_1.[3] This monotonicity criterion is consistent with the assumption of the Pareto principle and the absence of externalities in the SWF (Amiel and Cowell 1994).

The first-order dominance criterion $\succeq_Q$ is sometimes considered to be less than ideal[4] and so is of interest to consider the second-order criterion. This requires the following.

[2]See Gastwirth (1971). Alternative definitions are available. For example we may define a quantile correspondence $\tilde{Q} : \mathsf{F} \times [0,1] \mapsto \Xi$ such that $\tilde{Q}(F;q) = \{x : F(x) = q\}$ where $\Xi := \{\{x : a \leq x \leq b\} : a, b \in \mathsf{X}\}$ - Cf Kendall and Stuart (1977), pp. 39-41. This redefinition does not affect the results that follow.

[3]Formally

$$(G \succeq_Q F) \Leftrightarrow (\forall W \in \mathsf{W}_1 : W(G) \geq W(F))$$

(Saposnik 1981, 1983).

[4]One objection is on practical grounds: in empirical applications it often happens that neither distribution first-order dominates the other although Bishop et al. (1991) argue

Definition 4 For all $F \in \mathsf{F}$ and for all $0 \leq q \leq 1$, the cumulative income functional is defined by:

$$C(F;q) := \int_{\underline{x}}^{Q(F;q)} x dF(x). \tag{4}$$

where $\underline{x} := \inf \mathsf{X}$.

By definition $C(F;0) = 0$, $C(F;1) = \mu(F)$. For a given $F \in \mathsf{F}$ the graph of $C(F,q)$ against q describes the generalised Lorenz curve (GLC), which characterises another principal welfare property: if every cumulant in distribution G is greater than the corresponding cumulant in distribution F then distribution G will be assigned a higher welfare level by every SWF in class $\mathsf{W_2}$.[5] From the fundamental concept C one can derive two other important analytical distributional tools for drawing welfare-conclusions from income data, namely the relative Lorenz curve (Lorenz 1905):

$$L(F;q) := \frac{C(F;q)}{\mu(F)} \tag{5}$$

and the absolute Lorenz curve (Moyes 1987):

$$A(F;q) := C(F;q) - q\mu(F) \tag{6}$$

The (relative) Lorenz curve – the graph of $L(F;q)$ against q, closely related to the first moment function [6] – encapsulates the intuitive principle of the distributional-shares ranking referred to in Section 2.1 illustrated in Figure 2. We will examine the implementation of (5) and (6) in Section 4 below.

Further interpretations of the basic properties of the C-criterion can be obtained by restricting the admissible SWFs to a subset of $\mathsf{W_2}$. Take the subclass that have the additional property that proportional increases in all incomes yield welfare improvements:

$$\left\{ W \left| W \in \mathsf{W_2}; \forall F \in \mathsf{F}, k > 1 : W\left(F^{(\times k)}\right) > W(F) \right. \right\}. \tag{7}$$

that in international comparisons the second-order criterion $\succeq_C$ defined below does not resolve many of the "incomparable cases" where $G \perp_Q F$. There is also a theoretical objection in that $\succeq_Q$ does not employ all the standard principles of social welfare analysis: in particular it does not incorporate the principle of transfers (as does $\succeq_C$).

[5] Formally $(\forall F, G \in \mathsf{F} : \; G \succeq_C F) \Leftrightarrow (\forall W \in \mathsf{W_2} : W(G) \geq W(F))$ – see Kolm (1969), Marshall and Olkin (1979), Shorrocks (1983).

[6] This is a function $\Phi : \mathsf{X} \mapsto [0,1]$ defined for any $F \in \mathsf{F}$ as $\Phi(x) = L(F; F(x)) = \frac{1}{\mu(x)} \int^x y dF(y)$ – (Kendall and Stuart 1977).

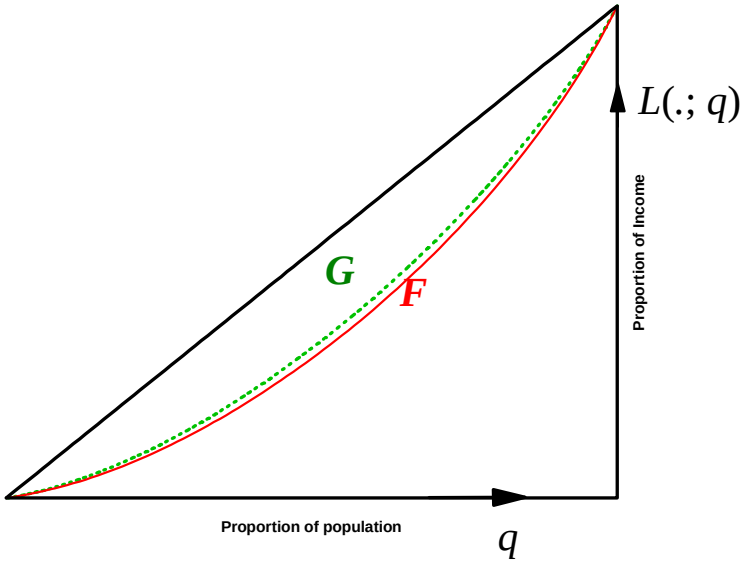

Figure 2: G (relative-) Lorenz-dominates F

where $F^{(\times k)}$ is a rescaling of F defined by $F^{(\times k)}(x) = F\left(\frac{x}{k}\right)$. Then distribution G dominates F for SWFs in this restricted class if and only if $G \succeq_L F$ and $\mu(G) \geq \mu(F)$.[7] Alternatively take the subclass for which uniform absolute increases in all incomes yield welfare improvements:

$$\left\{W \left| W \in \mathsf{W}_2; \forall F \in \mathsf{F}, k > 0 : W\left(F^{(+k)}\right) > W(F)\right.\right\}. \tag{8}$$

where$F^{(+k)}$ is a translation of F given by $F^{(+k)}(x) = F(x-k)$. Then $G \succeq_A F$ (see Figure 3) and $\mu(G) \geq \mu(F)$ if, and only if, $W(G) \geq W(F)$ for all W in W_2 that also satisfy (8).

[7]The basic insights of the income-cumulation function were originally obtained for $\mathsf{F}(\mu)$ the set of distributions with a given mean μ :

$$(\forall F, G \in \mathsf{F}(\mu) : \ G \succeq_L F) \Leftrightarrow (\forall W \in \mathsf{W}_2 : W(G) \geq W(F))$$

- see Atkinson (1970).

