

Quantile versions of the Lorenz curve

Luke A. Prendergast and Robert G. Staudte

*Department of Mathematics and Statistics
La Trobe University, Melbourne, Vic. Australia 3086*

e-mail: luke.prendergast@latrobe.edu.au; r.staudte@latrobe.edu.au

Abstract: The classical Lorenz curve is often used to depict inequality in a population of incomes, and the associated Gini coefficient is relied upon to make comparisons between different countries and other groups. The sample estimates of these moment-based concepts are sensitive to outliers and so we investigate the extent to which quantile-based versions can capture income inequality and lead to robust procedures. Distribution-free interval estimates of the associated coefficients of inequality are obtained, as well as sample sizes required to estimate them to a given accuracy. Convexity, transference and robustness of the measures are examined and illustrated.

Keywords and phrases: Gini index, inequality measures, influence function, quantile density.

Received November 2015.

Contents

| | | |
|---|---|------|
| 1 | Introduction | 1897 |
| 2 | Quantile analogues of the Lorenz curve | 1898 |
| | 2.1 Definitions and basic properties | 1898 |
| | 2.2 Coefficients of inequality | 1900 |
| | 2.3 Transference of income | 1902 |
| | 2.4 Example of transference: Income size-dependent levy | 1903 |
| | 2.5 Example of transference: Fixed levy | 1903 |
| 3 | Estimation of inequality measures | 1905 |
| | 3.1 Empirical quantile inequality curves | 1905 |
| | 3.2 Example using grouped data | 1906 |
| | 3.3 Empirical coefficients of inequality | 1907 |
| | 3.4 Simulation studies | 1907 |
| | 3.5 Distribution-free confidence intervals for the quantile coefficients of inequality | 1911 |
| 4 | Robustness properties | 1913 |
| | 4.1 Influence functions of quantile inequality curves | 1914 |
| | 4.2 Influence of contamination at on the graph $\{p, L_i(p)\}$ | 1915 |
| | 4.3 Influence functions of quantile coefficients of inequality | 1917 |
| 5 | Convexity of the quantile inequality curves | 1918 |
| | 5.1 Non-convex example | 1918 |
| | 5.2 Convex examples | 1919 |
| 6 | Summary and further research | 1921 |

| | |
|--|------|
| Appendix: Proof of Proposition 3 | 1922 |
| Acknowledgements | 1923 |
| References | 1923 |

1. Introduction

The Lorenz curve and the associated Gini coefficient are routinely employed for comparisons of income inequality in various countries. There are also numerous applications of them in the biological, computing, health and social sciences. These concepts have nice mathematical properties, and thus are the subject of numerous theoretical studies; for a recent review see [33]. However, when it comes to statistical inference for the Lorenz curve and the Gini coefficient, thorny issues arise. An excellent review of existing methods and new proposals for estimating the standard error of the Gini coefficient are investigated by [14]. As this author notes, such methods will not work when the variance of the income distribution is large or fails to exist, and of course this means that they are undermined by outliers in the data. Indeed, [6] show that many inequality measures in the econometrics literature have unbounded influence functions.

There are methods available for resolving these inferential obstacles. One is to choose a parametric income model and then to find optimal bounded influence estimators for the parameters; for example, [49] do this for the gamma and Pareto models. And, [48] shows how to robustly choose between parametric models and then find robust estimates of inequality indices based on a single data sample, even if it has been grouped or truncated. In a series of papers [7, 8, 9] investigate damaging effects of data contamination on transfer properties of various inequality indices, as well as dealing with the effects of truncation of non-positive and/or large data values. They propose semi-parametric models for overcoming these issues.

We go one step further here, redefining the basic concept of the Lorenz curve in terms of quantiles instead of moments, and then determining what has been gained and lost in terms of conceptual clarity, inference and estimator resistance to contamination. Examples of this approach are the standardized median in lieu of the standardized mean, and quantile measures of skewness and kurtosis, rather than the classical moment-based measures, [42, 43, 44]. Ratios of quantiles based on one sample are often presented as measures of inequality, and inferential procedures for them are available in [39, 38].

The role of quantiles in inequality measures is long-standing, beginning when [23] observed that the definition of the Lorenz curve could be extended to all distributions having a finite mean μ by expressing the cumulative income as an integral of the quantile function. More recently [24] showed that the inequality coefficient of [25] could be made much more sensitive to shifts in income inequality if the mean in its denominator were replaced by the median. While this has the advantage of protecting the denominator of the coefficient from large outliers, it does not protect the numerator.

The effects of means versus medians on poverty indices are investigated by [32]. It is in this spirit that we begin in Section 2 by introducing three simple quantile versions of the Lorenz curve for distributions on the positive axis, and their associated coefficients of inequality. Numerous examples demonstrate how these curves and coefficients agree or disagree with the moment-based classical version. In particular, the effects of income transfer functions on the inequality coefficients are illustrated in Sections 2.4 and 2.5.

In Section 3 we study empirical versions of these inequality curves and their associated estimated coefficients. Confidence intervals for the coefficients are found which have good coverage properties for a wide range of income distributions. These intervals are distribution-free in the sense that they only require consistent estimates of the quantile function and its density, which are included. It is not surprising that these quantile coefficients of inequality are resistant to outliers, and in Section 4 we show that they have bounded influence functions.

While the quantile versions of the Lorenz curve are not always convex, they are so for many standard distributions used to model incomes, as explained in Section 5. A summary and further research problems are given in Section 6.

2. Quantile analogues of the Lorenz curve

2.1. Definitions and basic properties

Let \mathcal{F} be the class of all cumulative distribution functions F with $F(0) = 0$. Such F will be interpreted as ‘income’ distributions and $p = F(x)$ as the proportion of incomes less than or equal to x . Define the *quantile function* associated with $F \in \mathcal{F}$ at each $p \in [0, 1)$ by $Q(F; p) = F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. When the meaning of F is clear, we will sometimes write x_p or $Q(p)$ for $Q(F; p)$.

The mean income of those having proportion p of the smallest incomes is $\mu = \mu_p(F) = \int_0^{x_p} x dF(x)/p$, and the mean income of the entire population is defined by $\mu = \mu(F) = \lim_{p \rightarrow 1} \mu_p$. Let $\mathcal{F}_0 \subset \mathcal{F}$ be the set of F for which $\mu(F)$ exists as a finite number. For each $F \in \mathcal{F}_0$ the *Lorenz curve* of F is defined by $L_0(F; p) \equiv p\mu_p/\mu$, for $0 \leq p \leq 1$. The lowest proportion of incomes p have proportion $L_0(p)$ of the total wealth.

What we are proposing here is to replace μ_p , the mean of the proportion p of those with wealth less than x_p , by its median $x_{p/2} = Q(F; p/2)$. In addition, we replace the mean μ of the entire population by one of three quantile measures of its size: $x_{1/2}$, $x_{1-p/2}$, or $(x_{p/2} + x_{1-p/2})/2$. The robustness merits of this last divisor, a symmetric quantile average, are investigated by [4].

Definition 1. For $F \in \mathcal{F}$ the three quantile inequality curves $\{(p, L_i(p))\}$ are defined for $p \in [0, 1)$ by:

$$\begin{aligned} L_1(F; p) &\equiv p \frac{x_{p/2}}{x_{0.5}} \\ L_2(F; p) &\equiv p \frac{x_{p/2}}{x_{1-p/2}} \end{aligned} \tag{1}$$

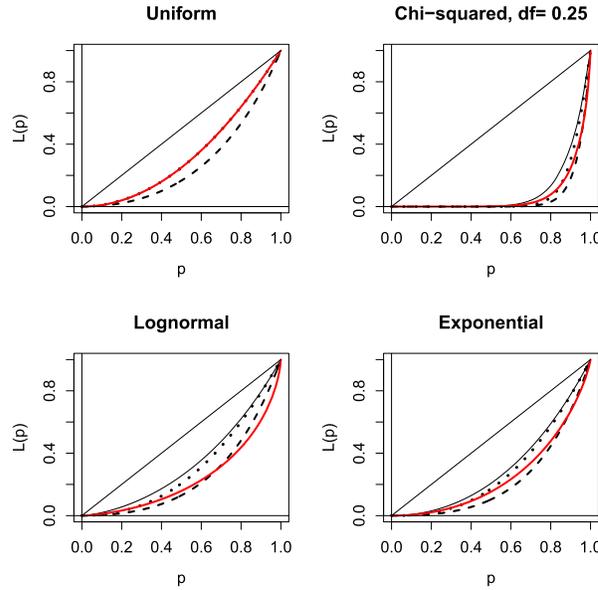


FIG 1. Graphs of $L_1(p)$ (solid line), $L_2(p)$ (dashed line), $L_3(p)$ (dotted line), defined in (1) for various models. The red line is the Lorenz curve.

$$L_3(F; p) \equiv 2p \frac{x_{p/2}}{(x_{p/2} + x_{1-p/2})} = \frac{2}{1/p + 1/L_2(F; p)} .$$

Also define $L_i(F; 1) = 1$ for $i = 1, 2$ and 3. We often abbreviate $L_i(F; p)$ to $L_i(p)$. Clearly $L_2(p) \leq L_1(p)$.

For each p the ordinate $L_1(p)$ compares the typical (median) wealth of the poorest proportion p of incomes with the typical (median) wealth of the entire population. The second $L_2(p)$ compares the bottom typical wealth with the top typical wealth; for example $L_2(0.2)$ corresponds to the popular ‘20-20 rule’, which compares the mean wealth of the lowest 20% of incomes with the largest 20%. For each p the third $L_3(p)$ gives the typical wealth of the poorest 100 p % incomes, relative to the mid-range wealth of the middle 100(1 - p) % of incomes. In all cases, extreme incomes are down-weighted because of multiplication by the factor p , as it is for the Lorenz curve $L_0(p) = p \mu_p / \mu$.

All of these quantile inequality curves $\{(p, L_i(p))\}$ are scale invariant and monotone increasing from $L_i(0) = 0$ to $L_i(1) = 1$, and all satisfy $L_i(p) \leq p$ for $0 \leq p \leq 1$. Each $L_i(p) \equiv p$ when all incomes are equal. None are strictly speaking ‘Lorenz’ curves, because they are not convex for all $F \in \mathcal{F}_0$, as examples will show. Nevertheless, for many commonly assumed income distributions F , they are convex, as shown in Section 5.

The third curve $\{(p, L_3(p))\}$ is the harmonic mean of $\{(p, L_2(p))\}$ and the diagonal line $\{(p, p)\}$ representing equal incomes, so $L_2(p) \leq L_3(p)$. Some examples of the quantile curves are depicted in Figures 1–2, which compares their

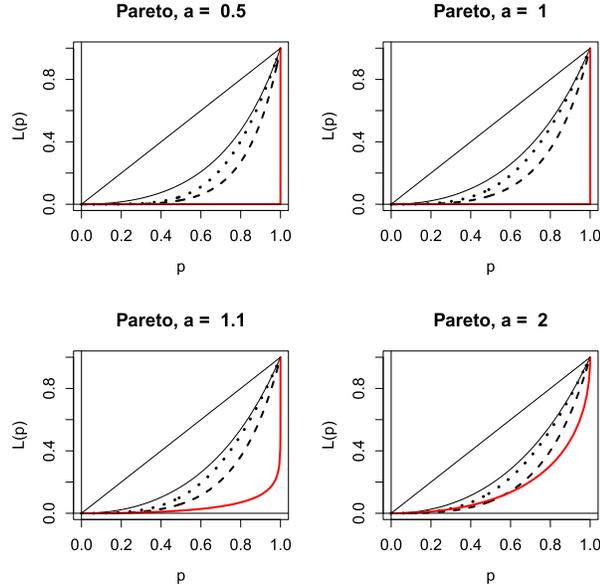


FIG 2. Graphs of $L_i(p)$ for Type II Pareto(a) Models, with the same notation as in Figure 1.

graphs with the Lorenz curve. Note that $L_0(p) \equiv L_1(p) \equiv L_3(p) \equiv p^2$ for the uniform distribution. And, $L_2(p) \approx p^3$ for the log-normal distribution. These plots show that the Lorenz curve is most sensitive to larger tailed income distributions, but these are exactly the situations where inference for them fails.

