Statistical Inference for Welfare under Complete and Incomplete Information

by

Frank Cowell
and
Maria-Pia Victoria-Feser

Discussion Paper
No. DARP 47
December 1999

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
Tel: 020-7405-7686

ABSTRACT

We show how a collection of results in the literature on the empirical estimation of welfare indicators from sample data can be unified. We also demonstrate how some of these ideas can be extended to empirically important cases where the data have been trimmed or censored.

Keywords: inequality measurement; income distribution; Lorenz curve; influence function, sampling variance; censoring; trimming

JEL classification: C13; D63

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

Maria-Pia Victoria-Feser is at to the Université de Genéve, Faculty of Psychology and Education, University of Geneva, 40, Bd du Pont d'Arve, CH-1211 Geneva 4
Tel: + 41 22 705 9104

Partially supported by the Human Capital and Mobility Programme of the EU grant #ERBCHRXCT94067 and ESRC grant reference R000237324: Robust Methods for Comparisons of Income Distribution.

---

1 Introduction

In income-distribution analysis data-sets incorporating different types of information are routinely encountered but the problems that these differences create are not always appropriately recognised in statistical applications. For the case of complete information – where a sample representative of the whole underlying distribution is available – there is a large collection of scattered results, although it is arguable that a unifying framework is lacking. For incomplete information – where the data have been "tampered with" in some way before the researcher gets them – the position is less satisfactory: comparatively little has been written to clarify the way standard statistical analysis should be modified in the light of the tampering, still less to turn the analysis into practical algorithms. The purpose of this paper is to show, first, that an elegant unified framework can be applied to the case of complete information and, second, that the same framework can be used to provide straightforward formulas for use with incomplete data.

Three basic types of incomplete information are commonly distinguished:

1. *Censored data*. This case refers to the situation in which, for reasons of confidentiality or practicality, some of the data in the lower and/or upper tail have been set to given values respectively $\underline{y}$ and $\overline{y}$. Typically the distribution outside those limits is reported as a point mass at the limits themselves. In some cases other statistics are available such as the mean of the censored parts of the distribution.

2. *Truncated data*. This is similar to the previous case, but one has no information about the "missing" data outside the $(\underline{y}, \overline{y})$ boundaries.

3. *Trimmed data*. Here a proportion of the data that are originally observed in the whole range of definition of income is removed for reasons such as robust estimation (Angrist and Krueger 1999, Cowell and Victoria-Feser 1998, Gottschalk and Moffitt 1995).

The various types of problem covered by this brief review raise different issues for the analysis of income-distribution statistics. We will show how these can be incorporated within our unified approach. The paper is organised as follows: section 2 introduces the basic definitions and techniques; section 3 handles the case of complete information and shows how the tools of section 2 can throw light on the underlying structure of the inference problem; section 4 shows how one may build upon these results to handle the problems of incomplete data. Section 5 concludes.

2 Methodology

2.1 Terminology and definitions

Let $\mathfrak{Q} := [0,1]$ be the set of population proportions and $\mathfrak{F}$ be the set of all absolutely continuous differentiable probability distributions with finite mean and variance. Let income be represented as a continuous random variable X with distribution $F \in \mathfrak{F}$ and support on the interval $\mathfrak{X} := [\underline{x}, \bar{x}] \subseteq \Re$, where $\Re$ is the real line.

In this approach a *statistic* is a functional T defined on the space of probability distributions. For example the mean of a distribution $F \in \mathfrak{F}$ is written as the functional

$$\mu(F) = \int_{\underline{x}}^{\bar{x}} x dF(x). \tag{1}$$

Other tools required for economic appraisal of distributions and for statistical inference can be introduced once some basic concepts have been defined for complete and incomplete information.

2.1.1 Sample

In estimating the statistics used to implement welfare-economic concepts in practice one uses $F^{(n)}$, a sample of size n drawn from the distribution F; this is a distribution consisting of n point-masses $\frac{1}{n}$, one at each observation in the sample. Denote the order statistics of the sample by $\left\{x_{[i]} : i = 1, ..., n\right\}$. How these are to be implemented appropriately depends upon the nature of the sample as we discuss in sections 3 and 4.

2.1.2 A typology of incomplete information

Let us look more carefully at the issue of classifying types of data incompleteness cited in the introduction. Assume that some of the sample in the tails of the distribution has been "excluded": the data here are considered to be unreliable, contaminated or have been removed altogether. There are two separate issues:

1. What determines the boundaries of the excluded subsets of the sample space?

2. What use is made of information in the excluded part of the sample?

There are two types of answer that are relevant to the first question – selection of a subset of $\mathfrak{X}$; selection of a subset of $\mathfrak{Q}$. In the first case the income-boundaries of the excluded subset $(\underline{y}, \overline{y})$ are determinate but the proportions of the excluded subsets are random. In the second case the boundaries of the excluded sample

	Information about Excluded Sample		
	None	*proportion*	*Multiple statistics*
limits $\mathbf{y}$ *fixed,* $\boldsymbol{\alpha}$ *random*	**A**	**B**	**C**
proportions $\boldsymbol{\alpha}$ *fixed,* $\mathbf{y}$ *random*	**D**	**E**	**F**

Table 1: A typology of imperfect information

are fixed by the numbers $(\underline{\alpha}, \bar{\alpha})$, and the incomes at the boundary of the excluded samples are random.

There are several possible answers to the second question, as indicated by the columns in Table 1. This then results in six possible cases, at least four of which appear in the income-distribution literature. Case A is the standard form of truncation. B covers "censoring": in this case there are point masses at boundaries $\mathbf{y}$ corresponding to the population-share of the excluded sample; but if one voluntarily discards information about these point masses (see page 28 below) one has case A. Case C is an extension of standard estimation problem with grouped data (Gastwirth, Nayak, and Krieger 1986). D represents the case of trimming. E and F are of less immediate relevance, because trimming is usually done on a voluntary basis – for robustness reasons for example.[1] However, for any case in the second row of Table 1, because the boundaries of the excluded sample are random, some standard statistical procedures are no longer valid.

2.1.3 Basic Tools: Complete information

We require three functionals from $\mathfrak{F}\times\mathfrak{Q}$ to $\Re$. To introduce them let $q \in \mathfrak{Q}$ denote an arbitrary population proportion. Then the *quantile functional* is defined by:

$$Q(F;q) := \inf\{x|F(x) \geq q\} =: x_q \tag{2}$$

(Gastwirth 1971), and the *cumulative income functional* is defined by:

$$C(F;q) := \int_{\underline{x}}^{x_q} x dF(x) =: c_q \tag{3}$$

(Cowell and Victoria-Feser 1996c); in particular note that $C(F;1) = \mu(F)$. Analogously define:

$$S(F;q) := \int_{\underline{x}}^{x_q} x^2 dF(x) =: s_q. \tag{4}$$

[1]However, case E does have a role to play in determining the statistical properties in case D – see section 4.3 below.

Note that the sample analogues are obtained by replacing F by the empirical distribution $F^{(n)}$.

In the case of *linear* functionals it is convenient to define the following two operators, where θ_1, θ_2 are real functions of the random variable X:

$$\operatorname{cov}(\theta_1(x), \theta_2(x); F) := \int \theta_1(x)\theta_2(x)dF(x) - \int \theta_1(x)dF(x) \int \theta_2(x)dF(x) \tag{5}$$

$$\operatorname{var}(\theta_1(x); F) := \operatorname{cov}(\theta_1(x), \theta_1(x); F) \tag{6}$$

2.1.4 Basic Tools: Trimmed data

Here we assume that determinate proportions $\underline{\alpha}$ and $1 - \bar{\alpha}$ have been removed from the bottom and from the top of the distribution respectively. Let us define the total amount trimmed as $\alpha := \underline{\alpha} + (1 - \bar{\alpha})$ and the *trimming indicator function* as

$$a(u) := \begin{cases} \frac{1}{1-\alpha} & Q(F; \underline{\alpha}) < u \leq Q(F; \bar{\alpha}) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

such that

$$\int_{\underline{x}}^{\overline{x}} a(u)dF(u) = 1.$$

The trimming indicator function enables to define $\tilde{F}_\alpha$, the trimmed distribution, as

$$\tilde{F}_\alpha(x) := \begin{cases} 0 & \text{if } x < Q(F, \underline{\alpha}) \\ \frac{F(x)-\underline{\alpha}}{1-\alpha} & \text{if } Q(F, \underline{\alpha}) \leq x < Q(F, \overline{\alpha}) \\ 1 & \text{if } x \geq Q(F, \overline{\alpha}) \end{cases} . \tag{8}$$

This is to be distinguished from F^*_α, the "censored" distribution formed from F with $(x_{\underline{\alpha}}, x_{\overline{\alpha}})$ as truncation points (case E in Table 1), thus

$$F^*_\alpha(x) := \begin{cases} 0 & \text{if } x < Q(F, \underline{\alpha}) \\ F(x) & \text{if } Q(F, \underline{\alpha}) \leq x < Q(F, \overline{\alpha}) \\ 1 & \text{if } x \geq Q(F, \overline{\alpha}) \end{cases} . \tag{9}$$

Using (8) the α-trimmed counterpart of (1) is given by

$$\mu_\alpha := \mu(\tilde{F}_\alpha) = \int_{\underline{x}}^{\overline{x}} a(u)udF(u). \tag{10}$$

Using (2) the α-trimmed income cumulations are given by

$$c_{\alpha,q} := C(\tilde{F}_\alpha; q) = \int_{\underline{x}}^{x_q} a(u) u dF(u) = \frac{1}{1-\alpha} \int_{x_{\underline{\alpha}}}^{x_q} u dF(u) \tag{11}$$

where $x_{\underline{\alpha}} := Q(F; \underline{\alpha})$. We also define

$$s_{\alpha,q} := S(\tilde{F}_\alpha; q) := \int_{\underline{x}}^{x_q} a(u) u^2 dF(u) = \frac{1}{1-\alpha} \int_{x_{\underline{\alpha}}}^{x_q} u^2 dF(u). \tag{12}$$

Once again, the sample analogues of these quantities are obtained by replacing F by the empirical distribution $F^{(n)}$. Let $\text{int}\,(z)$ be the largest integer less than or equal to z, and let

$$\kappa(n, q) := \text{int}[(n-1)q + 1] \tag{13}$$

denote the order of the observation corresponding to the quantile q. Then, given a sample $x_{[1]}, \ldots, x_{[n]}$ of (untrimmed) ordered observations, we have:

$$\hat{c}_{\alpha,q} = \frac{1}{1-\alpha} \int_{\inf\{x|F^{(n)}(x) > \underline{\alpha}\}}^{\inf\{x|F^{(n)}(x) \geq q\}} u dF^{(n)}(u) = \frac{1}{\kappa(n,\alpha)} \sum_{i=\kappa(n,\underline{\alpha})+1}^{\kappa(n,q)} x_{[i]}$$

2.2 Economic Tools

2.2.1 Welfare ranking

The functionals defined in section 2.1.3 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.