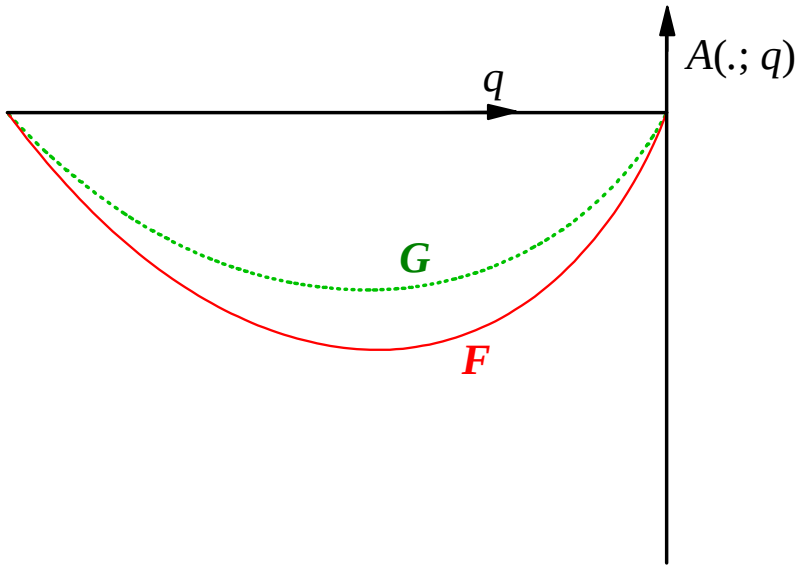

Figure 3: G absolute Lorenz-dominates F

3 Dirty Data: An Approach

3.1 Contamination and Robustness

To assume that data will automatically give a reasonable picture of the "true" picture of a distributional comparison would obviously be reckless in the extreme. A prudent applied researcher will anticipate that, because of miscoding and misreporting and other types of mistake, some of the observations will be incorrect, and this may have a serious impact upon distributional comparisons (Van Praag et al. 1983). Obviously if one had reason to suspect that this sort of error were extensive in the data sets under consideration the problem of distributional comparison might have to be abandoned because of unreliability. But it is possible that there might be a serious problem of comparison even if the amount of contamination were small, so that the data might be considered "reasonably clean".

Let us briefly review a standard model of this type of problem.[8] Suppose that the "true" distributions that we wish to compare are denoted by F and G; but because of the problem of data-contamination we cannot assume that the data we have to hand have really been generated by F and G. What we

[8]This approach is based upon the work of Hampel (1968, 1974), Hampel et al. (1986), Huber (1981).

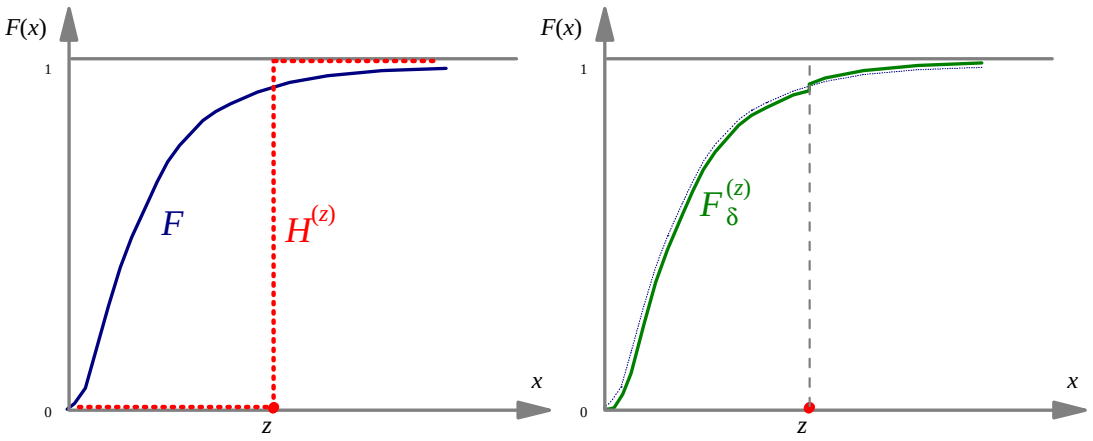

Figure 4: Contamination modelled as a mixture of distributions

actually observe instead of F is a distribution in some neighbourhood of F. An elementary case is illustrated in Figure 4 where a mixture distribution has been constructed by combining the "true" distribution F with a point mass at income z

$$F_\epsilon^{(z)} = [1-\epsilon]\, F + \epsilon H^{(z)} \tag{9}$$

where

$$H^{(z)}(x) = \begin{cases} 1 & \text{if } x \geq z \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The degenerate distribution $H^{(z)}$ represents a simple form of data contamination at point z; ϵ indicates the importance of the contamination; the convex combination $F_\epsilon^{(z)}$ is the observed distribution, and F remains unobservable.

As we have noted if ϵ were large we cannot expect to get sensible estimates of income-distribution statistics; but what if the contamination were very small? To address this question for any given statistic T one uses the influence function given by

$$IF(z;T,F) := \lim_{\epsilon \to 0} \left[\frac{T(F_\epsilon^{(z)}) - T(F)}{\epsilon} \right] \tag{11}$$

Then under the given model of data-contamination (9) the statistic T is robust if IF in (11) is bounded for all $z \in \mathsf{X}$.

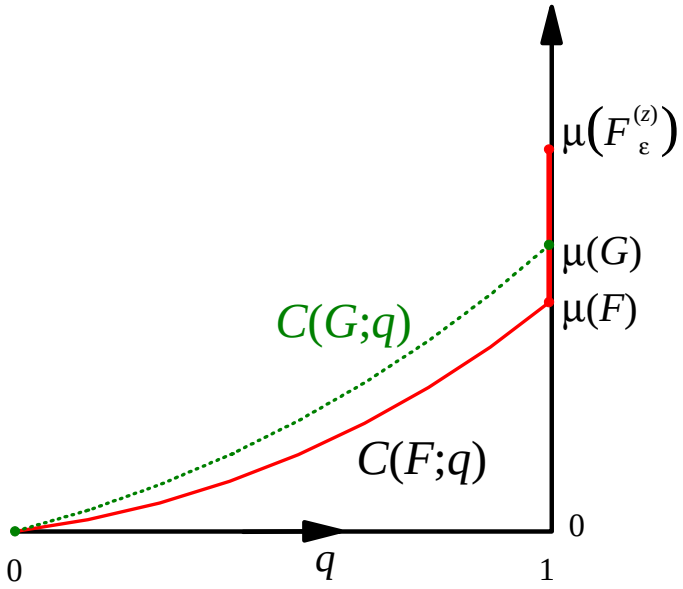

Figure 5: A small amount of contamination changes second-order dominance conclusions

In Cowell and Victoria-Feser (2002, 1996) we have shown that most inequality measures are non-robust, but that most poverty indices with exogenous poverty lines are robust – see also Monti (1991). However the non-robustness problem is more than pervasive than that which emerges in connection with inequality measures: the same type of approach can be used to show that while first-order dominance criteria are usually robust,[9] second- and higher-order dominance criteria (and associated ranking tools) are not. (Cowell and Victoria-Feser 2002). The result is illustrated in Figure 5 which depicts contamination of distribution F at a very high level of income. It is clear that $G \succeq_C F$ and that this conclusion would emerge from the bulk of the data; on the other hand the whole data set suggests that $G \perp_C F_\epsilon^{(z)}$ – no clear-cut distributional dominance.