2.2. Coefficients of inequality

The relative measure of dispersion due to [25] is defined for $F \in \mathcal{F}_0$ by $G_0 = E|X_1 - X_2|/(2\mu)$, where X_1, X_2 are independent and each distributed as F , and μ is the mean of F . It is known, see [40] for example, to equal twice the area between the Lorenz curve and the diagonal line; it is an indicator, on the scale of 0 to 1, of ‘how far’ the inequality graph is from the diagonal line representing equal incomes; the further it is, the larger the Gini coefficient.

Definition 2. For each of the L_i given in (1) define the respective coefficients of inequality by:

$$G_i \equiv G_i(F) \equiv 2 \int_0^1 \{p - L_i(F;p)\} dp \quad \text{for all } F \in \mathcal{F}. \quad (2)$$

Specific numerical comparisons of the G_i s are given in Table 1. It lists a variety of F ranging from uniform to very long-tailed distributions and the associated values of Gini’s index for the four G_i s. The Dagum distribution [11] is a popular distribution for modeling incomes; it has two shape parameters a_1

TABLE 1
 Values of G_i to 3 decimal places for various F .
 Also listed are the rankings of F induced by the various G_i . For background on these standard distributions, see [29, 30].

| F | $G_0(F)$ | R_0 | $G_1(F)$ | R_1 | $G_2(F)$ | R_2 | $G_3(F)$ | R_3 |
|-----------------------------|----------|-------|----------|-------|----------|-------|----------|-------|
| 1. Uniform | 0.3333 | 2 | 0.3333 | 6 | 0.455 | 3 | 0.3333 | 4 |
| 2. $\chi_{0.5}^2$ | 0.7628 | 12 | 0.6713 | 14 | 0.792 | 13 | 0.7200 | 14 |
| 3. χ_1^2 | 0.6366 | 8 | 0.5251 | 12 | 0.673 | 10 | 0.5721 | 11 |
| 4. χ_3^2 | 0.4244 | 5 | 0.3289 | 4 | 0.483 | 4 | 0.3608 | 5 |
| 5. χ_5^2 | 0.3395 | 4 | 0.2614 | 3 | 0.406 | 3 | 0.2852 | 3 |
| 6. Lognormal | 0.5205 | 7 | 0.3328 | 5 | 0.510 | 5 | 0.3882 | 6 |
| 7. Pareto(0.5) ¹ | 1.0000 | 15 | 0.5151 | 11 | 0.704 | 11 | 0.6102 | 12 |
| 8. Pareto(1) | 0.9973 | 14 | 0.4548 | 10 | 0.636 | 9 | 0.5279 | 10 |
| 9. Pareto(1.5) | 0.7500 | 10 | 0.4343 | 9 | 0.609 | 8 | 0.4970 | 9 |
| 10. Pareto(2) | 0.6667 | 9 | 0.4240 | 8 | 0.595 | 7 | 0.4810 | 8 |
| 11. Weibull(0.25) | 0.9375 | 13 | 0.7311 | 15 | 0.843 | 14 | 0.7871 | 15 |
| 12. Weibull(0.5) | 0.7500 | 11 | 0.5700 | 13 | 0.720 | 12 | 0.6293 | 13 |
| 13. Weibull(1) | 0.5000 | 6 | 0.3933 | 7 | 0.550 | 6 | 0.4316 | 7 |
| 14. Weibull(4) | 0.1591 | 1 | 0.1364 | 1 | 0.222 | 1 | 0.1343 | 1 |
| 15. Dagum ² | 0.3352 | 3 | 0.2597 | 2 | 0.3884 | 2 | 0.2713 | 2 |

¹. The Lorenz curve and Gini coefficient are not defined for distributions with $\mu = +\infty$, but if the definition were so extended, $L_0(p)$ would be 0 for $0 < p < 1$ and the associated coefficient of inequality would be 1.

². The two shape and scale parameters for the Dagum [11] distribution are 4.273, 0.36 and 14.28 respectively.

and a_2 and a scale parameter b . We use the Type I Dagum distribution with $a_1 = 4.273$, $a_2 = 0.36$ and $b = 14.28$, which has been used to model US family incomes from 1969 in [34]. The rankings of different F s by these four measures of inequality are similar and the Spearman rank correlation of G_0 with G_i for $i = 1, 2$ and 3 are respectively 0.85, 0.90 and 0.90, for this list of F s.

Proposition 1. *Let $F \in \mathcal{F}$ have density $f = F'$ and denote its median by $m = F^{-1}(0.5)$. Choose two incomes Y_1, Y_2 independently and randomly from those incomes less than the median, and let $V = \max\{Y_1, Y_2\}$ be the larger. Then G_1 defined by (2) is the average relative distance of V from the median: $G_1 = E[(m - V)/m]$.*

Further define $W = F^{-1}(1 - F(V))$, so if $V = x_r$ is the r th quantile of F , $W = x_{1-r}$. Then $G_2 = E[(W - V)/W]$ and $G_3 = E[(W - V)/(V + W)]$.

Proof. Let Y have the conditional distribution of X given $X \leq m$; then its distribution function $F_Y(y) = 2F(y)$, for $0 \leq y \leq m$ and the distribution of V is determined by $F_V(v) = F_Y^2(v) = 4F^2(v)$, for $0 \leq v \leq m$. Consider the integral of L_1 in (2), and make the change of variable $v = F^{-1}(p/2)$ to obtain:

$$1 - G_1 = \int_0^1 \frac{2p F^{-1}(p/2)}{m} dp = \int_0^m \frac{8v F(v) f(v)}{m} dv = \frac{1}{m} \int_0^\infty v dF_V(v) . \quad (3)$$

The results for G_2 and G_3 are obtained in a similar manner. □

Proposition 1 shows that $G_1 \leq G_2$ and $G_3 \leq G_2$ for all F . It also allows for simple alternative interpretations of the three quantile inequality coefficients

defined in (2) which can be compared with Gini's original definition as a relative measure of concentration.

The Gini measure has been criticized for placing too much emphasis on the central part of the distribution. As Proposition 1 shows, the quantile versions can be criticized for the same reason, for they depend on the *maximum* of two randomly chosen incomes from the lower half of the population. This maximum arises because when making the change of variable in (3), the p is changed to $2F(v)$, part of the density of $V = \max\{Y_1, Y_2\}$. If for example, L_1 were redefined (without the multiplier p) to be $L_1^*(p) = x_{p/2}/x_{0.5}$ taking values in $[0,1]$, and G_1 redefined to $G_1^* = 1 - \int_0^1 L_1^*(p) dp$, then the calculation in (3) would become $1 - G_1^* = \int_0^m 2y f(y) dy/m = \int_0^m y dF_Y(y) = E[Y]/m$, where Y has the conditional distribution of X , given that X is less than its median. Thus $G_1^* = (m - E[Y])/m$ the average relative distance of a *single* randomly chosen income less than the median from the median.

2.3. Transference of income

The effect of income transfers on inequality measures is of great interest to economists, see [33] and [21]. The basic idea [12] is that if one transfers income from some having income above the mean to others having income below the mean while preserving income order, then the coefficient of inequality should reflect this by decreasing. Our definition to follow requires that after transference no quantile should be further from the median.

Definition 3. Given $X \sim F_X \in \mathcal{F}$, and let $m \equiv x_{0.5} = F_X^{-1}(0.5)$ be the median. We define a median preserving transfer (of income) function $y = t(x)$ as one which is non-decreasing and satisfies $t(x) \geq x$ for $x < m$, $t(m) = m$ and $t(x) \leq x$ for $x > m$. The graph $\{(x, t(x))\}$ lies on or above the diagonal for $x < m$, passes through (m, m) , and lies on or below the diagonal for $x > m$.

For such t we have $Y = t(X) \sim F_Y$, where $F_Y(y) = P(t(X) \leq y)$ for all y . Hence $y_p = Q(F_Y; p) \geq Q(F_X; p) = x_p$ for all $0 < p < 0.5$ and $y_p = Q(F_Y; p) \leq Q(F_X; p) = x_p$ for all $0.5 < p < 1$. The effect on the quantile inequality curves is then easily seen: $L_1(F_X; p) = p x_{p/2}/x_{0.5} \leq p y_{p/2}/y_{0.5} = L_1(F_Y; p)$; that is, the transfer function can only increase $L_1(p)$ at each p . This implies the associated coefficient of inequality (2) satisfies $G_1(F_X) \geq G_1(F_Y)$. We say that L_1 preserves the ordering induced by the transfer function. The reader can verify that for $i = 2, 3$ the other quantile inequality curves satisfy $L_i(F_X; p) \leq L_i(F_Y; p)$ and hence $G_i(F_X) \geq G_i(F_Y)$. For any non-trivial transfer function we will have $G_i(F_X) > G_i(F_Y)$, a positive reduction in the coefficient of inequality.

The above definition does not require the existence of the mean $\mu = \int x dF(x)$, which is useful in theoretical papers, but in practice μ will be finite and one would normally require the transfer function to preserve the mean as well. The definition is strong in that for such t each $L_i(p)$ is ordered, and weaker definitions would only require that each G_i be ordered. [1] discuss income transfer functions which are both median and mean preserving in some detail.

2.4. Example of transference: Income size-dependent levy

First we consider a levy that is dependent on income size. For example, suppose that we wish to bring the poorest $p100\%$ up to the $p100$ th income percentile by imposing a levy on the $q100\%$ richest individuals. Let $s_1 = \sum_{x < x_p} (x_p - x)$ denote the total that needs to be distributed; that is, the sum of the differences between x_p and the lowest incomes. Similarly, let $s_2 = \sum_{x > x_q} (x - x_q)$ denote the total of the difference between richest incomes and x_q . Then, provided it is less than one, $p_l = s_1/s_2$ equals the proportion each income above x_q required so the total levied can be distributed to those under x_p . For any $x > x_q$, the levy imposed is $p_l \times (x - x_q)$, so a flat percentage $100 p_l\%$ levy is imposed on all incomes above x_q .

As an example, we sample 10,000 values from the Dagum(4.273,14.28,0.36) distribution [11]. For these data we have $x_{0.2} = 5.036$ and we wish to make this the minimum income in the levy-adjusted incomes by imposing a levy on those with incomes above $x_{0.5} = 9.364$.

TABLE 2
Summary measures of pre-adjusted and post-adjusted data following an income size-dependent levy.

| | Min. | $x_{0.25}$ | Median | Mean | $x_{0.75}$ | Max. |
|------------|-------|------------|--------|--------|------------|---------|
| Unadjusted | 0.027 | 5.873 | 9.364 | 10.380 | 13.480 | 110.000 |
| Adjusted | 5.036 | 5.873 | 9.364 | 10.380 | 12.920 | 96.140 |

In Table 2 we provide summary measures of pre-adjusted and post-adjusted data following the size-dependent levy. For the original data, the minimum was 0.027 and we wish to increase the minimum to 5.036. To achieve this, we need to take $p_l = 0.137$ (13.7%) of the difference between each income and 5.036 for each income above 9.364. For example suppose $x = 10.0$, then the levy imposed is 1.37. In the final row of the table we provide summary results for the adjusted data. The median has been preserved and so to has the first quartile, however the minimum has increased to 5.036. The total redistributed was 3991.72.

Some might consider a size-dependent levy as described above as unfair or difficult to implement, and prefer a fixed levy on incomes above a certain threshold, as considered next.

2.5. Example of transference: Fixed levy

Suppose one wants to increase all incomes less than a specific threshold b (say the poverty line) so that they equal b . That is; $t(x) = b$ for $0 < x \leq b$. This requires an amount per person of $d = b - (\int_0^b x dF(x))/F(b)$ to be found, say, by transference from those with incomes above the median or some higher threshold c . One possibility is to charge a levy of amount d on those with income exceeding c , leading to the following transfer function $Y = t(X) \sim F_Y$:

$$y = t(x) = \begin{cases} b, & 0 \leq x < b; \\ x, & b \leq x < c; \\ x - d, & c \leq x . \end{cases} \tag{4}$$

In the interest of fairness one could also charge a proportional amount for those with income between c and $c + d$ so that $Y = c$ for $c < x < c + d$, but this unnecessarily complicates our presentation.