For example, using (2), for a given $F \in \mathfrak{F}$, the graph $\{q, Q(F, q)\}$ describes Pen's parade that forms the basis for first-order distributional dominance results. Furthermore the functional (3) is used to define the following standard concepts. For a given $F \in \mathfrak{F}$, the graph $\{q, C(F, q)\}$ describes the *generalised Lorenz curve* (GLC), the basis for second-order distributional dominance results (Shorrocks 1983). The scale normalisation of the GLC by the mean (1) gives the (relative) Lorenz functional:

$$L(F; q) := \frac{C(F; q)}{\mu(F)} \tag{14}$$

and the graph $\{q, L(F; q)\}$ gives the *relative Lorenz curve* (RLC).[2] An alternative normalisation of the GLC yields the absolute counterpart to (14)

$$A(F; q) := C(F; q) - q\mu(F) \tag{15}$$

and the graph $\{q, A(F; q)\}$ is the *absolute Lorenz Curve* (ALC) (Moyes 1987).

[2]Beach and Davidson (1983) use a different, related concept to underpin these expressions, the first moment function $\Phi : \mathfrak{X} \mapsto [0, 1]$ given by $\Phi(x) = L(F; x) = \frac{1}{\mu(F)} \int^x y dF(y)$ (Kendall and Stuart 1977).

2.2.2 Welfare indices

The term "welfare indices" is used here to cover a number of specific tools of distributional analysis such as social-welfare functions, inequality measures and poverty indices. Many of the welfare indices that are commonly used can be expressed in the following quasi-additively decomposable form

$$W_{\mathrm{QAD}}(F) := \int \varphi\left(x, \mu(F)\right) dF(x) \tag{16}$$

where $\varphi : \mathfrak{X} \times \mathfrak{X} \to \Re$ is piecewise differentiable. For example, almost all commonly-used inequality indices other than the Gini can be written in the form

$$\Omega\left(W_{\mathrm{QAD}}(F), \mu(F)\right) \tag{17}$$

where Ω is continuous and monotonic increasing in its first argument. In fact a substantial proportion of the applied welfare-economic literature uses the more restrictive form of welfare index found by replacing W_{QAD} in (16) by the additively decomposable form

$$W_{\mathrm{AD}}(F) := \int \phi\left(x\right) dF(x). \tag{18}$$

Most of the welfare indices that cannot be put in the form (16) can however be expressed in the explicitly rank-dependent form

$$W_{\mathrm{RD}}(F) := \int \psi\left(x, \mu(F), F(x)\right) dF(x) \tag{19}$$

where ψ is piecewise differentiable. The class (19) encompasses the Gini coefficient and the Sen poverty index (Sen 1976).

2.3 Statistical techniques

2.3.1 The influence function

The principal analytical tool employed here is the *influence function* (IF).[3] The primary usage of the IF is to characterise the sensitivity of a statistic to point contamination in the data (Hampel 1971, 1974; Hampel et al. 1986). So, assume that T is a functional $\mathfrak{F} \to \Re^m$, that $F \in \mathfrak{F}$ is an arbitrary distribution, that $H^{(z)} \in \mathfrak{F}$ is a degenerate distribution that consists of a single point mass at z and that $\delta \in \mathfrak{Q}$. The mixture distribution

$$G := [1 - \delta]F + \delta H^{(z)} \tag{20}$$

[3] Also called the influence curve

can be taken as a representation of contamination of a "true" distribution (F) by the point mass, where δ represents the relative size or importance of the contamination. The IF measures the impact of the contamination on the statistic T for infinitesimal δ, namely

$$\mathrm{IF}(z;T,F) := \lim_{\delta \downarrow 0} \left[\frac{T(G) - T(F)}{\delta} \right] \tag{21}$$

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

The IF's relevance to the present analysis is that it may be used to derive asymptotic results such as asymptotic covariance matrices. Again let the distribution G be "near" F; then the first-order von-Mises expansion of T at F evaluated in G is given by

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

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

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

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

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

$$\int \mathrm{IF}(z;T,F)\mathrm{IF}^{T}(z;T,F) dF(z) \tag{22}$$

Regularity conditions can be found in Reeds (1976), Boos and Serfling (1980) and Fernholz (1983). It should be stressed that it is usually not easy to prove the asymptotic normality by proving that the remainder is negligible, other means are easier. The IF is used here as a simple tool to derive the asymptotic covariances, and for the asymptotic normality of the statistics considered in this paper, one can refer to e.g. Hoeffding (1948) Moore (1968), Shorack (1972), Stigler (1974), Goldie (1977), Serfling (1980).

2.3.2 Background results

Several useful results on IFs applied to income-distribution analysis are available from previous work in different contexts. In particular we have two key properties for the fundamental functionals with complete information (Cowell and Victoria-Feser 1996c). Let f be the density function for the distribution function F, then::

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

$$\begin{aligned} \mathrm{IF}(z; Q(\cdot, q), F) &= \frac{q - \iota(Q(F;q) \geq z)}{f(Q(F;q))} \\ &= \frac{q - \iota(x_q \geq z)}{f(x_q)}. \end{aligned} \tag{23}$$

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

$$\begin{aligned} \mathrm{IF}(z; C(\cdot, q), F) &= qQ(F;q) - C(F;q) + \iota(q \geq F(z))[z - Q(F;q)] \\ &= qx_q - c_q + \iota(x_q \geq z)[z - x_q]. \end{aligned} \tag{24}$$

where $\iota(\cdot)$ is the indicator function giving $\iota(D) = 1$ if D is true and 0 otherwise.

3 Complete Information

Our baseline situation is that which is implicitly assumed in the bulk of the literature – that where information is available about the whole distribution. There is no truncation or censoring of the sample. Although there is a collection of results for this (Cowell 1999, 2000) they have not previously been placed within a single coherent framework.

3.1 Ranking criteria

In order to implement ranking criteria empirically a standard approach is as follows. Choose a finite collection of population proportions $\Theta \subset \mathfrak{Q}$; then for each $q \in \Theta$ one can compute the sample quantiles and cumulants required for empirical implementation of first- and second-order dominance relations.

3.1.1 First order

The sample quantiles are given by

$$\hat{x}_q := Q\left(F^{(n)}; q\right) = x_{[\kappa(n,q)]} \tag{25}$$

We then have the standard first-order result:

Theorem 1 *For any $q, q' \in \Theta$ such that $q \leq q'$ the asymptotic covariance of $\sqrt{n}\hat{x}_q$ and $\sqrt{n}\hat{x}_{q'}$ is*

$$\frac{q\,[1 - q']}{f(x_q) f(x_{q'})} \tag{26}$$

Proof. First note that $\iota(x_q \geq z)\iota(x_{q'} \geq z) = \iota(x_q \geq z)$ and $\int \iota(x_q \geq z)dF(z) = q$. Then, using Lemmas 1 and 2 we have

$$\begin{aligned} & \int \mathrm{IF}(z; Q(F;q), F)\mathrm{IF}(z; Q(F;q'), F)dF(z) \\ = & \frac{1}{f(x_q)f(x_{q'})} \int \left[qq' + [1-q']\,\iota(x_q \geq z) - q\iota(x_{q'} \geq z)\right] dF(z) \\ = & \frac{qq' + q\,[1-q'] - q'q}{f(x_q)f(x_{q'})}. \end{aligned}$$

■

The obvious difficulty with this is that in order to implement it one needs a suitable estimator for f.

3.1.2 Second order

We now use the influence function approach to develop the asymptotic covariance matrix of the sample cumulative income functional

$$\hat{c}_q := C\left(F^{(n)}; q\right) = \frac{1}{n}\sum_{i=1}^{\kappa(n,q)} x_{[i]} \tag{27}$$

which corresponds to the sample quantiles (25) and which is in turn used to make statistical inference using Lorenz curves. The set of pairs $\{(q, \hat{c}_q) : q \in \Theta\}$ gives points on the empirical generalised Lorenz curve, and $\hat{c}_1$ is the sample mean $\mu\left(F^{(n)}\right)$; the relative and absolute Lorenz curves are found by normalisation as in (14) and (15).

Following Beach and Davidson (1983) assume that the underlying distribution satisfies $F \in \mathfrak{F}$. Then we have:

Theorem 2 [4] *The set of sample income cumulations $\{\hat{c}_q : q \in \Theta\}$ is asymptotically multivariate normal. For any $q, q' \in \Theta$ such that $q \leq q'$ the asymptotic covariance of $\sqrt{n}\hat{c}_q$ and $\sqrt{n}\hat{c}_{q'}$ is*

$$\omega_{qq'} := s_q + [qx_q - c_q]\,[x_{q'} - q'x_{q'} + c_{q'}] - x_q c_q \tag{28}$$

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

$$\int \mathrm{IF}(z; C(F;q), F)\mathrm{IF}(z; C(F;q'), F)dF(z) = \int [qx_q - c_q + \iota(x_q \geq z)[z - x_q]]$$
$$[q'x_{q'} - c_{q'} + \iota(x_{q'} \geq z)[z - x_{q'}]]\,dF(z) \tag{29}$$

[4] This is a slightly modified version of Theorem 1 in Beach and Davidson (1983).