3.2 Ranking criteria: trimming

Because ranking criteria can be misleading in the presence of data contamination it is desirable to have a procedure that enables one to control systematically for suspect values that may distort distributional comparisons

[9]For further discussion of the statistical implementation of first order criteria see Ben Horim (1990) and Stein, Pfaffenberger, and French (1987).

using second-order ranking criteria. A natural approach would be to use an established tool in the statistical literature, the "trimmed mean" and extend the idea to Lorenz curve analysis. The trimmed mean of distribution F with trimming parameter α is

$$\begin{aligned} \bar{X}_\alpha(F) &= \tfrac{1}{1-2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y dF(y) \\ &= \tfrac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(t) dt. \end{aligned} \tag{12}$$

where $\alpha \in [0, \frac{1}{2})$ is the balanced trimming proportion. This estimator of location has intuitive appeal: one removes the αn smallest and the αn largest observations in a sample of size n, and calculates the mean of the remaining observations: notice that $\lim_{\alpha \to 0.5} \bar{X}_\alpha(F) = Q(F, 0.5)$ – in the limiting case as α approaches 50% the trimmed estimate of the mean approaches the median.

Likewise consider trimmed Lorenz Curves as estimators of Lorenz curves. One has to interpret the quantile and income-cumulation functions (3) and (4). α-trimming the data means that $Q(F; q) \in (Q(F; \alpha), Q(F; 1-\alpha))$ and thus $q \in (\alpha, 1-\alpha)$. However, it makes sense to consider a more general trimming method which would in particular include the single-tail trimming case. Indeed, this case is appropriate when one can form an a priori judgment about the nature of the contamination, for example when contamination is assumed to affect only the lower tail of the distribution. Let $\underline{\alpha}$ and $1-\bar{\alpha}$ be the lower and upper trimming and $\alpha := \underline{\alpha} + (1-\bar{\alpha})$ be the total trimmed amount. Then the α-trimmed generalized Lorenz, Lorenz and absolute Lorenz curves[10] for $q \in (\underline{\alpha}, 1-\bar{\alpha})$ are respectively given by

$$c_{\alpha,q} := C_\alpha(F; q) = \frac{1}{1-\alpha} \int_{Q(F;\underline{\alpha})}^{Q(F;q)} u dF(u) \tag{13}$$

$$l_{\alpha,q} := L_\alpha(F; q) = \frac{C_\alpha(F; q)}{C_\alpha(F; 1-\bar{\alpha})}, \tag{14}$$

$$A_\alpha(F; q) = (1 - \bar{\alpha} - \underline{\alpha}) \cdot C_\alpha(F; q) - C_\alpha(F; 1-\bar{\alpha}) \cdot (q - \underline{\alpha}). \tag{15}$$

– Cf equations (4), (5) and (6). From equations (13-15) we have $C_\alpha(F; \underline{\alpha}) = 0$, $L_\alpha(F; \underline{\alpha}) = 0$, $A_\alpha(F; \underline{\alpha}) = 0$ and $L_\alpha(F; 1-\bar{\alpha}) = 1$, $A_\alpha(F; 1-\bar{\alpha}) = 0$.

[10] See the similar concept of restricted dominance discussed by Atkinson and Bourguignon (1989).

The IFs of these trimmed Lorenz curves will be bounded for all q because extreme values in the data are automatically removed, for all $\underline{\alpha}, 1-\bar{\alpha} > 0$. Trimmed Lorenz curves can be thought of as Lorenz curves on a restricted sample in which $100\underline{\alpha}$ percent of the bottom observations and $100(1-\bar{\alpha})$ percent of the top observations have been trimmed away[11]. Estimates can be obtained by replacing F by the empirical distribution $F^{(n)}(x) = \frac{1}{n}\sum_{i=1}^{n} H^{(x_i)}(x)$.

3.3 Confidence Intervals

When comparing distributions using ranking criteria, it is also important to be able to provide confidence intervals for the later. In Cowell and Victoria-Feser (1999), formulas are given for several distributional statistics including ranking criteria for full and trimmed samples. In particular, we have that the asymptotic covariance of $\sqrt{n}C_\alpha(F^{(n)};q)$ and $\sqrt{n}C_\alpha(F^{(n)};q')$ with $q \leq q'$ is given by $\omega_{qq'}/(1-\alpha)^2$ where

$$\begin{aligned}\omega_{qq'} \quad : \quad &= [qQ(F;q) - \underline{\alpha}Q(F;\underline{\alpha}) - [1-\alpha]\, c_{\alpha,q}] \\ &[[1-q']\, Q(F;q') - [1-\underline{\alpha}]\, Q(F;\underline{\alpha}) + [1-\alpha]\, c_{\alpha,q'}] - \\ &[Q(F;q)\,[1-\alpha]\, c_{\alpha,q} - [1-\alpha]\, s_{\alpha,q}] \\ &+Q(F;\underline{\alpha})\,[(q-\underline{\alpha})Q(F;q) - [1-\alpha]\, c_{\alpha,q}] \end{aligned} \tag{16}$$

with $s_{\alpha,q} := S(F;q) = \frac{1}{1-\alpha}\int_{Q(F;\underline{\alpha})}^{Q(F;q)} u^2 dF(u)$. For the Lorenz curve ordinates the asymptotic variance is

$$v_{qq'} = \frac{1}{(1-\alpha)^2\mu_\alpha^4}\left[\mu_\alpha^2\omega_{qq'} + c_{\alpha,q}c_{\alpha,q'}\omega_{\overline{\alpha}\overline{\alpha}} - \mu_\alpha c_{\alpha,q}\omega_{q'\overline{\alpha}} - \mu_\alpha c_{\alpha,q'}\omega_{q\overline{\alpha}}\right]$$

with $\mu_\alpha = C_\alpha(F;1-\overline{\alpha})$. These covariances can be estimated by their empirical counterpart – see Appendix 6.1.