Now $F_Y(y)$ jumps from 0 to $F(b)$ at b , equals $F(y)$ for $b \leq y < c$, jumps at c from $F(c)$ to $F(c + d)$ and equals $F(y + d)$ for $c \leq y$. Therefore the quantile function $Q(F_Y; p)$ for the transferred income Y is given by

$$Q(F_Y; p) = \begin{cases} b, & 0 \leq p < F(b); \\ F^{-1}(p), & F(b) \leq p < F(c); \\ c, & F(c) \leq p < F(c + d); \\ F^{-1}(p) - d, & F(c + d) \leq p. \end{cases} \quad (5)$$

At this point it is convenient to introduce the p th cumulative income by $C(F; p) = \int_0^{x_p} y dF(y)$, where $x_p = Q(F; p)$. As [7] point out, this function is fundamental to analysis of Lorenz curves, and $C(1; F) = \mu$ and $L_0(F; p) = C(F; p)/C(1; F)$. We want to determine $C(F; p)$ for the Type II Pareto distribution having shape parameter $a > 1$ and scale parameter $\sigma > 0$.

Now $1 - F_{a,\sigma}(x) = (1 + x/\sigma)^{-a}$, which has mean $\mu = \sigma/(a - 1)$ and p th quantile $Q(F_{a,\sigma}; p) = \sigma\{(1 - p)^{-1/a} - 1\}$. Integrating by parts we obtain

$$C(F_{a,\sigma}; p) = \int_0^{\sigma x_p} y dF_{a,\sigma}(y) = \frac{\sigma}{a - 1} \{p - a(1 - p)x_p\}, \quad (6)$$

where $x_p = Q(F_{a,1}; p)$. The mean income of the poorest proportion p is $\mu_p = C(F_{a,\sigma}; p)/p$.

For the transfer problem with $F_{a,\sigma}(b) = p < 0.5$, we have $b = \sigma x_p$, so (6) implies

$$d = b - \mu_p = \frac{\mu}{p} \{(a - p)x_p - p\}.$$

This amount can be obtained by a levy d on each income greater than $c = x_{1-p}$.

For the Pareto distribution with parameters $a = 2$, $\sigma = 100,000$, the median income is 41,421.36 and the mean income is $\mu = 100,000$. For $p = 0.2$, say, the quantities of interest are the poverty line $b = 11,803.40$, the mean cumulative income $\mu_{0.2} = 5,572.80$ and $d = 6,230.60$. All those having income greater than the 0.8 quantile 123,606.30 would need to pay an impost of $d = 6,230.60$.

The absolute and relative effects of such a transfer function are depicted in Figure 3 for two income distributions, Pareto with $a = 1.1$ and $a = 2$. For the first distribution, the change in the Gini coefficient G_0 is larger than for the G_2 and G_3 coefficients, but less than that for G_1 ; but the relative effect plot shows that the G_1 coefficient is most sensitive of the four, especially for p_0 near 0.25. For the second distribution both G_0 and G_1 are roughly the same in terms of sensitivity to changes by transference and again preferable to G_2 and G_3 .

Many other transfer functions and income distributions could be considered; what is politically feasible, fair and implementable transference functions are important applications beyond the scope of this work.

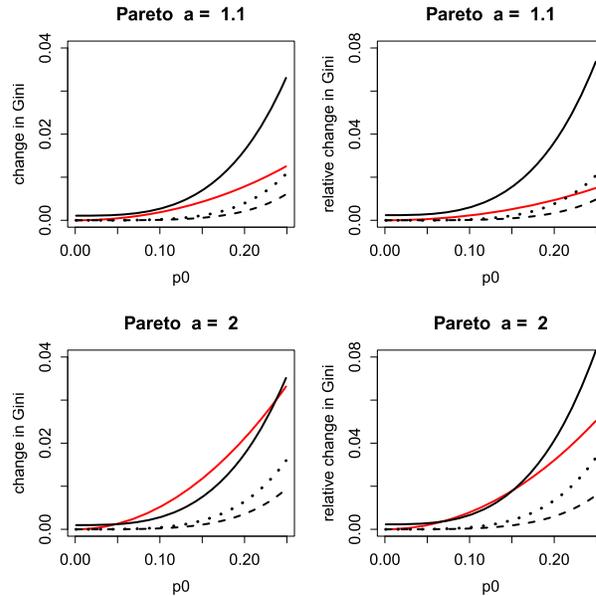


FIG 3. The left hand plots show the graphs of the absolute change in the inequality coefficients $G_i(F; p) - G_i(F_Y; p)$ caused by the income transference (4) for $i = 0$, red line; $i = 1$, solid line; $i = 2$, dashed line; and $i = 3$, dotted line. The right hand plots show the relative changes.

3. Estimation of inequality measures

In the last section we showed that a certain transfer of income from higher to lower incomes would lead to a measurable reduction in the inequality coefficients when the underlying distribution was known. In practice we want, for a given sample of incomes, to estimate these coefficients and their standard errors.

3.1. Empirical quantile inequality curves

Given data x_1, \dots, x_n with ordered values $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ let $L_0(0) = 0$ and $L_0(i/n) = \sum_{j \leq i} x_{(j)} / \sum_{j \leq n} x_{(j)}$ for $i = 1, \dots, n$. The empirical Lorenz curve is then defined as the graph of the piecewise linear connection of the points $(i/n, L_0(i/n))$, $i = 0, 1, \dots, n$. The empirical distribution function defined for each x by $F_n(x) = \{\sum_{i=1}^n I(X_i \leq x)\} / n$. It has inverse $Q(F_n; p) = F_n^{-1}(p) = x_{([np]+1)}$ for $0 \leq p < 1$, and so empirical versions of the quantile curves (1) can be expressed in terms of the n order statistics. Such curves are discontinuous, but there are several continuous quantile estimators available, including kernel density estimators [41] and the linear combinations of two adjacent order statistics studied by [27]. Many of the latter are implemented on the software package *R* [15], and here we use the Type 8 version of the `quantile` command recommended by [27]. It linearly interpolates between the points $(p_{[k]}, x_{(k)})$, where

$p_{[k]} = (k - 1/3)/(n + 1/3)$ for $k = 1, \dots, n$. It is the continuous function of p in $(0, 1)$ given by:

$$\hat{Q}(p) = \begin{cases} x_{(1)}, & 0 < p \leq p_{[1]} \\ x_{(k)} + b_k (p - p_{[k]}), & p_{[k]} < p \leq p_{[k+1]}, \quad k = 1, \dots, n - 1 \\ x_{(n)}, & p_{[n]} < p < 1. \end{cases}$$

where $b_k = (n + 1/3)(x_{(k+1)} - x_{(k)})$. Often we abbreviate $\hat{Q}(p)$ to \hat{x}_p .

Definition 4. All of the L_i curves defined by (1) are functions of the quantile function $Q(F; p)$, so given the estimator $\hat{x}_p = \hat{Q}(p)$ one can by substitution obtain estimators of each of the $L_i(p)$ for any p in $(0, 1)$; we call these estimators $\hat{L}_i(p)$, for $i = 1, 2$, and 3 . Also let $\hat{L}_i(0) = 0$ and $\hat{L}_i(1) = 1$ for $i = 1, 2$, and 3 .

3.2. Example using grouped data

Often for issues of privacy, income data is commonly reported in a summarized form. Recently, [35] showed how the Lorenz curve and Gini coefficient could be computed when provided with intervals of quantiles and mean incomes within these intervals. Their method can be adapted to find quantile inequality curves and coefficients. Table 3 presents grouped income data obtained from [5] which was used as an example in [35]. Their idea is to first create an approximate density function using linear interpolation within the reported closed intervals and an exponential tail for the final open interval. The slope within each interval is based on the difference between the mean and interval midpoints, a method that is an extension of the simple method of assuming uniformity within intervals that was considered by [46].

In Figure 4 are shown the linearly interpolated density (Plot A) based on the intervals reported in Table 3, the associated cumulative density (Plot B) and the resulting Lorenz and quantile-based curves L_1 , L_2 and L_3 . An advantage of the density in Plot A is that, due to its simplicity, one can obtain closed-form solutions for the cumulative distribution function and its inverse required for estimation. We used adaptive quadrature to compute the inequality coefficients $G_0 = 0.468$, $G_1 = 0.343$, $G_2 = 0.499$, $G_3 = 0.377$. The value of G_0 is approximately the Gini coefficient reported in [35].

TABLE 3
Mean values of incomes falling within income quantiles x_a USD and x_b USD from 2010 household income data.

| a | b | $[x_a, x_b)$ | Mean income |
|------|------|----------------------|-------------|
| 0 | 0.2 | [0, 20,000) | 10,994 |
| 0.2 | 0.4 | [20,000, 38,000) | 28,532 |
| 0.4 | 0.6 | [38,000, 61,500) | 49,167 |
| 0.6 | 0.8 | [61,500, 100,029) | 78,877 |
| 0.8 | 0.95 | [100,029, 180,485) | 130,121 |
| 0.95 | 1.0 | [180,485, ∞) | 287,201 |

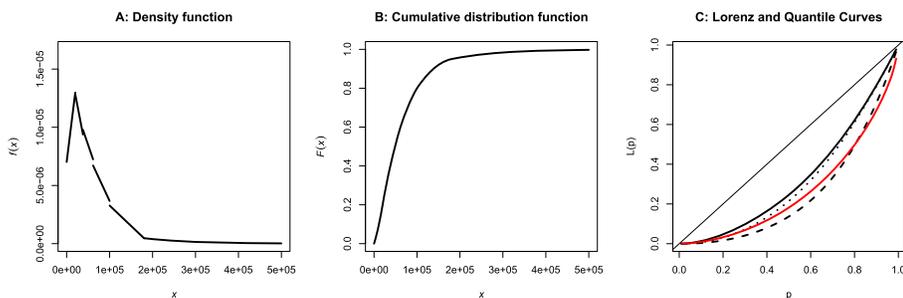


FIG 4. Approximate density function (Plot A) using linear interpolations and an exponential right tail, associated cumulative density function (Plot B) and the associated Lorenz and Quantile Curves (Plot A) L_0 (red line) and L_i -curves, (solid, dashed, and dotted lines, respectively).

3.3. Empirical coefficients of inequality

With few exceptions, such as the uniform distribution, one cannot analytically compute the $G_i(F)$ s, but using modern software packages such as R [15], it is easy to get very good approximations to them for many F of interest as follows.

Definition 5. Given a large integer J define a grid in $(0,1)$ with increments of size $1/J$ by $p_j = (j - 1/2)/J$, for $j = 1, 2, \dots, J$. Then evaluate the quantile function $Q(p_j)$ for p_j in the grid and find $G_i(J) \equiv (2/J) \sum_j \{p_j - L_i(p_j)\}$ for each $i = 1, 2$ and 3 .

Clearly one can make $G_i(J)$ as close to G_i as desired by choosing J sufficiently large. We will estimate $G_i(J)$, and hence G_i , as follows. Let $\hat{L}_i(p_j)$ be the estimated inequality curve value at p_j , for each p_j in the grid. Then $\hat{G}_i(J)$ is defined by

$$\hat{G}_i(J) \equiv (2/J) \sum_j \{p_j - \hat{L}_i(p_j)\} . \tag{7}$$

In our computations, we used $J = 1000$. Hereafter we write G_i for $G_i(J)$ and \hat{G}_i for $\hat{G}_i(J)$, but it is understood that these are computed on a grid with increments $1/J$.

3.4. Simulation studies

It will be shown that despite the values of the quantile coefficients of inequality $G_i(F)$ varying greatly over the range of F in Table 1, the standard errors of estimation are relatively stable. By ‘standard error’ of \hat{G}_i , we mean the square root of the mean squared error. Initial simulations suggested that $\text{Bias}[\hat{G}_i] = o(n^{-1/2})$ and $\text{Var}[\hat{G}_i] = O(1/n)$ so in Figure 5 we show some examples of $\sqrt{n} \text{SE}[\hat{G}_i(F)]$, plotted as a function of $\ln(n)$, for n ranging from 20 to 1600. These plots are based on 1000 replications at each of the selected values of n for various F . In all four plots it is seen that the standard errors of

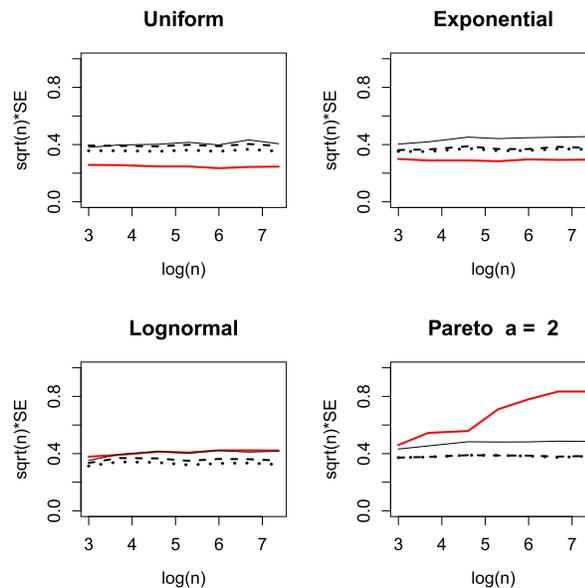


FIG 5. $\sqrt{n}SE[\hat{G}_i(F)]$ plotted as a function of $\ln(n)$ for the Lorenz curve L_0 (red line) and L_i -curves, (solid, dashed, and dotted lines, respectively).