Given that $\iota(x_{q'} \geq z) = 1$ whenever $\iota(x_q \geq z) = 1$ the right-hand side of (29) becomes

$$[qx_q - c_q]\,[q'x_{q'} - c_{q'}] + \int_{\underline{x}}^{x_{q'}} [qx_q - c_q]\,[z - x_{q'}]dF(z)$$
$$+ \int_{\underline{x}}^{x_q} [q'x_{q'} - c_{q'} + z - x_{q'}]\,[z - x_q]dF(z) \tag{30}$$

Using the definitions in (2)-(4) of section 2.1.3 we find that (30) becomes

$$[qx_q - c_q]\,[x_{q'}(1 - q') + c_{q'}] - [x_q c_q - s_q] = \omega_{qq'} \tag{31}$$

■

This result can be implemented empirically by substituting in (28) the sample concepts $\hat{x}_q$, $\hat{c}_q$ and

$$\hat{s}_q := S(F^{(n)}; q) = \frac{1}{n} \sum_{i=1}^{\kappa(n,q)} x_{[i]}^2$$

obtained from (25), (27) and (4). It can also be used for the case of RLC and ALC. For the former, we first note that

$$L(F; q) = \frac{c_q}{\mu} = \frac{c_q}{c_1}$$

and then, using the standard result on limiting distributions of differentiable functions of random variables (Rao 1973), the asymptotic covariances of $(\sqrt{n})$ the RLC ordinates are given by

$$v_{qq'} = \frac{1}{\mu^4} \left[\mu^2 \omega_{qq'} + c_q c_{q'} \omega_{11} - \mu c_q \omega_{q'1} - \mu c_{q'} \omega_{q1}\right]. \tag{32}$$

3.2 QAD Welfare indices

Here we restrict ourselves to indices that are quasi-additively decomposable: the important case that lies outside this class is considered in section 3.3.

3.2.1 General result

Given a sample $x_1, \ldots, x_n$, the sample analogues of W_{QAD} defined in (16) are given by

$$\hat{w}_{\text{QAD}} := W_{\text{QAD}}(F^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(x_i, \frac{1}{n} \sum x_i\right) \tag{33}$$

The principal result for this class of indices is the following. Let φ_2 denote the partial differential of φ with respect to its second argument and

$$\Phi(F) := \int_{\underline{x}}^{\overline{x}} \varphi_2\left(x, \mu(F)\right) dF(x)$$

Theorem 3 *The asymptotic variance of* W_{QAD} *is*

$$\mathrm{var}(\varphi\left(x, \mu(F)\right); F) + 2\Phi(F)\mathrm{cov}(x, \varphi\left(x, \mu(F)\right); F) + \Phi(F)^2\mathrm{var}(x; F) \tag{34}$$

Proof. Substituting the mixture distribution (20) into (16), differentiating with respect to δ and evaluating at $\delta = 0$ we get

$$\begin{aligned} \mathrm{IF}(z; W_{\mathrm{QAD}}, F) &= \varphi\left(z, \mu(F)\right) - \int_{\underline{x}}^{\overline{x}} \varphi\left(x, \mu(F)\right) dF(x) \\ &\quad + \left[z - \mu(F)\right] \int_{\underline{x}}^{\overline{x}} \frac{\partial \varphi\left(x, \mu(F)\right)}{\partial \mu(F)} dF(x) \\ &= \varphi\left(z, \mu(F)\right) - W_{\mathrm{QAD}}(F) + \left[z - \mu(F)\right] \Phi(F) \end{aligned} \tag{35}$$

Using (1), the asymptotic variance is given by

$$\int_{\underline{x}}^{\overline{x}} \mathrm{IF}(z; W_{\mathrm{QAD}}, F)^2 dF(z) \tag{36}$$

Substituting from (35) into (36) we get (34)

■

The asymptotic variance given in (34) can be easily estimated by replacing F by the empirical distribution.

3.2.2 Examples: inequality and poverty measures

Take the *generalised entropy* family of inequality measures:

$$I^{\theta}_{\mathrm{GE}}(F) = \frac{1}{\theta^2 - \theta} \left[\int_{\underline{x}}^{\overline{x}} \left[\frac{x}{\mu(F)} \right]^{\theta} dF(x) - 1 \right]$$

which belongs to the more restrictive decomposable class of indices generated by (18). For convenience take the transformed statistic

$$W(F) = 1 + \left[\theta^2 - \theta\right] I^{\theta}_{\mathrm{GE}}(F) = \frac{1}{\mu(F)^{\theta}} \int_{\underline{x}}^{\overline{x}} x^{\theta} dF(x).$$

We have

$$\begin{aligned}\varphi\left(x,\mu(F)\right) &= \left[\frac{x}{\mu(F)}\right]^{\theta} \\ \varphi_2\left(x,\mu(F)\right) &= \frac{1}{\theta-1}\left[\frac{x^{\theta}}{\mu(F)^{\theta+1}}\right] = \theta\frac{\varphi\left(x,\mu(F)\right)}{\mu(F)}\end{aligned}$$

so that the asymptotic variance of $W(F)$ is

$$\frac{\text{var}(x^{\theta};F) + 2\gamma(F)\text{cov}(x,x^{\theta};F) + \gamma(F)^2\text{var}(x;F)}{\mu(F)^{2\theta}} \tag{37}$$

where

$$\gamma(F) := \frac{\theta}{\mu(F)}\int_{\underline{x}}^{\overline{x}} x^{\theta}dF(x),$$

which corresponds to the results in Cowell (1989).

As a second example consider the mean deviation[5], an index which does not belong to the class of indices generated by (18), but does belong to (17).

$$T_{\text{MD}}(F) := \int |x-\mu(F)|\,dF(x)$$

In this case we have

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

where $\iota_z := \iota(z \geq \mu(F))$ and $\bar{q} := F(\mu)$. So the asymptotic variance of T_{MD} is

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

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

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

(Gastwirth 1974)

Finally consider the broadly defined class of poverty measures given by

$$P(F) := \int p\left(x, \zeta\left(F\right)\right) dF(x)$$

where $\zeta(F)$ is the poverty line (an exogenous poverty line can be obtained as the special case where ζ is a constant functional) and p is a poverty evaluation function that is nonincreasing in x and takes the value zero for $x \geq \zeta(F)$. We have

$$\begin{aligned} \mathrm{IF}(z; P, F) &= p\left(z, \zeta(F)\right) - P(F) \\ &\quad + P_\zeta(F)\mathrm{IF}(z; \zeta, F) \end{aligned} \tag{38}$$

where

$$P_\zeta(F) := \int \frac{\partial p\left(x, Z\right)}{\partial Z} dF(x)$$

is the impact on measured poverty of a small change in the poverty line (Cowell and Victoria-Feser 1996a). It is clear from (38) that the form for the asymptotic variance of the poverty index will depend on the precise way in which the poverty line depends on the income distribution. Let us take a form that covers many commonly encountered situations, i.e. where $\zeta(F)$ is a function of the mean of the distribution:

$$\zeta(F) = \zeta_0 + \beta\mu(F)$$

where ζ_0 and β are non-negative constants. Then we have

$$\mathrm{IF}(z; \zeta, F) = \beta\mathrm{IF}(z; \mu, F) = \beta\left[z - \mu(F)\right]$$

and the asymptotic variance of P is

$$\begin{aligned} &\int p\left(z, \zeta\left(F\right)\right)^2 dF(z) - P(F)^2 \\ &+2\beta P_\zeta(F)\mathrm{cov}(p(x), x; F) \\ &+\beta^2 P_\zeta^2(F)\mathrm{var}(x; F) \end{aligned} \tag{39}$$

where $\beta = 0$ in (39) gives the case of the exogenous poverty line. It is clear that in order to estimate the asymptotic variance of P, one needs information on the whole distribution. In other situations, especially for robustness reasons (Cowell and Victoria-Feser 1996a), it is better to consider $\zeta(F)$ as a function of a quantile of the distribution, i.e.

$$\zeta(F) = F^{-1}(q) = \zeta_0 + \beta x_q$$

$q \in \mathfrak{Q}$. In this case, using Lemma 2, the asymptotic variance of P is

$$\begin{aligned}\mathrm{IF}(z;P,F) &= \int p\left(z,\zeta(F)\right)^2 dF(z) - P(F)^2 \\ &+\beta\frac{2q}{f(x_q)}P_\zeta(F)P(F) - \beta\frac{2}{f(x_q)}P_\zeta(F)\int_{\underline{x}}^{x_q} p\left(z,\zeta(F)\right) dF(z) \\ &+\beta^2 P_\zeta(F)^2\frac{q(1-q)}{f(x_q)^2}\end{aligned}$$

Note that in most cases $\int_{\underline{x}}^{x_q} p\left(z,\zeta(F)\right) dF(z) = P(F)$ because $\zeta_0 + \beta x_q \leq x_q$, so that the asymptotic variance of P simplifies to

$$\begin{aligned}\mathrm{IF}(z;P,F) &= \int p\left(z,\zeta(F)\right)^2 dF(z) - P(F)^2 \\ &-2\beta\frac{(1-q)}{f(x_q)}P_\zeta(F)P(F) \\ &+\beta^2 P_\zeta(F)^2\frac{q(1-q)}{f(x_q)^2}\end{aligned}$$

It is estimable if one can get a density estimate at x_q.

3.3 The Gini coefficient

The general form (19) is somewhat cumbersome to work with. However, we can fairly easily derive results for the most important member of this class, namely the Gini coefficient. There are several equivalent forms of this index, but the most useful here is to define

$$I_{\mathrm{Gini}}(F) = 1 - 2\int_0^1 \frac{C(F;q)}{C(F;1)}dq \tag{40}$$

The representation in terms of the C-functional makes it easy to use the result of Theorem 2. We now have:

Theorem 4 *The asymptotic variance of $\sqrt{n}I_{\mathrm{Gini}}(F^{(n)})$ is given by $4\vartheta/\mu^4$ where*

$$\begin{aligned}\vartheta = \mu^2\int_0^1\int_0^q \omega_{q'q}dq'dq + \mu^2\int_0^1\int_q^1 \omega_{qq'}dq'dq + \\ \omega_{11}\left[\int_0^1 c_q dq\right]^2 - 2\mu\int_0^1 c_q dq\int_0^1 \omega_{q1}dq\end{aligned} \tag{41}$$