3.4 Choosing the trimming proportions

The sampling properties of the key distributional statistics can provide simple choice criterion. Indeed, let $\tilde{\tilde{F}}_\alpha$ be the trimmed distribution[12] and T a

[11] This is a practice that is sometimes adopted in pragmatic discussion of inequality trends. See also the discussion of related issues by Howes (1996).

[12] The trimmed distribution is:

$$\tilde{F}_\alpha(x) := \begin{cases} 0 & \text{if } x < Q(F,\underline{\alpha}) \\ \frac{F(x)-\underline{\alpha}}{1-\alpha} & \text{if } Q(F,\underline{\alpha}) \leq x < Q(F,\overline{\alpha}) \\ 1 & \text{if } x \geq Q(F,\overline{\alpha}) \end{cases} .$$

statistic of interest and consider the following concept of efficiency

$$\kappa(\alpha) := \frac{\text{var } T(F)}{\text{var } T(\tilde{F}_\alpha)} \tag{17}$$

Clearly one would expect a higher value of α to reduce $\kappa(\alpha)$. The implied trade-off of robustness against efficiency enables the researcher to make an informed choice about the extent of trimming that may be reasonable in making distributional comparisons.

Now (17) clearly implies that this choice is conditional upon specification of T: which statistic would be appropriate? It seems reasonable to require that this be one of second-order distributional dominance, but this raises a further difficulty: there is an uncountable infinity of statistics $C(\cdot; q)$ and selecting one, or a few of these would appear to arbitrary. However, there is a simple argument to suggest that one particular case is especially important. Not all values of q in the unit interval will be relevant in computing efficiency under trimming: the very process of trimming "nibbles away" some of the interval. If one is interested in trimming of arbitrary size then it seems to be of particular interest to examine cases where $T(\tilde{F}_\alpha)$ is well defined for arbitrary α. In the case of a balanced trim this implies focusing attention on $C(\cdot; 0.5)$ or its relative-Lorenz counterpart $C(\cdot; 0.5)/\mu(\cdot)$.

$\kappa(\alpha)$ also depends on the underlying income distribution F. For the purposes of illustrating the technique and to obtain an idea of the efficiency losses involved we used a number of examples of the Dagum type I distribution given by

$$f(x; \beta, \lambda, \delta) = (\beta + 1)\lambda\delta x^{-(\delta+1)}(1 + \lambda x^{-\delta})^{-(\beta+1)}. \tag{18}$$

Two examples are illustrated in Figure 6. From these two simulated datasets, we computed the sampling variances for the trimmed and untrimmed cases, with lower, upper and balanced trims.

The results are illustrated in Figure 7 where the vertical axis gives estimated values of $\kappa(\alpha)$ as defined in (17). One can see that the efficiency loss depends on the underlying model and the type of trim. For small trimming quantities it is not very large and for larger trimming quantities it can be either quite large or reasonable. It is however, difficult to draw a general conclusion and the results presented here can at most provide a rough guideline.

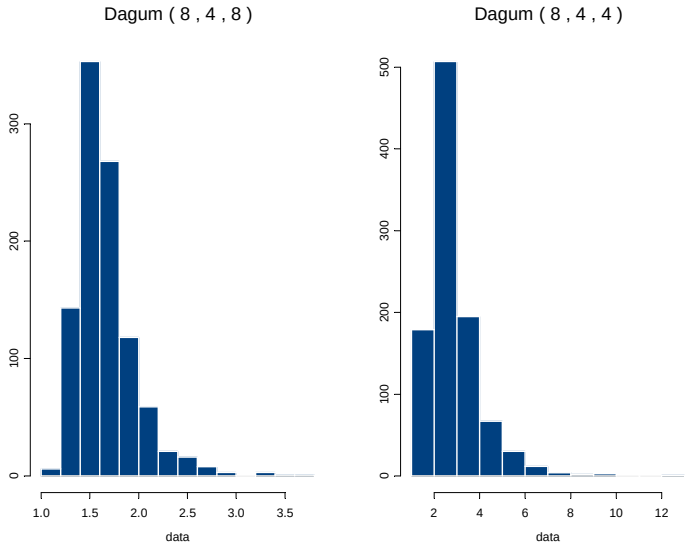

Figure 6: The Dagum Distribution

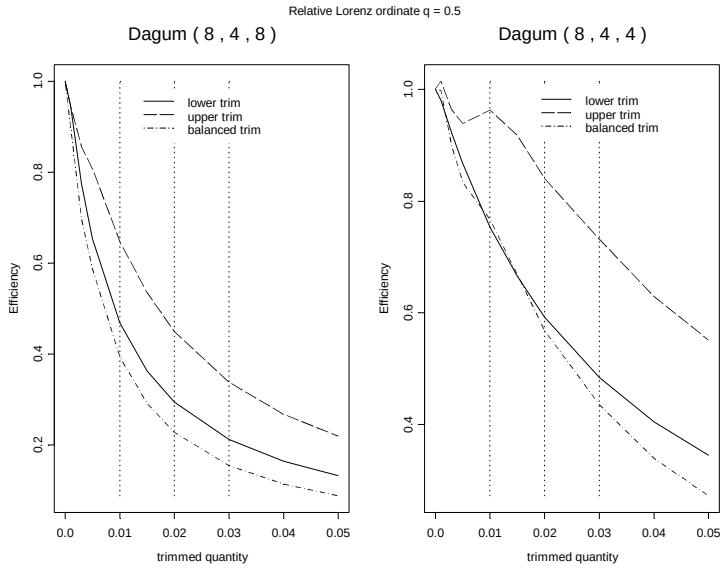

Figure 7: Efficiency Under Trimming for $C(\cdot, 0.5)/\mu(\cdot)$.

4 Empirical examples

The trimming approach offers a practical tool for the comparison of income distribution when one wants an explicit control for taking account of the influence of outliers. We use the analysis of sections 3.1 and 3.2 to examine more carefully two aspects of conventional wisdom concerning comparisons of income distribution. In each case the data are taken from the LIS (Luxembourg Income Study) data-base and refer to real income per equivalent adult distributed amongst individuals – see Appendix, section 6.2.