$\hat{G}_2(F) \approx \hat{G}_3(F) \approx 1/(2\sqrt{n})$ while $\hat{G}_1(F)$ is a little larger. This enables one to choose a sample size which guarantees a desired standard error for each of the three estimators. Attempting to estimate Gini's coefficient of inequality by means of the Lorenz curve areas has no such simple sample size solution.

We also plot $\sqrt{n}SE[\hat{G}_i(F_a)]$ versus a in Figure 6, where F_a denotes the Pareto distribution with shape parameter a ranging from 0.25 : 2.5/0.1. Again all three standard errors of the estimated inequality coefficients derived from the L_i -curves are well behaved, but those for the Lorenz curve are quite irregular. For $a \leq 1$ the Lorenz curve is not defined because $E_a[X] = +\infty$ but if one defines the curve to be 0 in this case the corresponding measure of inequality is 1 and this can be estimated. Even if one restricts attention to $1 < a < 2$, these plots suggest that for increasing n the standard error of the estimated Gini coefficient, multiplied by \sqrt{n} , is growing with increasing n . The reason for this behavior is that for $a \leq 2$ the variance of $X \sim F_a$ is infinite. Assuming $a \leq 2$, the larger the sample size, the more likely it is that an extreme outlier will be in the sample, and this will result in greater estimated variance of the Gini estimators.

The results in Table 4 suggest that one can choose the minimum sample size required to obtain $SE[\hat{G}_1] \leq c$; it is $n_1(c) = (0.55/c)^2$. So for example, for standard error $c = 0.01$, one needs $n \geq n_1 \approx 3000$. Note that this accuracy is achieved for all F in Table 4. Similarly for G_2, G_3 the required sample size is a little smaller $n_2(c) = (0.43/c)^2 = n_3(c)$.

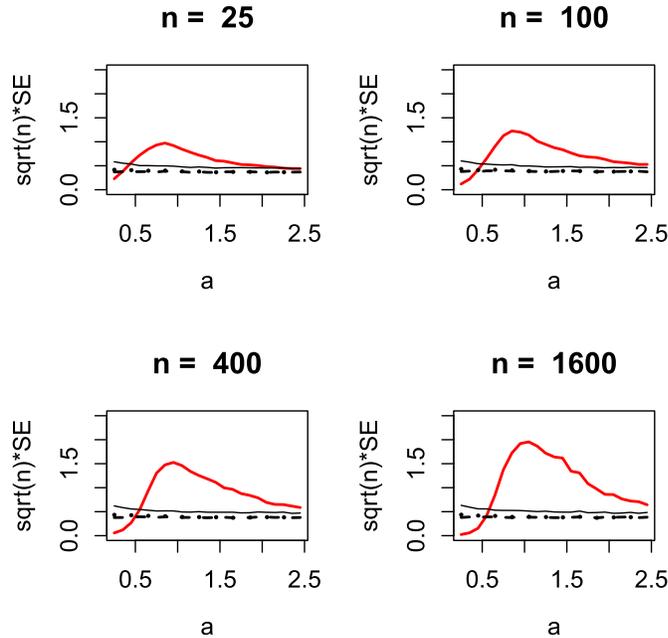


FIG 6. $\sqrt{n} SE[\hat{G}_i(F_a)]$, for Pareto(a) distributions, plotted as a function of a.

An online R script is provided as supplementary material. Given a set of data x from an arbitrary income distribution, it plots the empirical L_i -curves computes the associated inequality coefficients \hat{G}_i and gives the upper bounds on their standard errors.

Convergence to normality for the estimators of G_1 , G_2 and G_3 may also be of interest. For example, in the case of the Gini Index and data arising from an Expo(1) distribution, [22] show that the estimator is approximately normal, even for sample sizes as small as $n = 10$. We simulated 10,000 data sets from the Expo(1) distribution for each of $n = 10, 25, 50, 100$ and 500 . Histograms of the estimates for G_i ($i = 0, 1, 2, 3$) indicated approximate normality for \hat{G}_0 for $n = 10$, as previously noted, with similarly good results for \hat{G}_2 . While \hat{G}_1 and \hat{G}_3 did not achieve a comparable degree of approximate normality for $n = 10$, approximate normality was still achieved and by $n = 25$ all estimators were comparably very close to normal.

We investigated many other distributions for rapidity to normality and it is clear that, with respect to which of the \hat{G}_i ($i = 0, 1, 2, 3$) achieves approximate normality the quickest, this is very much dependent on the distribution itself. In fact, for some distributions we found that convergence to normality for \hat{G}_0 was extremely slow. For example, for the Pareto(2.5) distribution for which the G_0 exists due to finite first and second moments, \hat{G}_0 is skewed, even for quite large n , where as the quantile-based measures converge to normality quickly. In Figure 7 we provide Q-Q plots for 10,000 simulated \hat{G}_0 and \hat{G}_1 estimates for n

TABLE 4
 Standard errors of \hat{G}_i : $\sqrt{n} SE[\hat{G}_i]$ for $n = 25, 100$ together with the respective asymptotic SEs $\sigma_i = \lim \sqrt{n} SE[\hat{G}_i]$, based on numerical evaluation of the integrals in (14). The finite sample standard errors are based on 4000 samples of size n .

| F | G_1 | | | G_2 | | | G_3 | | |
|-------------------|-------|------|-----------|-------|------|-----------|-------|------|-----------|
| | 25 | 100 | $+\infty$ | 25 | 100 | $+\infty$ | 25 | 100 | $+\infty$ |
| 1. Uniform | 0.40 | 0.40 | 0.421 | 0.38 | 0.39 | 0.399 | 0.35 | 0.35 | 0.361 |
| 2. $\chi_{0.5}^2$ | 0.55 | 0.55 | 0.550 | 0.39 | 0.38 | 0.359 | 0.43 | 0.43 | 0.405 |
| 3. χ_1^2 | 0.50 | 0.53 | 0.521 | 0.40 | 0.41 | 0.402 | 0.42 | 0.44 | 0.427 |
| 4. χ_3^2 | 0.39 | 0.40 | 0.408 | 0.34 | 0.36 | 0.351 | 0.31 | 0.33 | 0.316 |
| 5. χ_5^2 | 0.32 | 0.33 | 0.337 | 0.30 | 0.32 | 0.305 | 0.26 | 0.27 | 0.253 |
| 6. Lognormal | 0.39 | 0.40 | 0.417 | 0.34 | 0.35 | 0.351 | 0.32 | 0.32 | 0.322 |
| 7. Pareto(0.5) | 0.53 | 0.54 | 0.540 | 0.38 | 0.39 | 0.351 | 0.41 | 0.42 | 0.370 |
| 8. Pareto(1) | 0.49 | 0.50 | 0.507 | 0.37 | 0.38 | 0.371 | 0.38 | 0.39 | 0.376 |
| 9. Pareto(1.5) | 0.46 | 0.47 | 0.492 | 0.36 | 0.38 | 0.379 | 0.36 | 0.38 | 0.380 |
| 10. Pareto(2) | 0.45 | 0.46 | 0.485 | 0.37 | 0.38 | 0.381 | 0.37 | 0.38 | 0.379 |
| 11. Weibull(0.25) | 0.55 | 0.53 | 0.540 | 0.35 | 0.34 | 0.330 | 0.40 | 0.39 | 0.384 |
| 12. Weibull(0.5) | 0.53 | 0.53 | 0.550 | 0.38 | 0.39 | 0.387 | 0.41 | 0.42 | 0.422 |
| 13. Weibull(1) | 0.44 | 0.45 | 0.461 | 0.37 | 0.38 | 0.382 | 0.36 | 0.37 | 0.370 |
| 14. Weibull(4) | 0.19 | 0.19 | 0.195 | 0.20 | 0.21 | 0.207 | 0.14 | 0.14 | 0.140 |
| 15. Dagum* | 0.33 | 0.35 | 0.260 | 0.32 | 0.33 | 0.388 | 0.27 | 0.28 | 0.271 |

*The shape and scale parameters for the Dagum distribution are 4.273, 0.36 and 14.28.

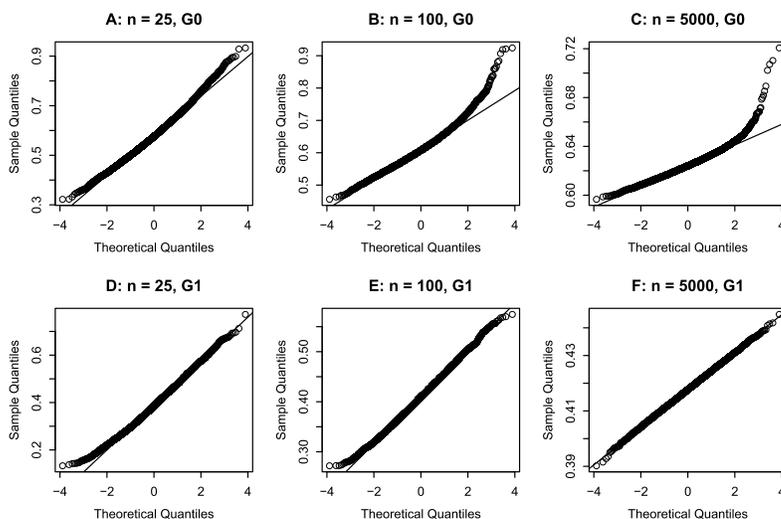


FIG 7. Normal Q-Q plots for 10,000 simulated \hat{G}_0 s and \hat{G}_1 s when n observations are sampled from the Pareto(2.5) distribution.

observations sample from the Pareto(2.5) distribution. For G_0 (top row) we see that although the estimator appears approximately normal for small $n = 25$, the estimator diverges from normality for $n = 100$ and even $n = 5000$. We also explored this further and even for $n = 50,000$ a skew was evident. However, for $n = 100,000$ the \hat{G}_0 was approximately normal once more. The problem was far greater for the Pareto(2) and not as bad for a shape parameter of four. This was

not the case for the estimator of G_1 where approximate normality was achieved quickly (bottom row). These results show that further research into the large sample properties of \hat{G}_0 are required.

3.5. Distribution-free confidence intervals for the quantile coefficients of inequality

Recall from (7) that for each $i = 1, 2, 3$ and large fixed J the estimated coefficient of inequality is $\hat{G}_i = (2/J) \sum_j \{p_j - \hat{L}_i(p_j)\}$. Now the estimate $\hat{L}_i(p_j)$, as a ratio of finite linear combinations of quantile estimates, is consistent for $L_i(p_j)$, so \hat{G}_i is also consistent for G_i . Further, [38] show that $n^{1/2}\{\hat{L}_i(p_j) - L_i(p_j)\}$ is asymptotically normal with mean 0 and variance depending on certain quantiles and quantile densities of the underlying F . The limiting *joint* normal distribution of estimates of a finite number of Lorenz curve ordinates are found by [3], assuming that $F \in \mathcal{F}_0 \cap \mathcal{F}'$, where \mathcal{F}' is specified in Definition 6. In the same way, for $F \in \mathcal{F}'$, the limiting joint normal distribution of the estimated ordinates $\hat{L}_i(p_j)$, $j = 1, \dots, J$ can be established. We do not need an analytic expression for the covariance matrix, because we only require the asymptotic normality of the estimated G_i , which being an average of the $p_j - \hat{L}_i(p_j)$, is immediate.