Proof. Using (32) and (40) it is clear that the asymptotic covariance of $\sqrt{n}I_{\mathrm{Gini}}(F^{(n)})$ is $4\vartheta/\mu^4$ where

$$\vartheta := \mu^4\int_0^1\int_0^1 v_{qq'}dq'dq \tag{42}$$

Expanding (42) and using (32) we get:

$$\begin{aligned}\vartheta &= \mu^4 \int_0^1 \int_0^q v_{q'q} dq' dq + \mu^4 \int_0^1 \int_q^1 v_{qq'} dq' dq \\ &= \int_0^1 \int_0^q \left[\mu^2 \omega_{q'q} + c_q c_{q'} \omega_{11} - \mu c_q \omega_{q'1} - \mu c_{q'} \omega_{q1}\right] dq' dq \\ &\quad + \int_0^1 \int_q^1 \left[\mu^2 \omega_{qq'} + c_q c_{q'} \omega_{11} - \mu c_q \omega_{q'1} - \mu c_{q'} \omega_{q1}\right] dq' dq \qquad (43a)\end{aligned}$$

which implies

$$\begin{aligned}\vartheta &= \mu^2 \int_0^1 \int_0^q \omega_{q'q} dq' dq + \mu^2 \int_0^1 \int_q^1 \omega_{qq'} dq' dq + \\ &\quad \omega_{11} \left[\int_0^1 c_q dq\right]^2 - 2\mu \int_0^1 c_q dq \int_0^1 \omega_{q1} dq \qquad (44)\end{aligned}$$

■

The expansion of (41) is given in the Appendix. The estimates of ϑ are easily obtained by making use of (46) leading to

$$\begin{aligned}\widehat{\vartheta} &= \widehat{\mu}^2 \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i} \widehat{\omega}_{q_j q_i} + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \widehat{\omega}_{q_i q_j}\right) + \\ &\quad \widehat{\omega}_{q_n q_n} \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i} x_{[j]}\right]^2 - 2\widehat{\mu} \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i} x_{[j]}\right) \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{\omega}_{q_i q_n}\right) \qquad (45)\end{aligned}$$

where $\widehat{\mu}$ is the sample mean. A $100\,(1-\beta)\%$ confidence interval for the Gini coefficient is then given by

$$\left(1 - \frac{2}{\widehat{\mu} n^2} \sum_{i=1}^{n} \sum_{j=1}^{i} x_{[j]} \mp z_{\beta/2} \sqrt{\frac{4\widehat{\vartheta}}{n\widehat{\mu}^4}}\right).$$

Note that one can also choose a number of proportions $k < n$ and change the indices i and j in (45) to respectively $\kappa(n, q_i)$ and $\kappa(n, q_j)$.

3.4 Empirical example 1

The results presented in this section can be used to compute confidence intervals for the points in a Lorenz curve or in a relative Lorenz curve. It should be stressed that when computing simultaneous confidence intervals, a correction should be made such that the so-called familywise error rate is kept constant. With Lorenz

curves, the points and hence the confidence intervals are numerous, so that for example a Bonferoni type correction would be too drastic a method. The search for an efficient correction procedure is not the subject of this paper and therefore no correction will be made here.

Given a sample $\left\{x_{[1]}, \ldots, x_{[n]}\right\}$ of ordered data we define the set of proportions $\Theta = \left\{q_i = \frac{i}{n} : i = 1, n\right\}$, then (28) is estimated by

$$\widehat{\omega}_{q_i q_j} = \frac{1}{n}\sum_{k=1}^{i} x_{[k]}^2 + \left[q_i x_{[i]} - \frac{1}{n}\sum_{k=1}^{i} x_{[k]}\right]\left[x_{[j]} - q_j x_{[j]} + \frac{1}{n}\sum_{k=1}^{j} x_{[k]}\right] \quad (46)$$

$$-x_{[i]}\frac{1}{n}\sum_{k=1}^{i} x_{[k]} \quad (47)$$

A $100\,(1-\beta)\%$ confidence interval for the ith Lorenz ordinate is then

$$\left(\frac{1}{n}\sum_{k=1}^{i} x_{[k]} \mp z_{\beta/2}\sqrt{\widehat{\omega}_{q_i q_i}/n}\right)$$

where $z_{\beta/2}$ is the $(1-\beta/2)$ quantile of the standard normal distribution. The calculations for the relative Lorenz ordinates are obtained in a straightforward manner using (46) and μ being estimated by the sample mean. One can also choose a number of proportions $k < n$ and in that case, the indices i and j in (46) are simply replaced by respectively $\kappa(n, q_i)$ and $\kappa(n, q_j)$.

The data we consider here come from the *Encuesta de Presupuestos Familiares* (EPF) in Spain for the period 1990-1991. The EPF is a continuous survey whose effective sample consists of 21,155 households and primarily aimed at the elaboration of the weights for the Retail Index Price Instituto Nacional de Estadistica (1992). The data are household incomes in thousand of pesetas (monetary, non-monetary and extraordinary income) standardized by the Oxford equivalence scale (Instituto Nacional de Estadistica 1992, p.34). Because the original sample size is very large, we chose a random subsample of size 500.

Figure 1 shows the Lorenz curve and the relative Lorenz curve as well as 95% confidence intervals for the Spanish data.

4 Incomplete information

4.1 A strategy

Recall, from the introduction and section 2.1.2 that the various types of incomplete information differ in terms of the way the excluded subset of the sample is demarcated and in the usage, if any, of information from the excluded subset. We will focus here on the first of these two issues and defer consideration of the

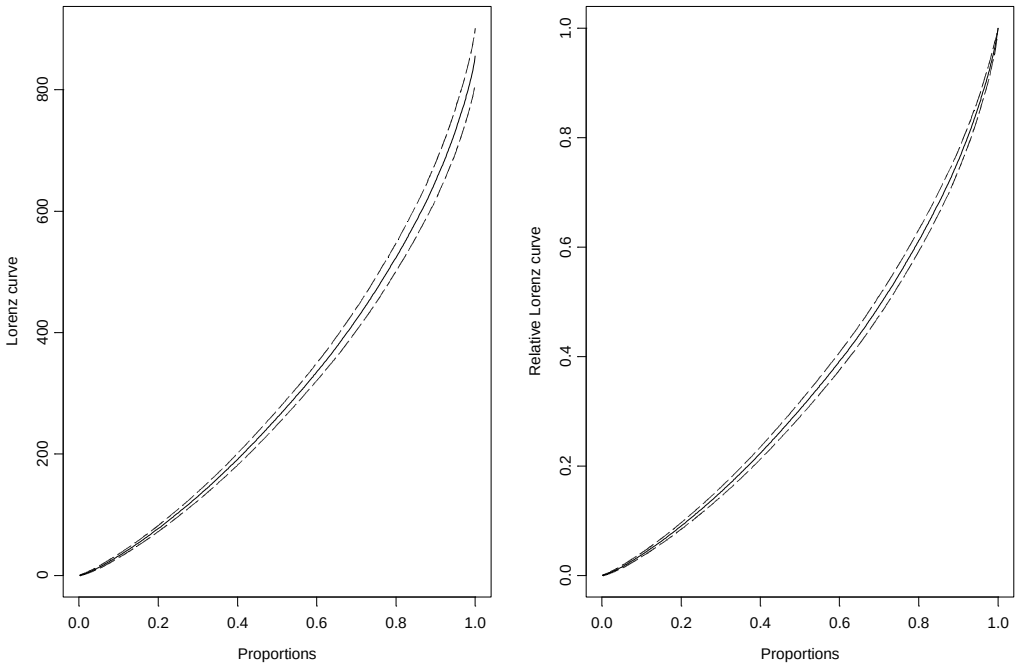

Figure 1: Lorenz curve and relative Lorenz curve with 95% confidence intervals for the Spanish data

further questions raised by the second issue until section 4.6: for the moment we suppose that no use is to be made of the excluded subsample and no attempt is made to model the distribution in the corresponding part of the population.

The cases of censored data and truncated data (cases A and B in Table 1) take $\underline{y}$ and $\overline{y}$ to be fixed for any sample. As long as no use is made of the observed number of these values (i.e. as long as B is treated like A), inference with in these cases is like inference in the complete information case with a redefined population: the limits of the support $(\underline{x}, \bar{x})$ are replaced by the narrower limits $(\underline{y}, \overline{y})$. However, in the case of trimmed data a fixed quantity of the sample is discarded or to be unobservable, as in the scheme outlined in section 2.1.4. If $(x_{\underline{\alpha}}, x_{\overline{\alpha}})$ denotes the range of the trimmed sample values, then $x_{\underline{\alpha}}$ and $x_{\overline{\alpha}}$ are *random* – see the discussion on page 2. This class of cases requires explicit modification of the theory.

The trimming phenomenon thus forms the principal alternative paradigm for our paper. Inference is carried out on the whole distribution conditional on the fact that a known proportion (in the tails) has been trimmed away. Given that the integration of $\mathrm{IF} \cdot \mathrm{IF}^T$ is required over the full distribution to derive the

asymptotic covariance matrix of functionals, this might appear to add a new difficulty, namely that there might be components of the asymptotic variance that cannot be estimated. If so, this would place a limitation on the applicability of nonparametric techniques because of the lack of information on the structure of the trimmed data. However, we will show that this supposition is groundless.

4.2 Lorenz criteria

The concepts that we need are the income cumulation $C(\tilde{F}_\alpha; q)$ for the trimmed distribution and its empirical counterpart. First we establish a key result that links the trimmed and untrimmed cases.

Lemma 4 *The* IF *for the cumulative income functional with trimmed data is*[6]*:*

$$\mathrm{IF}(z; C(\cdot; q), \tilde{F}_\alpha) = \frac{\mathrm{IF}(z; C(\cdot; q), F) - \mathrm{IF}(z; C(\cdot; \underline{\alpha}), F)}{1-\alpha} \tag{48}$$

$$= -c_{\alpha,q} + \frac{1}{1-\alpha}\left[q x_q - \underline{\alpha} x_{\underline{\alpha}} + \iota(x_q \geq z)[z - x_q] - \iota(x_{\underline{\alpha}} \geq z)[z - x_{\underline{\alpha}}]\right] \tag{49}$$

Proof. It is clear from (3) and (11) that the income cumulations based on the ordinary and trimmed distributions are related thus

$$C(\tilde{F}_\alpha; q) = \frac{C(F; q) - C(F; \underline{\alpha})}{1-\alpha}. \tag{50}$$

Given that $\underline{\alpha}$ and α are determinate in this case and that the relationship among the statistics in (50) is linear, (48) follows from the definition of the influence function. Using lemma 3 equation (49) follows immediately. ■

Given a sample $x_{[1]}, \ldots, x_{[n]}$ of ordered (untrimmed) data, the empirical income cumulation for the trimmed case is given by

$$\hat{c}_{\alpha,q} := C(\tilde{F}_\alpha^{(n)}; q) = \frac{1}{\kappa(n, 1-\alpha)} \sum_{i=1}^{\kappa(n,q)} x_{[i]} \iota(i > \kappa(n, \underline{\alpha}) + 1) \tag{51}$$

Assume that the set of population proportions satisfies $\Theta \subset [\underline{\alpha}, \bar{\alpha}]$. Then Lemma 4 enables to state the main trimming result which is as follows.