4.1 Cross-country comparison: Sweden and Germany

The received wisdom suggests that 1980s Sweden is more equal than Germany. However, is this actually borne out by the data, and what are the implications for standard welfare comparisons? To investigate this we use data for Sweden 1981 and (West) Germany 1983. Given standard definitions it immediately appears that $F_{\mathsf{Germany}} \succeq_C F_{\mathsf{Sweden}}$ so that there is no question but that the German income-distribution second-order (generalised-Lorenz) dominates that for Sweden. However we also find $F_{\mathsf{Sweden}} \perp_A F_{\mathsf{Germany}}$ and $F_{\mathsf{Sweden}} \succeq_{A_{0.005}} F_{\mathsf{Germany}}$: given a very slight trim of both tails (a half of one percent) Sweden absolute-Lorenz dominates Germany.

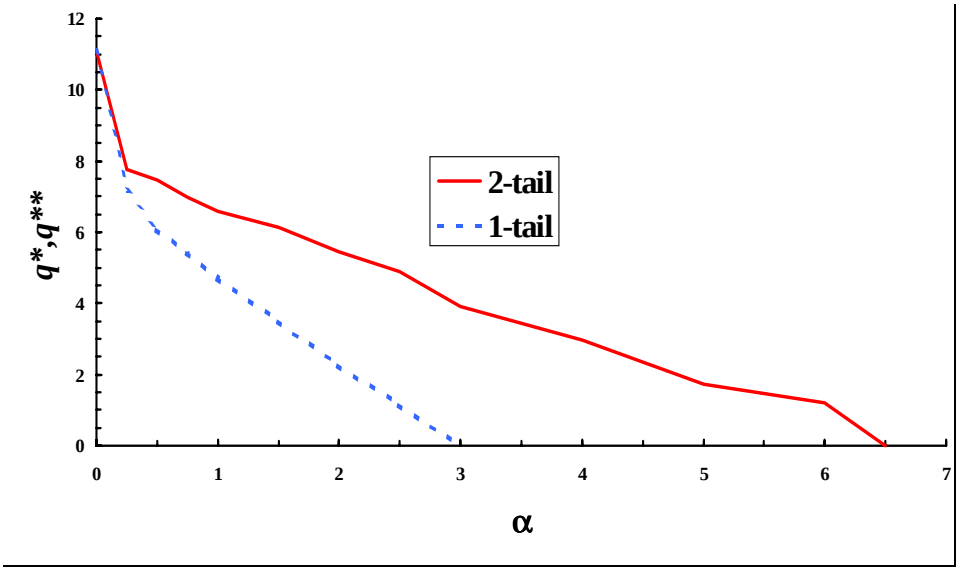

Figure 8: Is Sweden more equal than Germany?

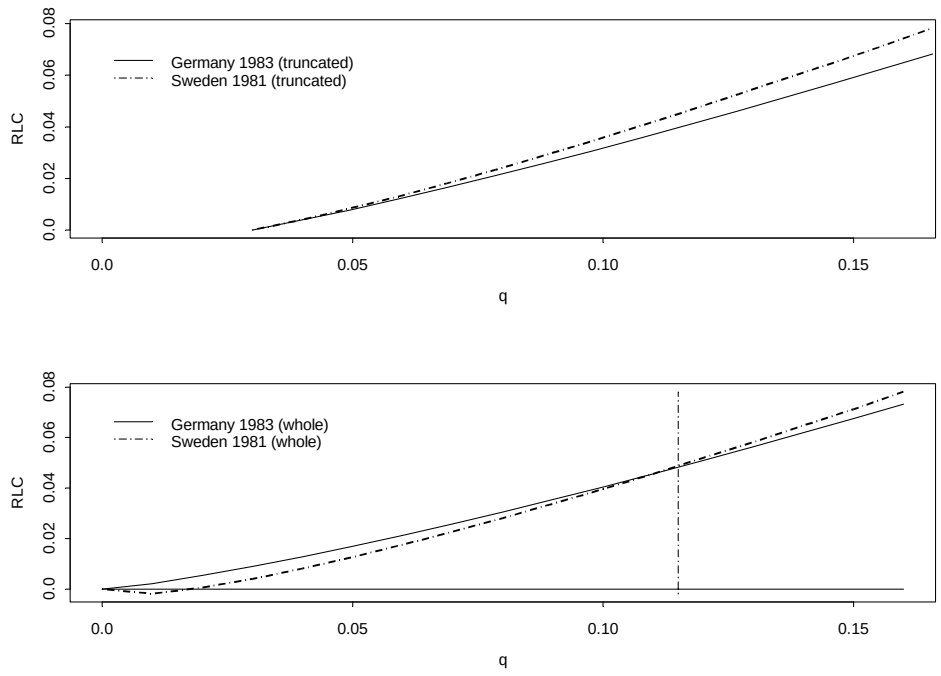

Figure 9: Germany vs Sweden: the effect of a 3% bottom-tail trim on the Lorenz comparison

What of inequality? As Figures 8 and 9 show there is an ambiguity for the raw data – $F_{\mathsf{Sweden}} \perp_L F_{\mathsf{Germany}}$. – which is due to a single intersection of the Lorenz curves. Figure 8 depicts the truncation profiles – the position of the switch-point (where the Lorenz curves intersect) for two types of trim expressed as functions of α – $q^{**}(\cdot)$ for the balanced two-tail trim (solid curve), and $q^{*}(\cdot)$ for the one-sided lower-tail trim (dotted curve). Let the points where the truncation profiles intersect the horizontal axis be α^{**} and α^{*} respectively. Then:

$$\begin{aligned} q^{**}(0) &= q^{*}(0) = 0.11 \\ q^{*}(\alpha) &= 0,\ \alpha \geq \alpha^{*} = 0.030 \\ q^{**}(\alpha) &= 0,\ \alpha \geq \alpha^{**} = 0.065 \end{aligned}$$

We have $F_{\mathsf{Sweden}} \succeq_{L_\alpha} F_{\mathsf{Germany}}$ only if a trim of 3% of the observations is carried out on the lower tail, or a balanced trim of 6.5%. May we say that Sweden is less unequal than Germany? Consider two points here.

First let us apply the analysis of section 3.3 in order to compute confidence intervals for the RLC of Germany 1983 and Sweden 1981 on 3% bottom-tail trimmed samples. The results are presented in Figure 10. The relative-Lorenz dominance is indeed significant, except for the first q. This result is not surprising because usually the sample sizes are large and therefore the standard errors are small. If dominance is not significant, then this should appear at the smallest or the largest q-values.