Large sample confidence intervals for G_i of the form $\hat{G}_i \pm 1.96 \times \hat{\sigma}_i$ are possible given a good estimate of the variance $\hat{\sigma}_i^2 = \text{Var}[\hat{G}_i]$. An efficient estimator is:

$$\hat{\sigma}_i^2 = \frac{4}{J^2} \sum_{j=1}^J \left\{ \widehat{\text{Var}} \left[\hat{L}_i(p_j) \right] + 2 \sum_{r < j} \widehat{\text{Cov}} \left[\hat{L}_i(p_r), \hat{L}_i(p_j) \right] \right\}. \tag{8}$$

Asymptotic variances and covariances for quantile estimators (see, for e.g. [13, Ch.7]), are given by $n \widehat{\text{Var}}(\hat{x}_p) \doteq p(1-p)g^2(p)$ and for $p < q$, $n \widehat{\text{Cov}}(\hat{x}_p, \hat{x}_q) \doteq p(1-q)g(p)g(q)$. Here $g(p) = 1/f(x_p)$ is the quantile density [36]. We estimate $g(p)$ directly using a kernel density estimator. Specifically, we used

$$\hat{g}(p; b) = \sum_{i=1}^n x_{(i)} \left\{ k_b \left(p - \frac{(i-1)}{n} \right) - k_b \left(p - \frac{i}{n} \right) \right\}, \tag{9}$$

where k is the [18] kernel, $k_b(\cdot) = k(\cdot/b)/b$, and b the bandwidth based on the quantile optimality ratio in [39]. Earlier work on this kernel estimator is due to [19], [50] and [31].

Next we obtain approximate variances and covariances of $\hat{L}_1(p_j)$. The results associated with $\hat{L}_2(p)$ are similarly obtained and only slightly more complicated. This may also be done for $\hat{L}_3(p)$, but it is much more complicated so we do not pursue intervals for L_3 further. For $p \leq q$ let $\sigma(p, q) = p(1-q)g(p)g(q)$ and $\sigma_p^2 = \sigma(p, p)$. Then, using the Delta method, approximations to the variances and covariances associated with estimation of L_1 are

$$\widehat{\text{Var}} \left[\hat{L}_1(p_j) \right] \approx \frac{1}{n} \cdot \frac{1}{x_{0.5}^2} \left[p_j^2 \sigma^2(p_j/2) + L_1^2(p_j) \sigma^2(0.5) \right]$$

TABLE 5
Distribution-free confidence intervals for G_1 and G_2 : Empirical coverage probabilities (first row) and widths (second row) based on 10,000 simulations of nominal 95% confidence intervals.

| Dist. | G_1 | | | G_2 | | |
|---------------|-----------|-----------|------------|-----------|-----------|------------|
| | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 100$ | $n = 500$ | $n = 1000$ |
| Uniform | 0.950 | 0.951 | 0.952 | 0.945 | 0.947 | 0.948 |
| | 0.165 | 0.074 | 0.052 | 0.156 | 0.069 | 0.049 |
| Chisq(0.5) | 0.996 | 0.987 | 0.980 | 0.990 | 0.983 | 0.976 |
| | 0.343 | 0.123 | 0.082 | 0.222 | 0.081 | 0.054 |
| Chisq(1) | 0.979 | 0.962 | 0.959 | 0.973 | 0.963 | 0.956 |
| | 0.236 | 0.099 | 0.069 | 0.179 | 0.074 | 0.052 |
| Chisq(3) | 0.964 | 0.956 | 0.958 | 0.962 | 0.958 | 0.955 |
| | 0.170 | 0.073 | 0.051 | 0.150 | 0.064 | 0.045 |
| Chisq(5) | 0.964 | 0.963 | 0.952 | 0.961 | 0.958 | 0.955 |
| | 0.144 | 0.062 | 0.043 | 0.134 | 0.057 | 0.040 |
| Lognormal | 0.968 | 0.962 | 0.957 | 0.969 | 0.958 | 0.956 |
| | 0.181 | 0.077 | 0.053 | 0.158 | 0.065 | 0.045 |
| Pareto(0.5) | 0.990 | 0.980 | 0.971 | 0.996 | 0.983 | 0.977 |
| | 0.335 | 0.117 | 0.079 | 0.391 | 0.083 | 0.055 |
| Pareto(1) | 0.978 | 0.965 | 0.960 | 0.983 | 0.968 | 0.964 |
| | 0.243 | 0.097 | 0.067 | 0.201 | 0.074 | 0.051 |
| Pareto(1.5) | 0.977 | 0.962 | 0.960 | 0.979 | 0.963 | 0.963 |
| | 0.222 | 0.091 | 0.063 | 0.177 | 0.072 | 0.049 |
| Pareto(2) | 0.974 | 0.959 | 0.956 | 0.974 | 0.963 | 0.960 |
| | 0.213 | 0.089 | 0.062 | 0.172 | 0.071 | 0.049 |
| Weibull(0.25) | 0.999 | 0.999 | 0.997 | 0.999 | 0.998 | 0.995 |
| | 0.586 | 0.156 | 0.097 | 0.339 | 0.092 | 0.058 |
| Weibull(0.5) | 0.991 | 0.972 | 0.968 | 0.989 | 0.972 | 0.968 |
| | 0.275 | 0.107 | 0.074 | 0.194 | 0.075 | 0.052 |
| Weibull(1) | 0.964 | 0.955 | 0.958 | 0.965 | 0.958 | 0.959 |
| | 0.191 | 0.083 | 0.058 | 0.160 | 0.069 | 0.048 |
| Weibull(4) | 0.973 | 0.966 | 0.964 | 0.969 | 0.964 | 0.961 |
| | 0.087 | 0.037 | 0.025 | 0.092 | 0.039 | 0.027 |
| Dagum* | 0.963 | 0.9589 | 0.957 | 0.959 | 0.958 | 0.954 |
| | 0.144 | 0.0629 | 0.044 | 0.137 | 0.059 | 0.041 |
| Grouped** | 0.965 | 0.959 | 0.953 | 0.964 | 0.954 | 0.954 |
| | 0.176 | 0.075 | 0.053 | 0.152 | 0.065 | 0.045 |

* The shape and scale parameters for the Dagum distribution were chosen to be 4.273, 0.36 and 14.28 respectively.

** Data simulated from the linearly interpolated distribution depicted in Figure 4.

$$\widehat{\text{Cov}} \left[\hat{L}_1(p_j), \hat{L}_1(p_r) \right] \approx \frac{1}{n} \cdot \frac{1}{x_{0.5}^2} \left[p_j p_r \sigma(p_j/2, p_r/2) + L_1(p_j) L_1(p_r) \sigma^2(0.5) - 2p_j L_1(p_j) \sigma(p_j/2, 0.5) - L_1(p_j) \sigma(p_r/2, 0.5) - L_1(p_r) \sigma(p_j/2, 0.5) \right]$$

and where we replace the unknown parameters with their respective estimates as described above. The resulting confidence interval can be computed efficiently and an R package is in development.

In Table 5 we present empirical coverage probabilities and widths for a variety of distributions. The confidence intervals are distribution-free in the sense that we compute asymptotic variances of quantiles as described earlier via di-

rect estimation of the quantile density function as in (9). We also sampled data from the interpolated distribution shown in Figure 4 and the results are found in the row labeled ‘Grouped’. A total of 10,000 simulation runs were used for each distribution and choice of sample size, with the focus on intervals for G_1 and G_2 . Even for $n = 100$, generally the coverage probability for the interval estimators of G_1 and G_2 are very good with a tendency to be conservative. Improved coverage is obtained for larger sample sizes. The interval widths are relatively stable across distributions and, as expected, decrease at the rate $1/\sqrt{n}$.

4. Robustness properties

In this section we show that the quantile inequality curves and their associated coefficients of inequality have bounded influence functions, which guarantees that a small amount of contamination can only have a limited effect on the asymptotic bias of estimators of these quantities. For background material on robustness concepts for functionals, see [26], although we attempt to make the presentation self-contained. We need to restrict $F \in \mathcal{F}$ to a smoother subclass:

Definition 6.

$$\mathcal{F}' = \{F \in \mathcal{F} : f = F' \text{ exists and is strictly positive.}\}$$

For $F \in \mathcal{F}'$ with inverse $x_p = Q(p) = F^{-1}(p)$, define the quantile density [47], [36] by

$$q(p) = \frac{\partial Q(F;p)}{\partial p} = \frac{1}{F'(Q(F;p))} = \frac{1}{f(x_p)}. \tag{10}$$

We also require the mixture distribution which places positive probability ϵ the point z (the contamination point) and $1 - \epsilon$ on the income distribution F . Formally, it is defined for each x by $F_\epsilon^{(z)}(x) \equiv (1 - \epsilon)F(x) + \epsilon I[x \geq z]$, where $I[\cdot]$ is the indicator function. The influence function for any functional T is then defined for each z as the $IF(z; T, F) \equiv \lim_{\epsilon \downarrow 0} \{T(F_\epsilon^{(z)}) - T(F)\} / \epsilon = \frac{\partial}{\partial \epsilon} T(F_\epsilon^{(z)})|_{\epsilon=0}$. The influence function of the p th quantile functional $T(F) = Q(F;p)$, where $F \in \mathcal{F}'$ of Definition 6, is well-known to be [45, p.59]

$$IF(z) \equiv IF(z; Q(\cdot; p), F) = \begin{cases} (p - 1) q(p), & z < x_p; \\ 0, & z = x_p; \\ p q(p), & z > x_p. \end{cases} \tag{11}$$

where $x_p = F^{-1}(p)$ and $q(p)$ is given by (10). The influence function in (11) is often replaced by the more compact $IF(z; x_p, F) = \{p - I(z < x_p)\} q(p)$, which differs from it at only one negligible point.

It is well known that $E_F[IF(Z)] = 0$ and $\text{Var}_F[IF(Z)] = E_F[IF^2(Z)] = p(1 - p) q^2(p)$. For those not familiar with such calculations, note that $IF[Z] =$

$q(p)((p-1)I\{Z < x_p\} + pI\{Z > x_p\})$ for continuous F , so

$$\begin{aligned} E_F[\text{IF}(Z)] &= q(p)\{(p-1)F(x_p) + p(1-F(x_p))\} = 0 \\ E[\text{IF}^2(Z)] &= q^2(p)\{(p-1)^2F(x_p) + p^2(1-F(x_p))\} \\ &= q^2(p)\{p(p-1)^2 + p^2(1-p)\} = p(1-p)q^2(p). \end{aligned}$$

One reason for calculating this variance is that it arises in the asymptotic variance of the functional applied to the empirical distribution F_n , namely $Q(F_n; p)$. That is, $n^{1/2}[Q(F_n; p) - Q(F; p)] \rightarrow N(0, p(1-p)q^2(p))$ in distribution; and an expression for the asymptotic variance is not always otherwise available.

4.1. Influence functions of quantile inequality curves

[7] show that the influence function of the Lorenz curve at the point p is unbounded, implying that a small amount of contamination can lead to a large bias in estimation; on the other hand the quantile inequality curves have bounded influence functions, provided only that $F \in \mathcal{F}'$. To see this, note that each $T_i(F) = L_i(F; p) = px_{p/2}/d_i(p)$, where $d_1(p) = x_{1/2}$, $d_2(p) = x_{1-p/2}$ and $d_3(p) = (x_{p/2} + x_{1-p/2})/2$ are all quantile functionals or an average of them.

Proposition 2. *The influence function of the functional defined by $T_i(F) = L_i(F; p)$ is given in terms of other influence functions by:*

$$\text{IF}(z; T_i, F) = p \left\{ \frac{\text{IF}(z; x_{p/2}, F)}{d_i(p)} - \frac{x_{p/2}\text{IF}(z; d_i(p), F)}{d_i^2(p)} \right\}.$$

This formula is derived for fixed p by noting that the influence function of each $L_i(F; p)$ is a constant multiple p times the derivative of a ratio of two functionals, which by elementary calculus yields (12). The derivation is completed by substitution of the respective d_i and their influence functions. For d_1 it is $\text{IF}(z; d_1(p), F) = \text{IF}(z; x_{1/2}, F)$, obtained from (11), and similarly for $d_2(p)$. For d_3 , we utilize $\text{IF}(z; d_3(p), F) = \{\text{IF}(z; x_{p/2}, F) + \text{IF}(z; x_{1-p/2}, F)\}/2$.