[6]Note that at $q = \overline{\alpha}$ for $\underline{\alpha} = \overline{\alpha}$ one gets the traditional trimmed mean which generalises the median as a robust estimator of location. One can verify that in that case (49) is equal to formula 2.18 in Huber (1981).

Theorem 5 *Given an original untrimmed sample of size n and lower and upper trimming proportions* $\underline{\alpha}, \bar{\alpha} \in \mathfrak{Q}$*, for any* $q, q' \in \Theta$ *such that* $q \leq q'$ *the asymptotic covariance of* $\sqrt{n}\hat{c}_{\alpha,q}$ *and* $\sqrt{n}\hat{c}_{\alpha,q'}$ *is given by* $\varpi_{qq'}/(1-\alpha)^2$ *where*

$$\begin{aligned}\varpi_{qq'} := [qx_q - \underline{\alpha}x_{\underline{\alpha}} - [1-\alpha]\, c_{\alpha,q}]\,[[1-q']\, x_{q'} - [1-\underline{\alpha}]\, x_{\underline{\alpha}} + [1-\alpha]\, c_{\alpha,q'}] - \\ [x_q\,[1-\alpha]\, c_{\alpha,q} - [1-\alpha]\, s_{\alpha,q}] + x_{\underline{\alpha}}\,[qx_q - \underline{\alpha}x_q - [1-\alpha]\, c_{\alpha,q}] \quad (52)\\ = \omega_{qq'} - [\omega_{\underline{\alpha}q} + \omega_{\underline{\alpha}q'}] + \omega_{\underline{\alpha}\underline{\alpha}} \quad (53)\end{aligned}$$

where $\omega_{qq'}$ *is defined in (28).*

Proof. Using Lemma 1 and (48) $(1-\alpha)^2\mathrm{cov}\left(C(\tilde{F}_\alpha^{(n)}; q), C(\tilde{F}_\alpha^{(n)}; q')\right)$ is given by

$$\int [\mathrm{IF}(z; C(\cdot; q), F) - \mathrm{IF}(z; C(\cdot; \underline{\alpha}), F)]\,[\mathrm{IF}(z; C(\cdot; q'), F) - \mathrm{IF}(z; C(\cdot; \underline{\alpha}), F)]\, dF(z) \quad (54)$$

but, expanding the integrand in (52) and using (29)-(31) it is clear that (54) becomes:

$$\omega_{qq'} - \omega_{\underline{\alpha}q} - \omega_{\underline{\alpha}q'} + \omega_{\underline{\alpha}\underline{\alpha}} \quad (55)$$

On simplifying (55) we get:

$$\begin{aligned}&s_q + [qx_q - c_q]\,[x_{q'} - q'x_{q'} + c_{q'}] - x_q c_q\\ &-s_{\underline{\alpha}} - [\underline{\alpha}x_{\underline{\alpha}} - c_{\underline{\alpha}}]\,[x_q - qx_q + c_q] + x_{\underline{\alpha}}c_{\underline{\alpha}}\\ &-s_{\underline{\alpha}} - [\underline{\alpha}x_{\underline{\alpha}} - c_{\underline{\alpha}}]\,[x_{q'} - q'x_{q'} + c_{q'}] + x_{\underline{\alpha}}c_{\underline{\alpha}}\\ &+s_{\underline{\alpha}} + [\underline{\alpha}x_{\underline{\alpha}} - c_{\underline{\alpha}}]\,[x_{\underline{\alpha}} - \underline{\alpha}x_{\underline{\alpha}} + c_{\underline{\alpha}}] - x_{\underline{\alpha}}c_{\underline{\alpha}}\end{aligned}$$

and so, using the definitions of $c_{\alpha,q}$ and $s_{\alpha,q}$ in section 2.1.4 we get (52). ■

This result can be used for the RLC case. Indeed, using the standard result on limiting distributions of differentiable functions of random variables (Rao 1973), the asymptotic covariances of $(\sqrt{n})$ the RLC ordinates are then given by

$$\upsilon_{qq',\alpha} = \frac{1}{(1-\alpha)^2\mu_\alpha^4}\left[\mu_\alpha^2\varpi_{qq'} + c_{\alpha,q}c_{\alpha,q'}\varpi_{\overline{\alpha}\overline{\alpha}} - \mu_\alpha c_{\alpha,q}\varpi_{q'\overline{\alpha}} - \mu_\alpha c_{\alpha,q'}\varpi_{q\overline{\alpha}}\right]\ . \quad (56)$$

4.3 QAD Welfare indices

Trimming data means that the information on the trimmed part is ignored and welfare indices are computed in the usual way but on the trimmed sample. This

means that the trimmed version of (16) becomes

$$\begin{aligned} W_{\mathrm{QAD}}(\tilde{F}_\alpha) &:= \int a(x)\varphi\left(x, \int a(x)x dF(x)\right) dF(x) \\ &= \frac{1}{1-\alpha}\int_{Q(F,\underline{\alpha})}^{Q(F,\overline{\alpha})} \varphi\left(x, \mu(\tilde{F}_\alpha)\right) dF(x) \end{aligned} \tag{57}$$

Given a sample $x_{[1]}, \ldots, x_{[n]}$ of ordered (untrimmed) data, we have the sample trimmed mean

$$\hat{\mu}_\alpha := \mu(\tilde{F}_\alpha^{(n)}) = \frac{1}{\kappa(n,1-\alpha)}\sum_{i=1}^{n} x_{[i]}\iota(\kappa(n,\underline{\alpha})+1<i<\kappa(n,\overline{\alpha}))$$

which is equal to the usual sample mean but on the trimmed sample. The sample analogues of $W_{\mathrm{QAD}}(\tilde{F}_\alpha)$ in (57) are then given by

$$\hat{w}_{\mathrm{QAD},\alpha} := W_{\mathrm{QAD}}(\tilde{F}_\alpha^{(n)}) := \frac{1}{\kappa(n,1-\alpha)}\sum_{i=1}^{n} \varphi\left(x_{[i]}, \hat{\mu}_\alpha\right)\iota(\kappa(n,\underline{\alpha})+1<i<\kappa(n,\overline{\alpha})) \tag{58}$$

which is the counterpart of (33) but applied to the trimmed sample.

To establish the principal result for this class of indices in the trimmed case it is convenient to use both $\tilde{F}_\alpha$, the trimmed distribution (8) and F_α^*, the"censored"distribution (9) – cases D and E in Table 1. Then we have

Lemma 5 *The influence function for the quasi-additive class of welfare measures,* $\mathrm{IF}(z; W_{\mathrm{QAD}}, \tilde{F}_\alpha)$*, is given by* $[1-\alpha]^{-1}$ *times*

$$\begin{aligned} &\varphi\left(\max\left(x_{\underline{\alpha}}, \min\left(z, x_{\overline{\alpha}}\right)\right), \mu(\tilde{F}_\alpha)\right) - \int \varphi\left(x, \mu(\tilde{F}_\alpha)\right) dF_\alpha^*(x) \\ &\quad + \mathrm{IF}(z, C(\cdot;\overline{\alpha}), \tilde{F}_\alpha)\int_{Q(F,\underline{\alpha})}^{Q(F,\overline{\alpha})} \varphi_2\left(x, \mu(\tilde{F}_\alpha)\right) dF(x) \end{aligned} \tag{59}$$

Proof. By evaluating the mixture distribution and applying (21) $(1-\alpha)\mathrm{IF}(z; W_{\mathrm{QAD}}, \tilde{F}_\alpha)$ is found as

$$\begin{aligned} &-(1-\alpha)W_{\mathrm{QAD}}(\tilde{F}_\alpha) + \varphi\left(z, \mu(\tilde{F}_\alpha)\right)\iota(z\le x_{\overline{\alpha}})\iota(z\ge x_{\underline{\alpha}}) - \varphi\left(x_{\overline{\alpha}}, \mu(\tilde{F}_\alpha)\right)\iota(z\le x_{\overline{\alpha}}) \\ &\quad + \varphi\left(x_{\underline{\alpha}}, \mu(\tilde{F}_\alpha)\right)\iota(z\le x_{\underline{\alpha}}) + \mathrm{IF}(z, C(\cdot;\overline{\alpha}), \tilde{F}_\alpha)\int_{Q(F,\underline{\alpha})}^{Q(F,\overline{\alpha})} \varphi_2\left(x, \mu(\tilde{F}_\alpha)\right) dF(x) \\ &\quad + \overline{\alpha}\varphi\left(x_{\overline{\alpha}}, \mu(\tilde{F}_\alpha)\right) - \underline{\alpha}\varphi\left(x_{\underline{\alpha}}, \mu(\tilde{F}_\alpha)\right) \end{aligned} \tag{60}$$

where the first two lines follow by analogy with (35). The third line of (60) is found by considering the way the mixture distribution affects the limits of integration in (57) using Lemma 2. Rearranging (60) gives (59). ■

Using this and the key lemma 4 we obtain the principal result:

Theorem 6 *The asymptotic variance of* W_{QAD} *for the trimmed distribution* $\tilde{F}_\alpha$ *is* $[1-\alpha]^{-2}$ *times*

$$\begin{aligned}
&\operatorname{var}(\varphi\left(x,\mu(\tilde{F}_\alpha)\right);F_\alpha^*)\\
+2\frac{\operatorname{cov}\left(x,\varphi\left(x,\mu(\tilde{F}_\alpha)\right);F_\alpha^*\right)}{1-\alpha}&\int_{Q(F,\underline{\alpha})}^{Q(F,\overline{\alpha})}\varphi_2\left(x,\mu(\tilde{F}_\alpha)\right)dF(x)\\
+\frac{\operatorname{var}(x;F_\alpha^*)}{[1-\alpha]^2}&\left[\int_{Q(F,\underline{\alpha})}^{Q(F,\overline{\alpha})}\varphi_2\left(x,\mu(\tilde{F}_\alpha)\right)dF(x)\right]^2
\end{aligned}\tag{61}$$

Proof. The result again follows using Lemma 1 by integrating $\mathrm{IF}(z;W_{\mathrm{QAD}},\tilde{F}_\alpha)^2$ over $\mathfrak{X}$. Observe that Lemma 4 implies

$$\mathrm{IF}(z,C(\cdot;\overline{\alpha}),\tilde{F}_\alpha)=\frac{\max\left(x_{\underline{\alpha}},\min\left(z,x_{\overline{\alpha}}\right)\right)-\mu(F_\alpha^*)}{1-\alpha}\tag{62}$$

and that

$$\max\left(x_{\underline{\alpha}},\min\left(z,x_{\overline{\alpha}}\right)\right)dF(z)=zdF_\alpha^*(z)\tag{63}$$

$$\varphi\left(\max\left(x_{\underline{\alpha}},\min\left(z,x_{\overline{\alpha}}\right)\right),\mu(\tilde{F}_\alpha)\right)dF(z)=\varphi\left(z,\mu(\tilde{F}_\alpha)\right)dF_\alpha^*(z)\tag{64}$$

Substituting from (62-64) into (59) and squaring, the result follows immediately. ■

Note that in (61) the variance and covariance terms for the linear functionals (see 5 and 6) are defined on the "censored" distribution F_α^* (9) as opposed to the trimmed distribution (8). All the components of (61) can be estimated from the trimmed sample.