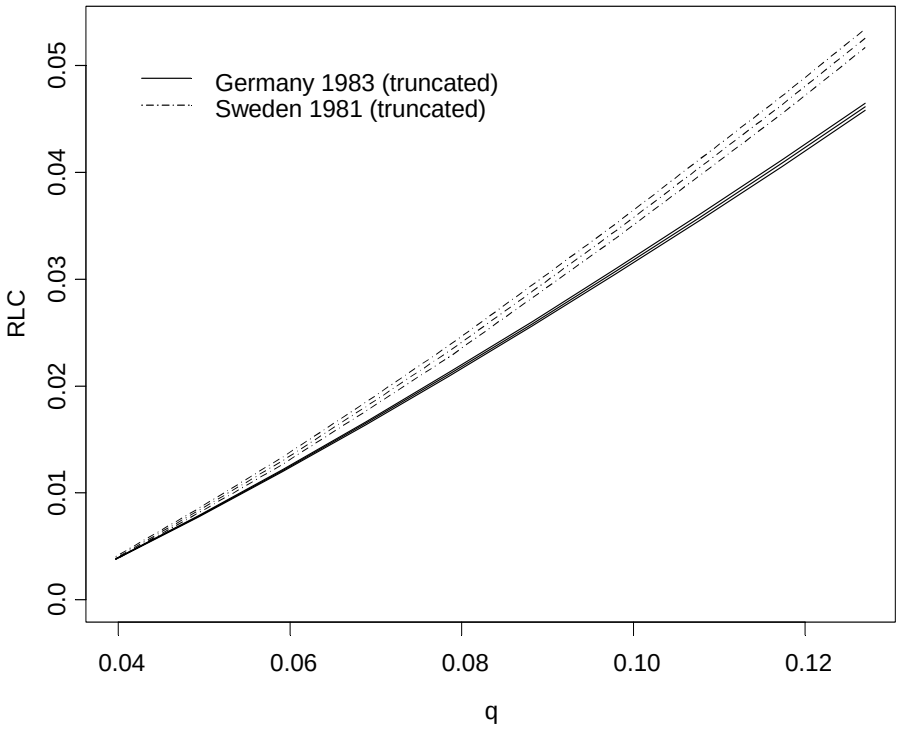

Figure 10: RLC of Germany versus Sweden with confidence intervals

Second, note the behaviour of the truncation profiles. Both q^{**} and q^{*} initially fall rapidly for α very close to zero and thereafter decrease more gently. So the Lorenz comparison is certainly very sensitive to presence or absence the first few observations (in either the 1- or 2-tail case) but the issue is clearly not just one of hypersensitivity to very small incomes. It seems unreasonable to suppose that the true picture is of strict Lorenz dominance in

that at least 1000 observations would have to be discarded from the German data ($n \simeq 42,000$) in order for this conclusion to obtain.

4.2 Inequality over time: the US in the 1980s

The same technique may of course be applied to comparisons within one country, but between two points in time. In the United States the conventional wisdom is perhaps even more sharp in its sketch of recent events – inequality rose over the 1980s. Again the fact is – perhaps surprisingly – that the raw data do not reveal an unambiguous increase in inequality, in the standard sense of relative-Lorenz dominance. It might appear that this is principally due to the presence of negative incomes in the first centile group: as we will see this is not quite the whole story. Note first that $F_{\mathsf{US86}} \perp_C F_{\mathsf{US79}}$ – we do not have first- or second-order distributional dominance (see Figure 11 – the generalised Lorenz curves intersect at about $q = 0.02, 0.10, 0.32$), but $F_{\mathsf{US79}} \succeq_A F_{\mathsf{US86}}$.

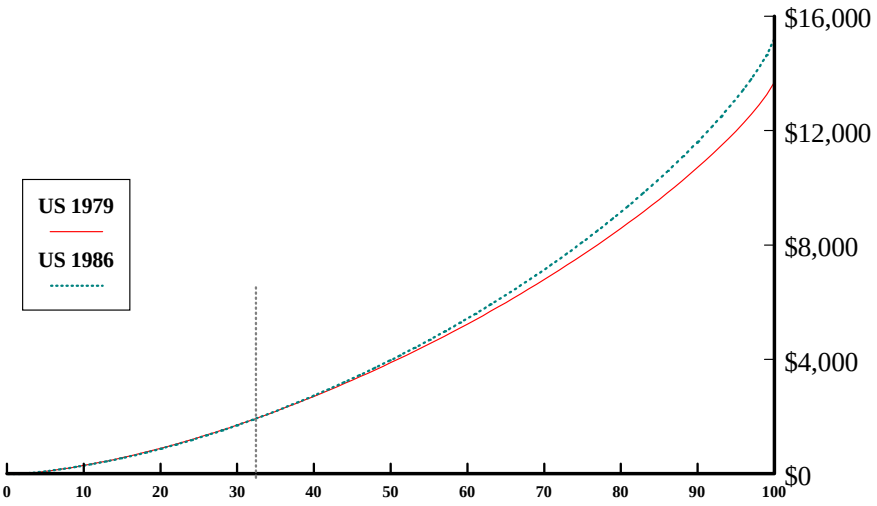

Figure 11: US 1986 does not second-order dominate US 1979

The trimming procedure is more complex. The problem of negative incomes is disposed of by a very modest (less than 0.5%) trim; but there remains a problem of multiple intersections of the Lorenz curves at the bottom tail (there are intersections between $q = 0.01$ and $q = 0.02$ and $q = 0.03$ and $q = 0.04$). Figure 12 plots $q^{**}(\alpha)$ and $q^{*}(\alpha)$ in this case: in view of the multiple intersections, these values are interpreted as the maximum switch point between the two Lorenz curves for each value of α. The outcome of the

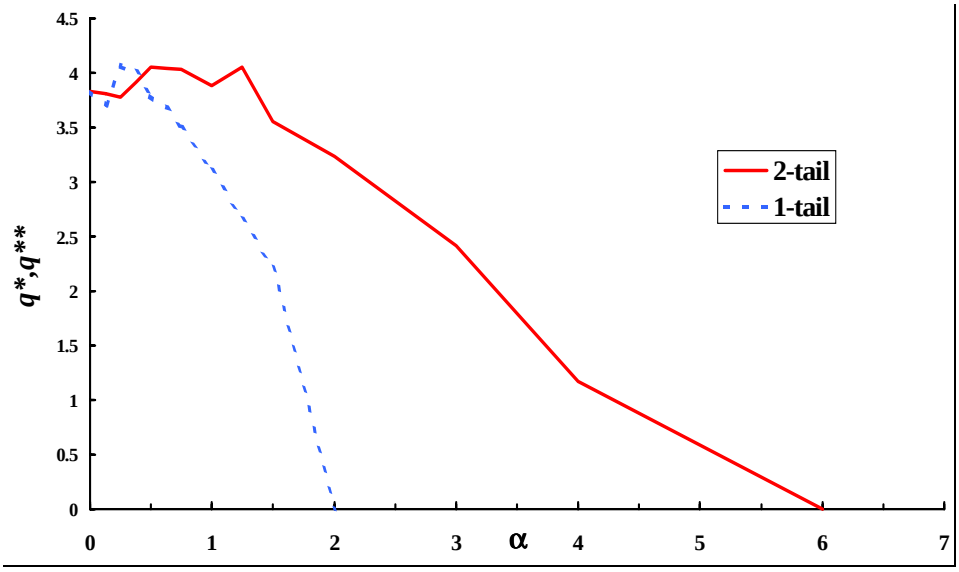

Figure 12: Did Inequality Rise in the US?