While these influence functions look complicated, they are easy to compute and plot using currently available software; an R script for doing so is in the supplementary material. Specific examples are shown Figure 8 when the underlying $F = F_a$ is the Type II Pareto distribution with shape parameter $a = 1$ and are plotted as functions of a possible contamination at z . For this distribution $Q(p) = p/(1-p)$ and $q(p) = 1/(1-p)^2$. To help explain their behavior as p varies, we examine the influence function of $L_1(p) = px_{p/2}/x_{0.5}$ at contamination z :

$$\text{IF}(z; L_1(p)) = c_p \left[\left\{ \frac{p}{2} - I\left(z < \frac{p}{(2-p)}\right) \right\} - p(2-p) \left\{ \frac{1}{2} - I(z < 1) \right\} \right] \quad (12)$$

where $c_p = 4p/(2-p)^2$. The expression in square brackets has maximum absolute value 1 for $z < 1$ and $1/8$ for $z \geq 1$, so for all $0 < p < 1$ the absolute influence on $L_i(F; p)$ of contamination is bounded by c_p for $z < 1$ and $c_p/8$

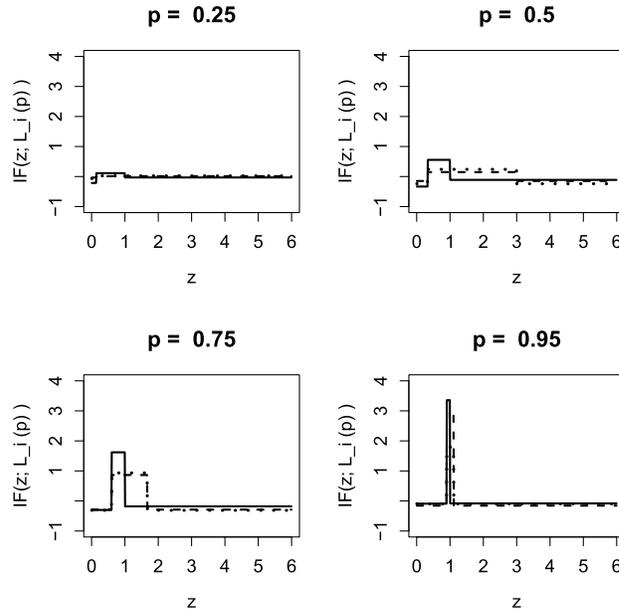


FIG 8. For various choices of p , $IF(z; L_i(p), F_1)$ is plotted as a function of z ; the solid, dashed and dotted lines correspond, respectively, to $i = 1, 2$ and 3 .

for $z \geq 1$, which explains why the upper left hand plot of Figure 8 shows small influence for all z . For larger values of p , as z increases to the median 1, the maximum influence approaches a peak; it then drops to a small negative and constant influence again as z increases past the median. This is to be expected, because when the median is pulled to the left by contamination, then $L_1(F; p) = p x_{p/2} / x_{0.5}$ is increased, but when the median is pulled to the right, the values of $L_1(F; p)$ are decreased. The maximum influence approaches 4 as $p \rightarrow 1$. The other two $L_i(F; p)$ are similarly affected by contamination at z , but to a lesser extent.

Plots of the influence functions of the quantile inequality curves for other Pareto(a) distributions (not shown) are similar to those in Figure 8, and again the peak is located at the median $F_a^{-1}(0.5) = 2^{1/a} - 1$. Similar influence function plots (not shown) were obtained for uniform, lognormal and Weibull distributions, again with finite peaks near their respective medians.

4.2. Influence of contamination at on the graph $\{p, L_i(p)\}$

We have found, for each fixed $0 < p < 1$, the influence functions $IF(z; L_i(p), F)$. Now we consider, for fixed z , the graph $\{(p, IF(z; L_i(p), F))\}$, which shows the influence of contamination at z on the respective inequality curves $\{(p, L_i(p))\}$. Examples are shown in Figure 9, again for F the Pareto ($a = 1$) distribution, and selected values of z .

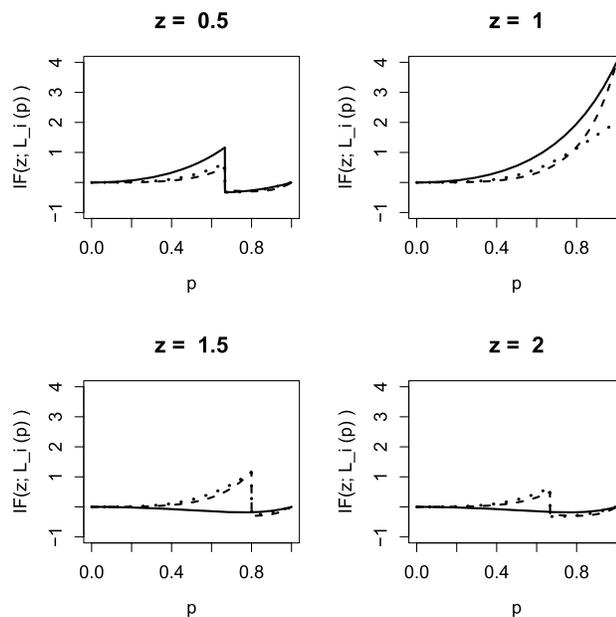


FIG 9. For various choices of z , $IF(z; L_i(p), F_1)$ is plotted as a function of p .

First we concentrate on only the solid lines corresponding to $L_1(p)$. Inspection of (12) shows that the discontinuity points are $x_{1/2} = 1$ and $x_{p/2} = p/(2-p)$. Now $z < x_{p/2}$ if and only if $p > 2F_1(z) = 2z/(1+z)$. Thus in the upper left plot of Figure 9 where $z = 0.5 < x_{1/2}$ there are only two cases of interest: $p < 2F_1(0.5) = 2/3$ and $p > 2/3$; in the first interval $(0, 2/3)$ the influence of contamination at $z = 0.5$ on the L_1 -curve is positive and increasing in p , but its influence is negative for p in $(2/3, 1)$.

For the top right plot $2F_1(z) = 1$ so the influence of contamination $z = 1$ at the median on the L_1 -curve is positive and increasing for all p .

For the other two plots z exceeds the median 1 and the influence function (12) reduces to $IF(z; L_1(p)) = 2(p-1)\{p/(2-p)\}^2$ which is not only free of $z > 1$ but negative for all p with a minimum -0.18 .

The influence of contamination at z on the graphs of $L_2(p)$, $L_3(p)$ is also shown in Figure 9 as dashed and dotted lines, respectively. Such influence is similar to that on $L_1(p)$ in the top two plots where z does not exceed the median. But in the lower plots where z exceeds the median, the contamination is positive and increasing on the interval $(0, 2(1-F_1(z)))$ and negative for larger p . After substituting z into $(0, 2(1-F_1(z)))$ where F_1 equals the Pareto(1) distribution, for the bottom left plot this interval is $(0, 0.952)$, and for the bottom right it is $(0, 0.8)$. One can see that increasing the values of z only diminishes its effect of contamination on the graphs of L_2 and L_3 .

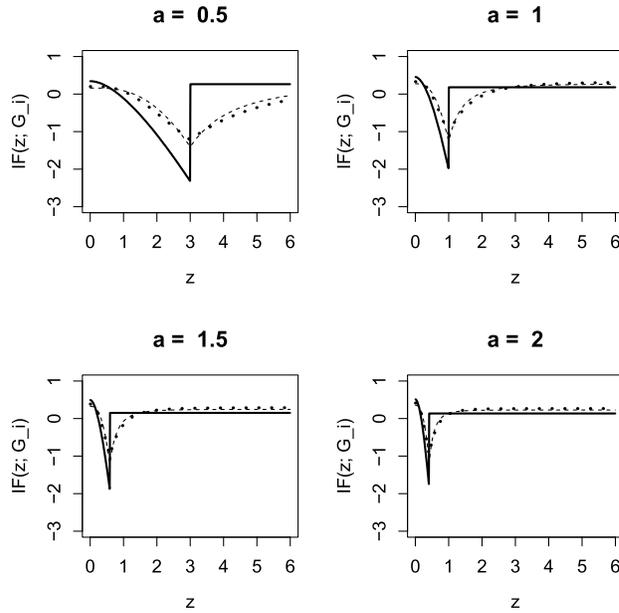


FIG 10. The solid, dashed and dotted lines correspond, respectively, to the influence functions of G_i for Pareto(a) distributions for $i = 1, 2$ and 3 .

4.3. Influence functions of quantile coefficients of inequality

The influence functions of the inequality coefficients associated with the L_i -curves are easily found, because the functional $G_i(F) = 1 - 2 \int_0^1 L_i(F; p) dp$, which contains an average of $L_i(F; p)$ values over $p \in (0, 1)$.

Proposition 3. For each $i = 1, 2$ and 3 the influence function of the inequality coefficients G_i are given respectively by

$$IF(z; G_i, F) = -2 \int_0^1 IF(z; L_i(\cdot; p), F) dp . \tag{13}$$

One only needs to justify taking the derivative $G_i(F_\epsilon^{(z)})$ with respect to ϵ at $\epsilon = 0$ under the integral sign. An argument based on the Leibniz Integration Rule is given in the Appendix.

Figure 10 gives plots of the influence functions $IF(z; G_i, F_a) = -2 \int_0^1 IF(z; L_i(\cdot; p), F_a) dp$ of the inequality coefficients $G_i(F_a)$ when F_a is the Pareto(a) distribution for selected values of a . The biggest influence of contamination occurs at $z = F_a^{-1}(0.5) = 2^{1/a} - 1$.

The mean and variance of $IF(z; G_i, F)$ are given by

$$E_F[IF(Z; G_i, F)] = -2 \int_0^1 E[IF(Z; L_1(\cdot; p), F)] dp = 0$$

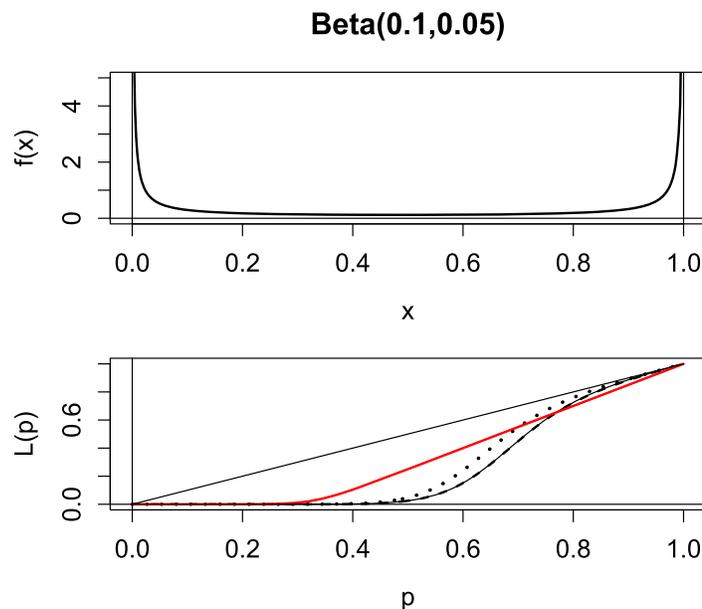


FIG 11. The top plot shows the density of the Beta(0.1,0.05) distribution. Below it are the corresponding L_i curves. The solid, dashed and dotted lines correspond, respectively, to $i = 1, 2$ and 3 ; The red line is the Lorenz curve.

$$\text{Var}_F[\text{IF}(Z; G_i, F)] = 4 \text{E} \left[\left\{ \int_0^1 \text{IF}(Z; L_i(\cdot; p), F) dp \right\}^2 \right]. \quad (14)$$

These quantities are easy to compute numerically; examples of the asymptotic standard error $\text{SE}[\hat{G}_i] = \{\text{Var}_F[\text{IF}(Z; G_1, F)]\}^{1/2}$ determined by (14) are shown in Table 4.

5. Convexity of the quantile inequality curves

One of the nice mathematical properties of the Lorenz curve $\{p, L_0(F; p)\}$ is that it is convex for all distributions $F \in \mathcal{F}_0$. The quantile-based versions (1) are defined for all F in the larger class \mathcal{F} , but need not be convex. In particular, empirical versions are often not convex over $(0, 1)$. The following examples demonstrate that for the more commonly assumed income distributions, the quantile inequality curves are convex. See [29, 30] for background material on these distributions.