4.4 The Gini coefficient

With trimmed data, the Gini coefficient can be expressed as

$$I_{\mathrm{Gini}}(F)=1-2\int_{\underline{\alpha}}^{\overline{\alpha}}\frac{C(\tilde{F}_\alpha,q)}{C(\tilde{F}_\alpha,\overline{\alpha})}dq\tag{65}$$

Using the results of Theorem 5, we can obtain

Theorem 7 *The asymptotic variance of* $\sqrt{n}I_{\mathrm{Gini}}(\tilde{F}_\alpha^{(n)})$ *is*

$$\frac{4\vartheta_\alpha}{\mu_\alpha^4(1-\alpha)^2}$$

where

$$\vartheta_\alpha = \mu_\alpha^4(1-\alpha)^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{\alpha}}^{\overline{\alpha}} \upsilon_{qq',\alpha} dq' dq$$
$$= \mu_\alpha^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{\alpha}}^{q} \varpi_{q'q} dq' dq + \mu_\alpha^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{q}^{\overline{\alpha}} \varpi_{qq'} dq' dq +$$
$$\varpi_{\overline{\alpha}\overline{\alpha}} \left[\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q} dq \right]^2 - 2\mu_\alpha \int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q} dq \int_{\underline{\alpha}}^{\overline{\alpha}} \varpi_{q\overline{\alpha}} dq \tag{66}$$

Proof. In the case of trimmed samples we use (52) and (65). It is clear that the asymptotic covariance of $\sqrt{n} I_{\text{Gini}}(\tilde{F}_\alpha^{(n)})$is $4\vartheta/\mu^4$ where

$$\vartheta_\alpha := \mu_\alpha^4(1-\alpha)^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{\alpha}}^{\overline{\alpha}} \upsilon_{qq',\alpha} dq' dq \tag{67}$$

Using the method of Theorem 4 and substituting (56) in (67) we get

$$\begin{aligned} \vartheta_\alpha &= \mu_\alpha^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\underline{\alpha}}^{q} \varpi_{q'q} dq' dq + \mu_\alpha^2 \int_{\underline{\alpha}}^{\overline{\alpha}} \int_{q}^{\overline{\alpha}} \varpi_{qq'} dq' dq + \\ &\varpi_{\overline{\alpha}\overline{\alpha}} \left[\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q} dq \right]^2 - 2\mu_\alpha \int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q} dq \int_{\underline{\alpha}}^{\overline{\alpha}} \varpi_{q\overline{\alpha}} dq \end{aligned} \tag{68}$$

■

Once again, the development of (66) is in the Appendix (see A.2) and the estimates of ϑ_α are easily obtained by making use of (69), with $\widehat{\mu}_\alpha$ being the trimmed sample mean.

4.5 Empirical example 2

Let $n_\alpha = \kappa(n, 1-\alpha)$, and the observed ordered trimmed sample be $x_{[1]}, \ldots, x_{[n_\alpha]}$, a $100\,(1-\beta)\%$ confidence interval for the Gini coefficient is then given by

$$\left(1 - \frac{2}{\widehat{\mu}_\alpha n_\alpha^2} \sum_{i=1}^{n_\alpha} \sum_{j=1}^{i} x_{[j]} \mp z_{\beta/2} \sqrt{\frac{4\widehat{\vartheta}_\alpha}{n_\alpha(1-\alpha)\widehat{\mu}_\alpha^4}} \right)$$

From the Spanish data, we computed 95% confidence intervals for the Gini coefficient for the whole sample and a trimmed sample where 5% of the data have been removed from each side of the distribution. For the later we compute the confidence intervals both by ignoring and by considering the trimming. For the whole sample, we get $(0.2696; 0.3141)$ with a Gini estimate of 0.2918. For the trimmed sample, we get respectively $(0.2098; 0.2352)$ without a correction for the

trimming and $(0.2041; 0.2409)$ using the proper formula. The Gini estimate is 0.2225. We can draw two important conclusions. First the Gini estimate in the whole sample is larger and significantly different than in the trimmed sample. This might due to the presence of very large income values (see Cowell and Victoria-Feser 1996b). Second, the confidence interval in the trimmed sample is slightly larger when one uses the proper formula. This reflects the fact that the trimming introduces variability in the estimates.

If one defines the set of proportions $\Theta = \left\{ q_i = \underline{\alpha} + \frac{i(1-\alpha)}{n_\alpha} : i = 1, n_\alpha \right\}$, then (53) can be estimated by

$$\begin{aligned} \widehat{\varpi}_{q_i q_j} &= \left[q_i x_{[i]} - \underline{\alpha} x_{[1]} - [1-\alpha] \frac{1}{n_\alpha} \sum_{k=1}^{i} x_{[k]} \right] \\ &\quad \left[[1-q_j] x_{[j]} - [1-\underline{\alpha}] x_{[1]} + [1-\alpha] \frac{1}{n_\alpha} \sum_{k=1}^{j} x_{[k]} \right] - \\ &\quad \left[x_{[i]} [1-\alpha] \frac{1}{n_\alpha} \sum_{k=1}^{i} x_{[k]} - [1-\alpha] \frac{1}{n_\alpha} \sum_{k=1}^{i} x_{[k]}^2 \right] + \\ &\quad x_{[1]} \left[q_i x_{[i]} - \underline{\alpha} x_{[i]} - [1-\alpha] \frac{1}{n_\alpha} \sum_{k=1}^{i} x_{[i]} \right] \end{aligned} \tag{69}$$

A $100\,(1-\beta)\%$ confidence interval for the ith Lorenz ordinate is then

$$\left(\frac{1}{n_\alpha} \sum_{k=1}^{i} x_{[k]} \mp z_{\beta/2} \sqrt{\widehat{\varpi}_{q_i q_i} / n_\alpha (1-\alpha)} \right).$$

The calculations for the relative Lorenz ordinates are obtained in a straightforward manner using (69) and μ_α being estimated by the observed trimmed sample mean. Note that one can also choose a number of proportions $k < n_\alpha$ between $\underline{\alpha}$ and $\overline{\alpha}$ and change the indices i and j in (69) to respectively $\kappa(n_\alpha, q_i)$ and $\kappa(n_\alpha, q_j)$.

From the Spanish data, we compute the standard errors of the cumulative incomes for the whole sample and a trimmed sample where 5% of the data have been removed from each side of the distribution. For the latter we compute the standard errors both by ignoring and by considering the trimming. The results are presented in Figure 2. One notices that the standard errors for the trimmed sample without correction are the lowest overall. They are certainly not correct because trimming leads to an information loss which should appear as larger standard errors than in the untrimmed case. This is actually the case when one looks at the standard errors for the trimmed sample with correction.

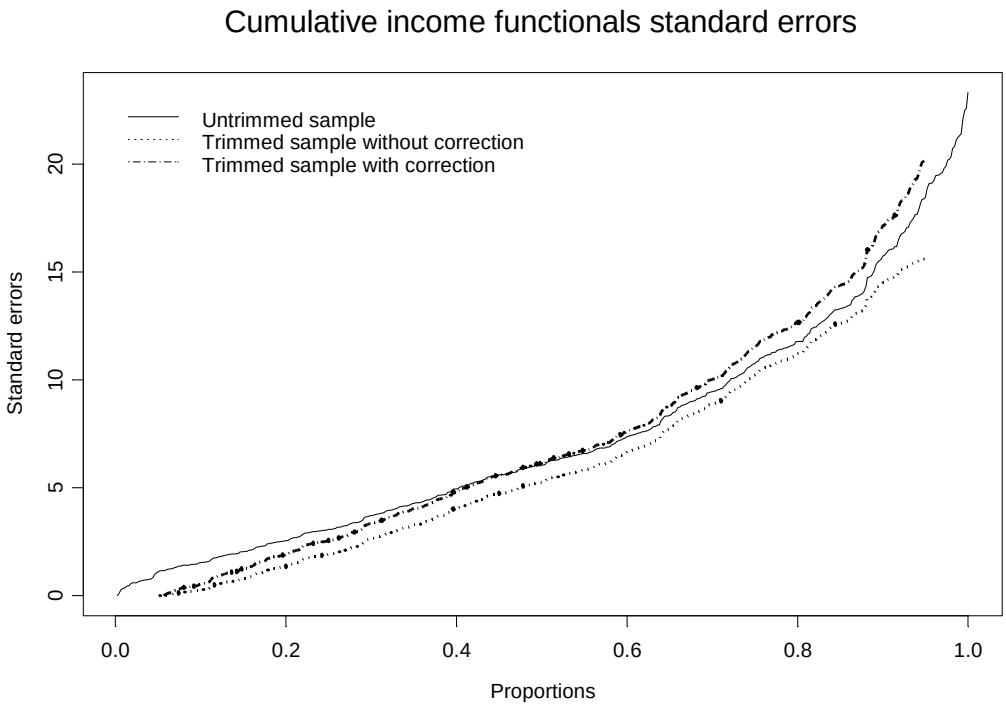

Figure 2: Standard error comparisons for complete and trimmed samples, with or without trimming correction

4.6 Information on the Excluded Subsample

In sections 4.2-4.5 we have focused on cases where information in the excluded part of the sample has been discarded. However, it is arguable that some information on the excluded subsample could be used in computing welfare statistics: how one goes about this depends on the type of data incompleteness one has. The key differences between the various situations are described in Table 1

4.6.1 Limits y fixed, α random

Consider the cases in the first row of Table 1. Here we can have two approaches to the excluded sample:

- where one is willing to make inference on the relevant quantities (such LC ordinates or the Gini index) for the whole population,
- where one makes inference on the same quantities for the part of the population from which the non-excluded part of the sample is generated.