α-trimming procedure is interesting in that – by contrast to the Germany-versus Sweden example – neither $q^{**}(.)$ nor $q^{*}(.)$ is monotonic. By dropping some 200 to 300 observations (2 percent) in the single-tailed trim, or 600 to 700 observations ($4\frac{1}{2}$ percent) in the two tail trim one may then conclude that $F_{\mathsf{US79}} \succeq_{L_\alpha} F_{\mathsf{US86}}$. – see Figure 13.

However there are interesting points in common with the Germany-versus-Sweden example. First, for values of α in the range $[0, 0.01]$ one finds a relationship between the switch-point and α which is clearly different from the relationship that holds in the neighbourhood of the points α^{**} and α^{*}. Second, the shape of the two-tail trim truncation profile follows closely that of the one-tail trim.[13] Thirdly, all the action appears to come from the lower tail: in the distributional comparisons reported in subsections 4.1 and 4.2 we also carried out an upper-tail experiment; here the hypothesis is that the data contamination is concentrated in the high incomes, and can be interpreted as potentially misreported data. However in this case the ranking results turned out to be insensitive to the trim.

[13]On multiplying by $\frac{1}{2}$ the horizontal scale of the graph of $q^{**}(\cdot)$ one finds that it lies extremely close to that of $q^{*}(\cdot)$: dropping $2\alpha\%$ of the sample in a two-tail trim has almost exactly the same impact on the Lorenz intersection as dropping $\alpha\%$ of the sample in a lower-tail trim.

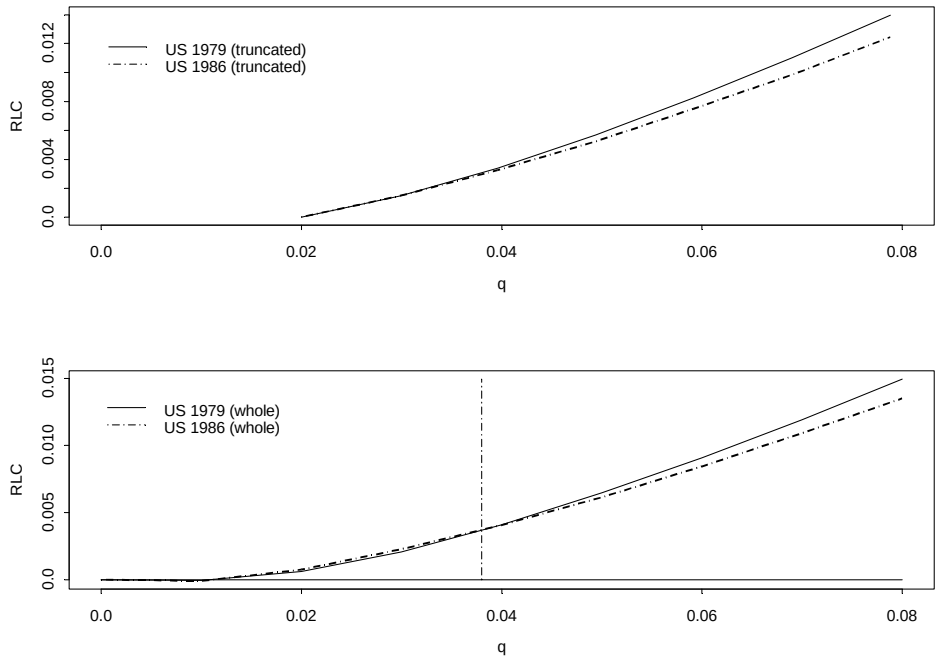

Figure 13: US inequality: the effect of a 2% bottom-tail trim on Lorenz comparisons

5 Conclusions

Given that second-order distributional-dominance criteria are known to be non-robust it is important to have practical methods of coping with the impact of potentially "dirty" data in either tail of an income distribution. One-tail or two-tail (balanced) trimming provides an obvious way to extend the simple distributional-dominance criteria. In effect the researcher has the option of trading off efficiency of the distributional-dominance statistic with robustness. In this way one can place intuition about comparisons of empirical Lorenz curves on an appropriate analytical foundation.

References

Amiel, Y. and F. A. Cowell (1994). Monotonicity, dominance and the Pareto principle. Economics Letters 45, 447–450.

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

Atkinson, A. B. and F. Bourguignon (1989). The design of direct taxation and family benefits. Journal of Public Economics 41, 3–29.

Ben Horim, M. (1990). Stochastic dominance and truncated sample data. Journal of Financial Research 13, 105–116.

Bishop, J. A., J. P. Formby, and P. D. Thistle (1991). Rank dominance and international comparisons of income distributions. European Economic Review 35, 1399–1409.

Buhmann, B., L. Rainwater, G. Schmaus, and T. Smeeding (1988). Equivalence scales, well-being, inequality and poverty: Sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. Review of Income and Wealth 34, 115–142.

Coulter, F. A. E., F. A. Cowell, and S. P. Jenkins (1992). Equivalence scale relativities and the extent of inequality and poverty. Economic Journal 102, 1067–1082.

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

Cowell, F. A. and M.-P. Victoria-Feser (1996). Robustness properties of inequality measures. Econometrica 64, 77–101.

Cowell, F. A. and M.-P. Victoria-Feser (1999). Statistical inference for welfare indices under complete and incomplete information. Distributional Analysis Discussion Paper 47, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (2002). Welfare rankings in the presence of contaminated data. Econometrica (forthcoming).

Dalton, H. (1920). Measurement of the inequality of incomes. Economic Journal 30 (9), 348–361.