5.1. Non-convex example

Figure 11 shows that for the very U-shaped Beta distribution with parameters (0.1, 0.05) only the Lorenz curve is convex. This distribution appears to have

TABLE 6

Examples of distributions $F(x)$ and associated quantile functions and their densities. In general, we denote $x_p = Q(p) = F^{-1}(p)$, but for the normal $F = \Phi$ with density φ , we write $z_p = \Phi^{-1}(p)$. the support of each F is $(0, +\infty)$, except for the normal and Type I Pareto, the latter having support on $[1, +\infty)$.

| | $1 - F(x)$ | $Q(p)$ | $q(p)$ |
|-----------------------|---|------------------------------|---|
| Exponential | e^{-x} | $-\ln(1 - p)$ | $(1 - p)^{-1}$ |
| Normal | $\Phi(-x)$ | z_p | $\frac{1}{\varphi(z_p)}$ |
| Lognormal | $\Phi(-\ln(x))$ | e^{z_p} | $\frac{e^{z_p}}{\varphi(z_p)}$ |
| Type I Pareto(a) | x^{-a} | $\frac{1}{(1-p)^{1/a}}$ | $\frac{1}{a(1-p)^{1/a+1}}$ |
| Type II Pareto(a) | $(1 + x)^{-a}$ | $\frac{1}{(1-p)^{1/a}} - 1$ | $\frac{1}{a(1-p)^{1/a+1}}$ |
| Weibull(β) | e^{-x^β} | $\{-\ln(1 - p)\}^{1/\beta}$ | $\frac{\{-\ln(1-p)\}^{1/\beta-1}}{\beta(1-p)}$ |
| Dagum(a_1, a_2) | $1 - \left\{1 + \left(\frac{x}{b}\right)^{-a_1}\right\}^{-a_2}$ | $b(p^{-1/a_2} - 1)^{-1/a_1}$ | $\frac{[u^{-1/a_2} - 1]^{-1/a_1} u^{-1/a_2}}{a_1 a_2 (u^{-1/a_2} - 1)^{1+1/a_1}}$ |

a symmetric density, but in fact is quite asymmetric, with mean $2/3$, and the quartiles $0.050, 0.997$, and 1.000 , to three decimal places. The inequality coefficients are $G_0 = 0.329$, $G_1 = 0.453$, $G_2 = 0.455$ and $G_3 = 0.403$. Note that the Gini coefficient $G_0 < 1/3$, its value for the uniform distribution, a non-intuitive result to us.

Other plots, not shown, for parameters $(0.05, 0.1)$, $(0.1, 0.1)$ and $(0.05, 0.05)$ indicate that all four L_i curves are convex.

5.2. Convex examples

Example 1. Uniform.

Starting with $Q(p) \equiv p$, we find $L_1(p) = p^2 = L_3(p)$ and $L_2(p) = p^2/(2 - p)$, all clearly convex functions of p in $(0, 1)$.

Example 2. Exponential.

Here $Q(p) = -\ln(1 - p)$, so $L_1(p) = -p \ln(1 - p/2) / \ln(2)$ where $L_1''(p) = (4 - p) / [(p - 2)^2 \ln(2)] > 0$. Similarly, $L_2(p) = p \ln(1 - p/2) / \ln(p/2)$ and $L_3(p) = 2p \ln(1 - p/2) / \ln[p(1 - p/2)/2]$ and it is not difficult to show that both $L_2''(p) > 0$ and $L_3''(p) > 0$ so that $L_1(p)$, $L_2(p)$ and $L_3(p)$ are all convex.

Example 3. Lognormal.

It is ‘obvious’ from the lower left plot in Figure 1 that all three $L_i(p)$ curves are convex on $(0, 1)$ for the lognormal distribution. Proving it using the calculus is not as straightforward as one might expect. Note that $Q(p) = e^{z_p}$,

$q(p) = e^{z_p}/\varphi(z_p)$. Further, observe that $L_1(p) = p \exp(z_{p/2})$ and that $\exp(z_{p/2})$ is not convex, so one cannot use the fact that two monotone increasing convex functions is convex. Taking derivatives,

$$\begin{aligned} L_1'(p) &= L_1(p) \left\{ \frac{1}{p} + \frac{1}{2\varphi(z_{p/2})} \right\} \\ L_1''(p) &= L_1(p) \left[\left\{ \frac{1}{p} + \frac{1}{2\varphi(z_{p/2})} \right\}^2 - \frac{1}{p^2} - \frac{\varphi'(z_{p/2})}{4\varphi^3(z_{p/2})} \right] \\ &= L_1(p) \left[\frac{1}{p\varphi(z_{p/2})} + \frac{1+z_{p/2}}{4\varphi^2(z_{p/2})} \right]. \end{aligned}$$

Thus $L_1''(p) > 0$ if and only if $4\varphi(z_{p/2}) + p(1+z_{p/2}) > 0$ and this again, while obvious from a plot, is not readily verified.

Next consider $L_2(p) = p \{ \exp(z_{p/2}) \exp(-z_{1-p/2}) \} = p \exp(2z_{p/2})$. The argument is very similar to that for L_1 :

$$\begin{aligned} L_2'(p) &= L_2(p) \left\{ \frac{1}{p} + \frac{1}{\varphi(z_{p/2})} \right\} \\ L_2''(p) &= L_2(p) \left[\left\{ \frac{1}{p} + \frac{1}{\varphi(z_{p/2})} \right\}^2 - \frac{1}{p^2} - \frac{\varphi'(z_{p/2})}{2\varphi^3(z_{p/2})} \right] \\ &= L_2(p) \left[\frac{2}{p\varphi(z_{p/2})} + \frac{2+z_{p/2}}{2\varphi^2(z_{p/2})} \right]. \end{aligned}$$

Thus $L_2''(p) > 0$ if and only if $4\varphi(z_{p/2}) + p(2+z_{p/2}) > 0$, a weaker condition than required for convexity of L_1 .

Finally, consider $L_3(p) = 2p/\{1+p/L_2(p)\} = 2p/\{1+\exp(-2z_{p/2})\}$. It suffices to show that $h(p) = 1/\{1+\exp(-2z_{p/2})\}$ is convex in p and this is readily verified.

Example 4. Type I Pareto.

For the Type I Pareto(a) distribution where $a > 0$, $Q(p) = (1-p)^{-1/a}$. Let $c_1 = (2-p)^{-1/a}/a$ which is positive. Then $L_1''(p) = c_1[(1-p/2)^{-1} + (1+1/a)p/(p-2)^2] > 0$ so that $L_1(p)$ is convex. Similarly, $L_2''(p) = c_1 p^{1/a}(1+1/a)[(1-p/2)^{-1} + p/(p-2)^2 + 1/p] > 0$ so that $L_2(p)$ is also convex. The expression for $L_3''(p)$ is much more complicated although plots and computational minimization reveal that convexity holds. For example, over all $p \in [0, 1)$ and $a \in (0, 10]$, $\min L_3''(p) = 0.169$ (at $p = 0.667$ and $a = 10$).

Example 5. Type II Pareto.

For the Type II Pareto(a) distribution where $a > 0$, $Q(p) = (1-p)^{-1/a} - 1$. We then have that

$$L_1''(p) = \frac{(1-p/2)^{-1/2}}{a^2(p-2)^2(2^{1/a}-1)} [p + a(4-p)] > 0$$

so that $L_1(p)$ is convex. Both $L_2''(p)$ and $L_3''(p)$ are complicated expressions although computational minimization reveals non-negative minimums over all p and $a \in (0, 10]$.

Example 6. Weibull.

For the Weibull distribution with shape parameter $\beta > 0$, we have

$$L_1''(p) = \frac{\ln(2)^{-1/\beta}}{\beta(p-2)^2} \ln\left(\frac{2}{2-p}\right)^{1/\beta-1} \left[4 - p - p \ln\left(\frac{2}{2-p}\right)^{-1} + \frac{p}{\beta} \ln\left(\frac{2}{2-p}\right)^{-1} \right].$$

The term $-p \ln(2/(2-p))$ is a decreasing function in p with limit equal to -2 as p approaches 0. Consequently, $L_1''(p) > 0$ so that $L_1(p)$ is convex. For $L_2(p)$ and $L_3(p)$, again we used computational minimization for all β values up to 100. Neither had a negative minimum so both were found to be convex.

Example 7. Dagum (Type I)

The Dagum Type I distribution described in Table 6 has two shape parameters, $a_1 > 0$ and $a_2 > 0$, and scale parameter $b > 0$. The second derivative of $L_1(p)$ can be written

$$L_1''(p) = \frac{1}{p} L_1(p) \left[a_1^2 a_2^2 p \left(p^{1/a_2} - 2^{1/a_2} \right)^2 \right]^{-1} \left[2^{1/a_2} + a_1 a_2 \left(2^{1/a_2} - p^{1/a_2} \right) \right]$$

so that $L_1''(p) > 0$ since $(2^{1/a_2} - p^{1/a_2}) > 0$. Consequently, L_1 is convex.

6. Summary and further research

We have shown that quantile versions of the Lorenz curve have most of the properties of the original definition, with two exceptions. The first exception is convexity, which is not satisfied for some very U-shaped distributions and many empirical ones. Nevertheless, for most distributions used to model population incomes, the quantile versions are convex. It would be highly desirable to find simple necessary and sufficient conditions in terms of the underlying income distribution for convexity of the quantile inequality curves. The second exception is the first order transference principle, which is mean-preserving. When replaced by a median-preserving definition, this principle is satisfied for all three quantile versions of the Lorenz curve. It would be of interest to explore whether the median-preserving definition has parallel results to the Fellman-Jakobsson Theorem and related results [20], [28], [17] and [21]. We illustrated the quantile methods on two transfer functions, a percentage levy and a fixed levy, and their effects for the grouped data model of [35] and the Pareto model, respectively.

The quantile versions of the Lorenz curve possess several advantages over the traditional measures. They are defined for *all* positive income distributions,

and their influence functions are bounded, while the influence functions of the traditional ones are not. This means that the quantile versions are more resilient in the presence of outliers.

In addition, we showed that the standard errors of estimates for the quantile analogues of the Gini coefficient do not depend much on the underlying income model, so that sample sizes can be chosen in advance to obtain desired standard errors. Simulation studies show that these sample inequality coefficients approach normality very rapidly, and reliable distribution-free confidence intervals for the inequality coefficients can be constructed for them. Along the way, we demonstrated that the standard estimators of the Gini coefficient are quite sensitive to the underlying model, and do not always approach normality nearly so rapidly as their quantile cousins, even when the underlying population has a finite variance.

Confidence bands for quantile versions of the Lorenz curve could utilize functionals of the quantile process, starting with the results in [16] and [10]. Applications to other fields which use diversity indices [37] are possible, as well as links to the ‘Lorenz dominance’ literature, see [2].

Appendix: Proof of Proposition 3

The interchange of limit (as $\epsilon \downarrow 0$) and integral is justified by the Leibniz Integral Rule. It requires that $h_i(p) \equiv \text{IF}(z; L_i(\cdot; p), F)$ be continuous in p , and bounded in absolute value for $p \in (0, 1)$ by an integrable function.

Proof for $i = 1$.

For L_1 , we have from Proposition 2 that

$$\begin{aligned} |h_1(p)| &\leq \frac{p}{x_{1/2}^2} \{x_{1/2} |\text{IF}(z; Q(\cdot, p/2), F)| + x_{p/2} |\text{IF}(z; Q(\cdot; 1/2), F)|\} \\ &\leq \frac{p}{x_{1/2}^2} \left\{ x_{1/2} \max\{p/2, 1 - p/2\} q(p/2) + \frac{x_{p/2} q(1/2)}{2} \right\}. \end{aligned}$$

The second term is bounded because $p Q(p/2) \leq x_{1/2}$ for $p \in (0, 1)$; and, for the first term we require only that $p q(p/2)$ be integrable on $(0, 1)$. By making the change of variable $x = F^{-1}(p/2)$ in $\int_0^1 p q(p/2) dp$ one finds that this integral is bounded by $4x_{1/2}$. Therefore $|h_1(p)|$ is bounded by an integrable function on $(0, 1)$, justifying (13) for L_1 .

Proof for $i = 2$.