According to the first approach, when one is in case A, one needs a parametric method to estimate the truncated part of the distribution. In case B, first order rankings could be analysed (although the estimation of the density is a problem) and for other quantities one has to act as in A. In case C, under some conditions on the available statistics on the censored part, second-order rankings and the Gini coefficient can be analysed. Using the second approach, the problem reduces to a redefinition of the target population. Under these circumstances no information on the excluded subsample is needed and the standard results for the complete data case can be used.

In the simplest case, the proportion of the censored part can be estimated (when given) and thus first-order rankings could be estimated. If the mean of the censored part is also given, then second order rankings can also be estimated, and thus the Gini coefficient as well. Once the resulting statistics are carefully written down, inference should be possible using the same methodology as for the other cases. For more complicated statistics, such as inequality indices involving nonlinear transformations of the data (such as φ, used in the definition of W_{QAD} – see 16) one can do little other than ignore the excluded subsample, unless the integrals of the relevant transformations over the censored parts, happen to be available.

The case of first-order rankings is not particularly interesting because of the problem of estimating the density. For second-order ranking criteria, we need the following statistics:

- $n_{\underline{\alpha}}$ (the sample quantity of the lower censored part),
- $n_{1-\overline{\alpha}}$ (the sample quantity of the upper censored part),
- n (the full sample size) or $n - n_{\underline{\alpha}} - n_{1-\overline{\alpha}}$,
- $n\hat{c}_{\mathrm{low}} := \sum_{i=1}^{n_{\underline{\alpha}}} x_{[i]}$, $n\hat{s}_{\mathrm{low}} := \sum_{i=1}^{n_{\underline{\alpha}}} x_{[i]}^2$

and for relative or absolute LC and the Gini coefficient also

- $n\hat{c}_{\mathrm{high}} := \sum_{n-n_{1-\overline{\alpha}}+1}^{n} x_{[i]}$, $n\hat{s}_{\mathrm{high}} := \sum_{n-n_{1-\overline{\alpha}}+1}^{n} x_{[i]}^2$.

For the first three quantities, one needs no more information than that the sample at hand is of size n, and $n_{\underline{\alpha}}$ values are equal to $\underline{y}$ and $n_{1-\overline{\alpha}}$ values are equal to $\overline{y}$ (case B of Table 1). The other two quantities are obviously more problematic. In some cases means will be available from data-providers (case C); otherwise satisfactory estimates may be achievable by using a parametric model of the tails.. If we have estimates of these two statistics, we can compute the uncensored part of the LC, RLC and Gini using the full sample results, i.e. for $q, q' \in (\underline{\alpha}, \overline{\alpha})$

$$\widehat{\omega}_{qq'} := \widehat{s}_q + [q\widehat{x}_q - \widehat{c}_q]\,[\widehat{x}_{q'} - q'\widehat{x}_{q'} + \widehat{c}_{q'}] - x_q\widehat{c}_q$$

with

$$\widehat{s}_q := \hat{s}_{\text{low}} + \frac{1}{n} \sum_{i=\kappa(n,\underline{\alpha})+1}^{\kappa(n,q)} x_{[i]}^2$$

$$\widehat{c}_q := \widehat{c}_{\text{low}} + \frac{1}{n} \sum_{i=\kappa(n,\underline{\alpha})+1}^{\kappa(n,q)} x_{[i]}$$

$$\widehat{x}_q := x_{[\kappa(n,q)]}$$

for the LC, and

$$\widehat{v}_{qq'} = \frac{1}{\widehat{\mu}^4} \left[\widehat{\mu}^2 \widehat{\omega}_{qq'} + \widehat{c}_q \widehat{c}_{q'} \widehat{\omega}_{11} - \widehat{\mu} \widehat{c}_q \widehat{\omega}_{q'1} - \widehat{\mu} \widehat{c}_{q'} \widehat{\omega}_{q1} \right]$$

with

$$\widehat{\mu} = \widehat{c}_{\overline{\alpha}} + \hat{c}_{\text{high}}$$

$$\widehat{s}_1 := \widehat{s}_{\overline{\alpha}} + \hat{s}_{\text{high}}$$

$$\widehat{\omega}_{11} := \widehat{s}_1 - \widehat{\mu}^2$$

$$\widehat{\omega}_{q1} := \widehat{s}_q + \left[q\widehat{x}_q - \widehat{c}_q \right] \widehat{\mu} - x_q \widehat{c}_q$$

for the RLC and the Gini.

As an example, take the Spanish data and compute the LC and RLC for the censored ($\underline{y} = 400$ and $\overline{y} = 1500$) sample as well as for the full sample. The required statistics are assumed to be known. The different Lorenz curves are presented in Figure 3. For the censored case, one gets exactly the full sample case from which the censored parts have been masked.

4.6.2 Proportions α fixed, y random

The second row of Table 1 presents a different situation to the researcher. Recall why one want to trim in the first place: outliers may seriously bias the point estimates as well as the variances of the distributional statistics that are of interest – see the results for the Gini coefficient in the Spanish example on page 22 above. Use of non-robust statistics – such as the mean – to summarise the data in the excluded subset (case F) may be inappropriate: otherwise the beneficial effect of trimming (i.e. more robust measures) would be lost.

However, note that from other aspects trimming (case D) is unsatisfactory and "censoring" (case E) is little better – except in freak cases these procedures will produce biased point estimates. Whether the data incompleteness is forced upon the researcher by the data-provider, or whether the researcher himself voluntarily discards some of the data, the way forward may require a special treatment of the excluded subset. An appropriate method may be to postulate a parametric model for the underlying income distribution in the tails – Cf (Cowell and Victoria-Feser

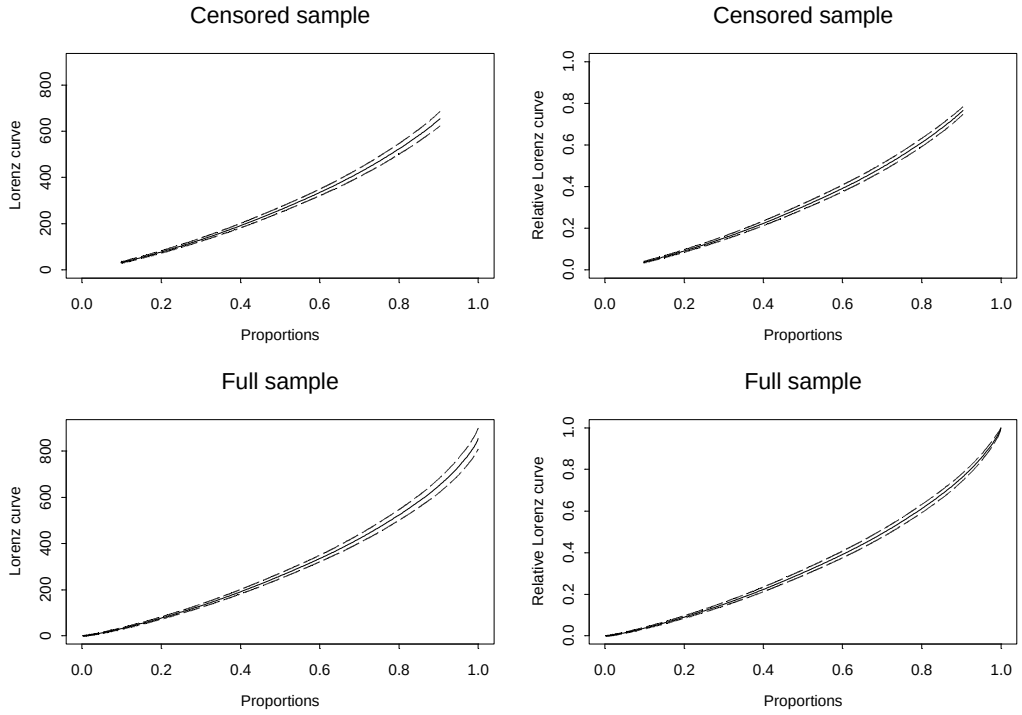

Figure 3: Use of information in the excluded subsample

1999). The richness of information in the excluded subset of the sample will determine the way in which this semi-parametric approach is to be implemented: but this takes us beyond the scope of the present paper.

5 Concluding Remarks

The influence function is a powerful tool for deriving the asymptotic variance of most of the welfare statistics used in income distribution analysis. Ranking criteria and summary indices of inequality and poverty can be handled with considerable economy of effort for both complete and incomplete data cases.

There are two basic paradigms for non-parametric approaches to the incomplete data problem: those based on cases where parts of the sample have been excluded according to some plan based on fixed income levels ("truncation" and "censoring") and those based on cases where the exclusions are made on the basis of fixed points in the set of income proportions ("trimming"). As we have seen the problem of finding asymptotic covariances of statistics in the first paradigm can be handled relatively straightforwardly – under one condition – as an ex-

tension of the complete-information case. The "trimming" paradigm presents a greater challenge, but the influence-function approach illustrates clearly how it may be derived from the complete-information case.

The condition referred to in the preceding paragraph concerns the use made of information in the excluded part of the sample. If one treats all of the excluded sample as though the information were irrecoverable (so that cases B and C in Table 1 were treated as case A) then indeed the problem is straightforward. However should one adopt this course? Making use of this information (where available) for truncated data can improve estimates of the sampling variance of Lorenz ordinates and the Gini coefficient. In the case of trimmed data this option is not available and explicit parametric modelling of the excluded part of the distribution may be required.

Finally, an appeal to data-providers. The practice of excluding or censoring some extreme observations from microdata sets on the grounds of confidentiality is understandable, but of course it causes serious problems for researchers on trends in income distribution. Publishing estimates of the mean and the variance in the excluded portion of the sample could greatly improve the point estimates and asymptotic covariances of some of the key distributional statistics, without compromising confidentiality.

References

Angrist, J. D. and A. B. Krueger (1999). Empirical strategies in labor economics. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3A, Chapter 23, pp. 1347.

Beach, C. M. and R. Davidson (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies 50*, 723–735.

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.