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

Esberger, S. E. and S. Malmquist (1972). En Statisk Studie av Inkomstutveklingen. Stockholm: Statisk Centralbyr å n och Bostadssyrelsen.

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

Gottschalk, P. and T. M. Smeeding (2000). Empirical evidence on income inequality in industrialized countries. In A. B. Atkinson and F. Bourguignon (Eds.), Handbook of Income Distribution, Chapter 3. Amsterdam: North Holland.

Hampel, F. R. (1968). Contribution to the Theory of Robust Estimation. Ph. D. thesis, University of California, Berkeley.

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.

Howes, S. R. (1996). The influence of aggregation on the ordering of distributions. Economica 63, 253–272.

Huber, P. J. (1981). Robust Statistics. New York: John Wiley.

Jenkins, S. P. and F. A. Cowell (1994). Dwarfs and giants in the 1980s: The UK income distribution and how it changed. Fiscal Studies 15 (1), 99–118.

Kendall, M. and A. Stuart (1977). The Advanced Theory of Statistics. London: Griffin.

Kolm, S.-C. (1969). The optimal production of social justice. In J. Margolis and H. Guitton (Eds.), Public Economics, pp. 145–200. London: Macmillan.

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

Marshall, A. W. and I. Olkin (1979). Inequalities: Theory and Majorization. New York: Academic Press.

Monti, A. C. (1991). The study of the Gini concentration ratio by means of the influence function. Statistica 51, 561–577.

Moyes, P. (1987). A new concept of Lorenz domination. Economics Letters 23, 203–207.

Pen, J. (1971). Income Distribution. London: Allen Lane, The Penguin Press.

Rawls, J. (1972). A Theory of Justice. Oxford: Oxford University Press.

Saposnik, R. (1981). Rank-dominance in income distribution. Public Choice 36, 147–151.

Saposnik, R. (1983). On evaluating income distributions: Rank dominance, the Suppes-Sen grading principle of justice, and Pareto optimality. Public Choice 40, 329–336.

Shorrocks, A. F. (1983). Ranking income distributions. Economica 50, 3–17.

Stein, W. E., R. C. Pfaffenberger, and D. W. French (1987). Sampling error in first-order stochastic dominance. Journal of Financial Research 10, 259–269.

Van Praag, B. M. S., A. J. M. Hagenaars, and W. Van Eck (1983). The influence of classification and observation errors on the measurement of income inequality. Econometrica 51, 1093–1108.

Wiles, P. J. D. (1974). Income Distribution, East and West. Amsterdam: North Holland.

Wiles, P. J. D. and S. Markowski (1971). Income distribution under communism and capitalism. Soviet Studies 22, 344–369,485–511.

6 Appendix

6.1 Computational Method

When using a database such as the LIS database from which the microdata cannot be recovered directly, to compute LC or RLC with confidence intervals at each chosen q, one can follow this procedure:

1. Define the percentiles p (say $p = 0, 0.01, 0.02, \ldots, 1$) and trimming proportions $\underline{\alpha}$ and $1 - \bar{\alpha}$

2. Extract from the database the personal incomes and the weights and sort incomes and weights by incomes

3. Define a new variable empperc which is made of the cumulative weights and divide all elements by the maximum, i.e. the last element. Keep only the incomes and the weights for which empperc is between $\underline{\alpha}$ and $\bar{\alpha}$. This defines the trimmed incomes trinc and weights trwgt.

4. Define tottrwgt as the sum of all trwgt and nbtrinc as the number of elements in trinc

5. Define a new variable trempperc which is made of the cumulative trwgt and divide all elements by the maximum, i.e. the last element. (and keep the value of the maximum of the cumulative trwgt in say totweight)

6. For each percentile $p > 0$ then do:

   (a) Select the elements of trinc and trwgt for which trempperc is between p and the previous p (for example for $p = 0.56$, trempperc is between 0.56 and 0.55). Call them respectively trincp and trwgtp

   (b) Define $m1_p$ as the sum of trincp·trwgtp divided by tottrwgt, $m2_p$ as the sum of trincp·trincp·trwgtp divided by tottrwgt and x_p as the maximum of trincp

7. Define $q = \underline{\alpha} + (1 - \alpha)p$. Then for $q > \underline{\alpha}$, $c_{\alpha,q}$ is estimated by the cumulative sum of the $m1_p$, $p \leq \frac{q-\underline{\alpha}}{(1-\alpha)}$ and $c_{\alpha,\underline{\alpha}} = 0$, $s_{\alpha,q}$ is estimated by the cumulative sum of the $m2_p$, $p \leq \frac{q-\underline{\alpha}}{(1-\alpha)}$ and $s_{\alpha,\underline{\alpha}} = 0$. Note that $\mu_\alpha = c_{\bar{\alpha}}$.

95% confidence intervals for the GLC and the RLC are respectively given $(c_q - 1.96\omega_{qq}; c_q + 1.96\omega_{qq})$ and $(c_q/\mu_\alpha - 1.96\upsilon_{qq}, c_q/\mu_\alpha + 1.96\upsilon_{qq})$ in which ω_{qq} (and therefore υ_{qq}) are estimated using the estimates of $c_{\alpha,q}$ and $s_{\alpha,q}$. Note that $m1_p$ and/or $m2_p$ can take very large values depending on the measurement scale of the incomes. It might be useful for numerical reasons to divide all incomes by a properly chosen quantity.

6.2 Data Specification

LIS permits comparison of different countries' income distributions based on consistent international definitions of income and the income receiver. Accordingly the same basic specifications were used both the (Germany, Sweden) and the (US 1979, US 1986) comparisons in section 4. The sample sizes were:

Germany 1983	42,752
Sweden 1981	9,625
US 1979	15.928
US 1986	12,600

The income distributions are formed using the following concept of equivalised incomes (Buhmann, Rainwater, Schmaus, and Smeeding 1988) (Coulter, Cowell, and Jenkins 1992):

$$y = \frac{hhy}{hhsize^\alpha}.$$

where hhy is net family (unit) income after tax, $hhsize$ is the number of persons in the family unit, $\alpha = 0.5$. Each observation is given a weight, $indwgt = hhsize * hweight$, to obtain distributions of income across individuals (Cowell 1984) (Danziger and Taussig 1979). The variable $hweight$, is the family unit sample weight.

For calculating distributions for different years and in dollars the following data from the IMF Year Book 1994 were used.

	1981	1983
Price level consumption		
Germany	106.3	115.6
Sweden	112.1	132.6
Dollar exchange rate		
Germany	2.26	2.553
Sweden	5.063	7.667