For $L_2(p) = p x_{p/2} / x_{1-p/2}$ we have

$$\begin{aligned} h_2(p) &\equiv \frac{p}{x_{1-p/2}^2} \{x_{1-p/2} \text{IF}(z; Q(\cdot, p/2), F) - x_{p/2} \text{IF}(z; Q(\cdot; 1 - p/2), F)\}, \text{ so} \\ |h_2(p)| &\leq \frac{p q(p/2)}{x_{1-p/2}} + \frac{p x_{p/2} q(1 - p/2)}{x_{1-p/2}^2}. \end{aligned} \tag{15}$$

The first term in the last line of (15) is bounded above by $p q(p/2)/x_{1/2}$, and it has already been shown that $p q(p/2)$ was integrable on $(0,1)$.

Next we show that the second term is bounded by an integrable function. Let $m = x_{1/2}$ and make the change of variable $x = F^{-1}(1 - p/2) = x_{1-p/2}$ to obtain:

$$\begin{aligned} \int_0^1 \frac{p x_{p/2} q(1-p/2)}{x_{1-p/2}^2} dp &= 4 \int_m^\infty \frac{\{1-F(x)\} F^{-1}(1-F(x))}{x^2} dx \\ &\leq 4m \int_m^\infty \frac{dx}{x^2} = 4. \end{aligned} \quad (16)$$

This shows that $h_2(p) = \text{IF}(z; L_2(\cdot; p), F)$ is bounded on $(0, 1)$ by an integrable function.

Proof for $i = 3$.

Let $m(p) = (x_{p/2} + x_{1-p/2})/2$, so $m(1) = m$ is the median, and $L_3(p) = p x_{p/2}/m(p)$. It is immediate that $\text{IF}(z; m(p), F) = \{\text{IF}(z; Q(\cdot, p/2), F) + \text{IF}(z; Q(\cdot, 1-p/2), F)\}/2$ and that $|\text{IF}(z; m(p), F)| \leq \{q(p/2) + q(1-p/2)\}/2$.

Consider bounding $h_3(p) = \text{IF}(z; L_3(\cdot; p), F)$ by an integrable function.

$$\begin{aligned} h_3(p) &\equiv \frac{p}{m^2(p)} \{m(p) \text{IF}(z; Q(\cdot, p/2), F) - x_{p/2} \text{IF}(z; m(p), F)\}, \text{ so} \\ |h_3(p)| &\leq \frac{p q(p/2)}{m(p)} + \frac{p x_{p/2} \{q(p/2) + q(1-p/2)\}}{2m^2(p)}. \end{aligned} \quad (17)$$

The first term $p q(p/2)/m(p) \leq 2p q(p/2)/x_{1-p/2}$, which has already been shown to be integrable. The third term $p x_{p/2} q(1-p/2)/(2m^2(p)) \leq 2p x_{p/2} q(1-p/2)/x_{1-p/2}^2$, shown to be integrable in (16). The second term $p x_{p/2} q(p/2)/(2m^2(p)) \leq p q(p/2)/x_{1-p/2}$, using the fact that $m^2(p) \geq x_{p/2} x_{1-p/2}$. Therefore $|h_3(p)|$ is bounded by an integrable function.

Acknowledgements

The authors thank the Editors and two referees who made numerous helpful suggestions that enlarged the scope and improved the clarity of this manuscript.

References

- [1] AABERGE, R., & ATKINSON, A.B. 2013. *The median as watershed*. Statistics Norway Discussion Paper 749. University of Norway.
- [2] AABERGE, R., & MOGSTAD, M. 2011. Robust inequality measures. *J. Econ. Inequal.*, **9**(3), 353–371.
- [3] BEACH, C.M., & DAVIDSON, R. 1983. Distribution-free statistical inference with Lorenz curves and income shares. *Rev. Econ. Studies*, **L**, 723–735.
- [4] BROWN, B.M. 1981. Symmetric quantile averages and related estimators. *Biometrika*, **68**(1), 235–242. [MR0614960](#)

- [5] BUREAU, U. S. CENSUS. 2014. *Historical income tables: Households*. <http://www.census.gov/hhes/www/income/data/historical/household>.
- [6] COWELL, F.A., & VICTORIA-FESER, M.P. 1996. Robustness properties of inequality measures. *Econometrica*, **64**(1), 77–101.
- [7] COWELL, F.A., & VICTORIA-FESER, M.P. 2002. Welfare rankings in the presence of contaminated data. *Econometrica*, **70**(3), 1221–1233. [MR1910417](#)
- [8] COWELL, F.A., & VICTORIA-FESER, M.P. 2003. Distribution-free inference for welfare indices under complete and incomplete information. *J. Econ. Inequal.*, **1**(3), 191–219.
- [9] COWELL, F.A., & VICTORIA-FESER, M.P. 2007. Robust stochastic dominance: A semi-parametric approach. *J. Econ. Inequal.*, **5**(1), 21–37.
- [10] CSÖRGŐ, M., GASTWIRTH, J.L., & ZITIKIS, R. 1998. Asymptotic confidence bands for the Lorenz and Bonferroni curves based on the empirical Lorenz curve. *J. Statist. Plan. & Inference*, **74**(1), 65:91. [MR1665121](#)
- [11] DAGUM, C. 1977. A new model of personal income distribution: Specification and estimation. *Economie Appliquée*, **30**, 413–437.
- [12] DALTON, H. 1920. The measurement of the inequality of incomes. *Economic Journal*, **30**, 348–361.
- [13] DASGUPTA, A. 2006. *Asymptotic Theory of Statistics and Probability*. Springer. DOI: [10.1007/978-0-387-75971-5](https://doi.org/10.1007/978-0-387-75971-5).
- [14] DAVIDSON, R. 2008. Reliable inference for the Gini index. *J. Econ.*, **150**, 30–40. [MR2525992](#)
- [15] DEVELOPMENT CORE TEAM, R. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- [16] DOSS, H., & GILL, R.D. 1992. An elementary approach to weak convergence for quantile processes. *J. Amer. Statist. Assoc.*, **87**, 869–877. [MR1185204](#)
- [17] EGGHE, L. 2009. The theorem of Fellman and Jakobsson: A new proof and dual theory. *Mathematical and Computer Modelling*, **50**, 1595–1605. [MR2556919](#)
- [18] EPANECHNIKOV, V.A. 1969. Nonparametric estimation of a multivariate probability density. *Theory Probab. Appl.*, **14**, 153–158.
- [19] FALK, M. 1986. On the estimation of the quantile density function. *Statist. Prob. Letters*, **4**, 69–73. [MR0829436](#)
- [20] FELLMAN, J. 1976. The effect of transformations of Lorenz curves. *Econometrica*, **44**, 823–824. [MR0469205](#)
- [21] FELLMAN, J. 2012. Properties of Lorenz curves for transformed income distributions. *Theor. Econ. Letters*, **2**, 487–493.
- [22] GAIL, M. H., & GASTWIRTH, J.L. 1978. A Scale-Free Goodness-of-Fit Test for the Exponential Distribution Based on the Gini Statistic. *J. Roy. Statist. Soc. B*, **40**(3). [MR0522218](#)
- [23] GASTWIRTH, J.L. 1971. A general definition of the Lorenz curve. *Econometrica*, **39**, 1037–1039.

- [24] GASTWIRTH, J.L. 2014. Median-based measures of inequality: Reassessing the increase in income inequality in the U.S. and Sweden. *Statistical Journal of the International Association for Official Statistics*, **30**(4), 311–320. DOI: [10.3233/SJI-140842](https://doi.org/10.3233/SJI-140842).
- [25] GINI, C. 1914. Sulla misura della concentrazione e della variabilit'a dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, **73**, 1203–1248. English translation (2005) in *Metron* Vol. 63, pp. 3–38. [MR2200970](https://doi.org/10.1017/S0026351605000000)
- [26] HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., & STAHEL, W.A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons. [MR0829458](https://doi.org/10.1002/9781118133000)
- [27] HYNDMAN, R.J., & FAN, Y. 1996. Sample quantiles in statistical packages. *The American Statistician*, **50**, 361–365.
- [28] JAKOBSSON, U. 1976. On the measurement of the degree of progression. *Journal of Public Economics*, **5**, 161–168.
- [29] JOHNSON, N.L., KOTZ, S., & BALAKRISHNAN, N. 1994. *Continuous Univariate Distributions*. Vol. 1. New York: John Wiley & Sons. [MR1299979](https://doi.org/10.1002/9781118133000)
- [30] JOHNSON, N.L., KOTZ, S., & BALAKRISHNAN, N. 1995. *Continuous Univariate Distributions*. Vol. 2. New York: John Wiley & Sons. [MR1326603](https://doi.org/10.1002/9781118133000)
- [31] JONES, M.C. 1992. Estimating densities, quantiles, quantile densities and density quantiles. *Ann. Inst. Statist. Math.*, **44**(4), 721–727.
- [32] KAMPKE, T. 2010. The use of mean values vs. medians in inequality analysis. *J. Econ. Social Meas.*, **35**, 43–62.
- [33] KLEIBER, C. 2005. *The Lorenz curve in Economics and Econometrics*. Technical Report 30. University of Dortmund, SFB 475.
- [34] KLEIBER, C. 2008. *Modeling Income Distributions and Lorenz Curves*. New York, NY: Springer. Chap. A Guide to the Dagum Distributions, pages 97–117.
- [35] LYON, M., CHEUNG, L. C., & GASTWIRTH, J. L. 2016. The advantages of using group means in estimating the Lorenz curve and Gini index from grouped data. *The Amer. Statist.*, **70**(1), 25–32. [MR3480667](https://doi.org/10.1080/01621459.2016.1191111)
- [36] PARZEN, E. 1979. Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.*, **7**, 105–131. [MR0529528](https://doi.org/10.1080/01621459.1979.10489444)
- [37] PATIL, G.P., & TAILLIE, C. 1982. Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.*, **77**, 548–561. [MR0675883](https://doi.org/10.1080/01621459.1982.10489444)
- [38] PRENDERGAST, L.A., & STAUDTE, R.G. 2015. When large n is not enough-Distribution-free Interval Estimators for Ratios of Quantiles. *arXiv preprint arXiv:1508.06321v2*.
- [39] PRENDERGAST, L.A., & STAUDTE, R.G. 2016. Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. *Stat*, **5**, 70–81. DOI: [10.1002/sta4.105](https://doi.org/10.1002/sta4.105).
- [40] SEN, P.K. 1986. The Gini coefficient and poverty indexes: some reconciliations. *J. Amer. Statist. Assoc.*, **81**, 1050–1057. [MR0867631](https://doi.org/10.1080/01621459.1986.10489444)
- [41] SHEATHER, S.J., & MARRON, J.S. 1990. Kernel quantile estimators. *J. Amer. Statist. Assoc.*, **85**, 410–416. [MR1141741](https://doi.org/10.1080/01621459.1990.10489444)

- [42] STAUDTE, R.G. 2013. Distribution-free confidence intervals for the standardized median. *STAT*, **2**(1), 184–196.
- [43] STAUDTE, R.G. 2014. Inference for quantile measures of skewness. *Test*, **23**(4), 751–768. [MR3274473](#)
- [44] STAUDTE, R.G. 2016. Inference for quantile measures of kurtosis, peakedness and tail-weight. *Comm. Statist. A: Theory and Methods*. DOI: [10.1080/03610926.2015.1056366](#).
- [45] STAUDTE, R.G., & SHEATHER, SIMON J. 1990. *Robust Estimation and Testing*. New York: Wiley. [MR1061581](#)
- [46] TILLÉ, Y., & LANGEL, M. 2012. Histogram-based interpolation of the Lorenz Curve and Gini Index for grouped data. *Am. Stat.*, **66**, 225–231.
- [47] TUKEY, J.W. 1965. Which part of the sample contains the information? *Proc, Math. Acad.Sci. USA*, **53**, 127–134. [MR0172387](#)
- [48] VICTORIA-FESER, M.P. 2000. Robust methods for the analysis of income distribution, inequality and poverty. *Int. Statist. Rev.*, **68**(3), 277–293.
- [49] VICTORIA-FESER, M.P., & RONCHETTI, E. 1994. Robust methods for personal-income distribution models. *Canadian J. Statist.*, **22**(2), 247–258. [MR1295691](#)
- [50] WELSH, A.H. 1988. Asymptotically efficient estimation of the sparsity function at a point. *Statistic. Prob. Letters*, **6**, 427–432. [MR0942656](#)