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

Cowell, F. A. (1999). Estimation of inequality indices. In J. Silber (Ed.), *Income Inequality Measurement: From Theory to Practice*. Dewenter: Kluwer.

Cowell, F. A. (2000). Measurement of inequality. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Chapter 2. Amsterdam: North Holland.

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 (1996c). Welfare judgements in the presence of contaminated data. Distributional Analysis Discussion Paper 13, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (1998). Statistical inference for Lorenz curves with censored data. Distributional Analysis Discussion Paper 35, STICERD, London School of Economics, London WC2A 2AE.

Cowell, F. A. and M.-P. Victoria-Feser (1999). Distributional analysis: A robust approach. In H. Glennerster (Ed.), *Putting Economics To Work*. Houghton St., London: STICERD.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics 19. New York: Springer.

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

Gastwirth, J. L. (1974). Large-sample theory of some measures of inequality. *Econometrica 42*, 191–196.

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.

Goldie, C. M. (1977). Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability 9*, 765–791.

Gottschalk, P. and R. A. Moffitt (1995). Trends in the covariance structure of earnings in the US: 1969 - 1987. mimeo, Brown University.

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.

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

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

Instituto Nacional de Estadistica (1992). Encuesta de presupuestos familiares 1990-1991. Metodologa, Madrid.

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

Moore, D. S. (1968). An elementary proof of asymptotic normality of linear functions of order statistics. *Annals of Mathematical Statistics 39*, 263–265.

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

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

Reeds, J. A. (1976). On the definition of von Mises functionals. Research Report S 44, Department of Statistics, Harvard University, Cambridge, Mass.

Sen, A. K. (1976). Poverty: An ordinal approach to measurement. *Econometrica 44*, 219–231.

Serfling, W. (1980). *Approximation Theorems in Mathematical Statistics*. New York: John Wiley.

Shorack, G. R. (1972). Functions of order statistics. *Annals of Mathematical Statistics 43*, 412–427.

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

Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. *Annals of Statistics 2*, 676–693.

A The Gini Coefficient

This appendix gives some derivations for the asymptotic variance for the Gini Coefficient in the complete information cases that enable one to relate our results to the literature; similar derivations are provided for the incomplete data case.

A.1 Complete Data

Taking terms in (44) of Theorem 4: separately and using (28) we have

$$\begin{aligned}
\int_0^1 \int_0^q \omega_{qq'} dq' dq &= \int_0^1 \int_0^q \left[s_{q'} + \left[q' x_{q'} - c_{q'} \right] \left[x_q - q x_q + c_q \right] - x_{q'} c_{q'} \right] dq' dq \\
&= \int_0^1 \int_0^q s_{q'} dq' dq + \int_0^1 x_q \int_0^q q' x_{q'} dq' dq - \\
&\quad \int_0^1 q x_q \int_0^q q' x_{q'} dq' dq + \int_0^1 c_q \int_0^q q' x_{q'} dq' dq - \\
&\quad \int_0^1 x_q \int_0^q c_{q'} dq' dq + \int_0^1 q x_q \int_0^q c_{q'} dq' dq - \\
&\quad \int_0^1 c_q \int_0^q c_{q'} dq' dq - \int_0^1 \int_0^q x_{q'} c_{q'} dq' dq
\end{aligned}$$

and

$$\begin{aligned}
\int_0^1 \int_q^1 \omega_{qq'} dq' dq &= \int_0^1 \int_q^1 \left[s_q + \left[q x_q - c_q \right] \left[x_{q'} - q' x_{q'} + c_{q'} \right] - x_q c_q \right] dq' dq \\
&= \int_0^1 s_q dq - \int_0^1 s_q q dq + \int_0^1 q x_q \int_q^1 x_{q'} dq' dq - \\
&\quad \int_0^1 q x_q \int_q^1 q' x_{q'} dq' dq + \int_0^1 q x_q \int_q^1 c_{q'} dq' dq - \\
&\quad \int_0^1 c_q \int_q^1 x_{q'} dq' dq + \int_0^1 c_q \int_q^1 q' x_{q'} dq' dq - \\
&\quad \int_0^1 c_q \int_q^1 c_{q'} dq' dq - \int_0^1 x_q c_q dq + \int_0^1 x_q c_q q dq
\end{aligned}$$

so that the sum $\int_0^1 \int_0^q \omega_{qq'} dq' dq + \int_0^1 \int_q^1 \omega_{qq'} dq' dq$ in (44) becomes

$$\begin{aligned}
&\int_0^1 \int_0^q s_{q'} dq' dq + \int_0^1 x_q \int_0^q q' x_{q'} dq' dq - \\
&\left[\int_0^1 q x_q dq\right]^2 + 2\left[\int_0^1 c_q dq\right]\left[\int_0^1 q x_q dq\right] - \\
&\int_0^1 x_q \int_0^q c_{q'} dq' dq - \left[\int_0^1 c_q dq\right]^2 - \\
&\int_0^1 \int_0^q x_{q'} c_{q'} dq' dq + \int_0^1 s_q dq - \int_0^1 s_q q dq + \\
&\int_0^1 q x_q \int_q^1 x_{q'} dq' dq - \int_0^1 c_q \int_q^1 x_{q'} dq' dq - \\
&\int_0^1 x_q c_q dq + \int_0^1 x_q c_q q dq
\end{aligned} \tag{70}$$

We also have

$$\omega_{11}\left[\int_0^1 c_q dq\right]^2 = \left[s_1 - \mu^2\right]\left[\int_0^1 c_q dq\right]^2 \tag{71}$$

and

$$\begin{aligned}
\int_0^1 \omega_{q1} dq &= \int_0^1 \left[s_q + \left[q x_q - c_q\right] c_1 - x_q c_q\right] dq \\
&= \int_0^1 s_q dq + \mu \int_0^1 q x_q dq - \\
&\quad \mu \int_0^1 c_q dq - \int_0^1 x_q c_q dq
\end{aligned} \tag{72}$$

so that, by substituting (70)-(72) into (44) we have

$$\begin{aligned}\vartheta &= \int_0^1\int_0^q s_{q'}dq'dq + \int_0^1 x_q\int_0^q q'x_{q'}dq'dq - \\ &\int_0^1 x_q\int_0^q c_{q'}dq'dq + \int_0^1 qx_q\int_q^1 x_{q'}dq'dq - \\ &\int_0^1 c_q\int_q^1 x_{q'}dq'dq - \int_0^1\int_0^q x_{q'}c_{q'}dq'dq - \\ &\left[\int_0^1 qx_qdq\right]^2 + \int_0^1 x_qc_qqdq + \\ &2(1-\mu)\left[\int_0^1 c_qdq\right]\left[\int_0^1 qx_qdq\right] - \int_0^1 s_qqdq + \\ &\left[2\int_0^1 c_qdq - 1\right]\left[\int_0^1 x_qc_qdq - \int_0^1 s_qdq\right] + \\ &\left[s_1 - \mu^2 + 2\mu - 1\right]\left[\int_0^1 c_qdq\right]^2 \end{aligned} \tag{73}$$

This result may be compared with, for example, Cowell (1989).

A.2 Incomplete data

Taking terms in (68) (72) of Theorem 7 separately and using (28) we have

$$\begin{aligned}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}\varpi_{qq'}dq'dq &= \int_{\underline{\alpha}}^{\overline{\alpha}}[1-q]\,x_q\int_{\underline{\alpha}}^{q}q'x_{q'}dq'dq - [1-\underline{\alpha}]\,x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}q'x_{q'}dq'dq + \\ &[1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}}c_{\alpha,q}\int_{\underline{\alpha}}^{q}q'x_{q'}dq'dq - \underline{\alpha}x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}[1-q]\,x_q(q-\underline{\alpha})dq - 0.5\underline{\alpha}^2\alpha^2x_{\underline{\alpha}}^2 - \\ &[1-\alpha]\,\underline{\alpha}x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}c_{\alpha,q}(q-\underline{\alpha})dq - [1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}}[1-q]\,x_q\int_{\underline{\alpha}}^{q}c_{\alpha,q'}dq'dq + \\ &[1-\underline{\alpha}]^2\,x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}c_{\alpha,q'}dq'dq - [1-\alpha]^2\int_{\underline{\alpha}}^{\overline{\alpha}}c_{\alpha,q}\int_{\underline{\alpha}}^{q}c_{\alpha,q'}dq'dq - \\ &[1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}x_{q'}c_{\alpha,q'}dq'dq + [1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}s_{\alpha,q'}dq'dq + \\ &x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}q'x_{q'}dq'dq - [1-\alpha]\,x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{\underline{\alpha}}^{q}c_{\alpha,q'}dq'dq\end{aligned}$$

and

$$\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{q}^{\overline{\alpha}}\varpi_{qq'}dq'dq = \int_{\underline{\alpha}}^{\overline{\alpha}} qx_q \int_{q}^{\overline{\alpha}} [1-q']\, x_{q'}dq'dq - [1-\underline{\alpha}]\, x_{\underline{\alpha}} \int_{\underline{\alpha}}^{\overline{\alpha}} qx_q(\overline{\alpha}-q)dq +$$
$$[1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}} qx_q \int_{q}^{\overline{\alpha}} c_{\alpha,q'}dq'dq - \underline{\alpha}x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{q}^{\overline{\alpha}} [1-q']\, x_{q'}dq'dq - 0.5\underline{\alpha}^2 x_{\underline{\alpha}}^2\alpha^2 -$$
$$[1-\alpha]\,\underline{\alpha}x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}}\int_{q}^{\overline{\alpha}} c_{\alpha,q'}dq'dq - [1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q}\int_{q}^{\overline{\alpha}} [1-q']\, x_{q'}dq'dq +$$
$$[1-\underline{\alpha}]\,[1-\alpha]\, x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q}(\overline{\alpha}-q)dq - [1-\alpha]^2\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q}\int_{q}^{\overline{\alpha}} c_{\alpha,q'}dq'dq -$$
$$[1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}} x_q c_{\alpha,q}(\overline{\alpha}-q)dq + [1-\alpha]\int_{\underline{\alpha}}^{\overline{\alpha}} s_{\alpha,q}(\overline{\alpha}-q)dq +$$
$$x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}} qx_q(\overline{\alpha}-q)dq - [1-\alpha]\, x_{\underline{\alpha}}\int_{\underline{\alpha}}^{\overline{\alpha}} c_{\alpha,q}(\overline{\alpha}-q)dq$$