Income Share Standard Errors and a Quantile Toolbox of Distributional Statistics*

by

Charles M. Beach

Department of Economics
Queen's University
Kingston. Ontario, Canada
K7L 3N6

email: **beach.chaz3@gmail.com**

and

Russell Davidson

Department of Economics and CIREQ
McGill University
Montreal, Quebec, Canada
H3A 2T7

Aix-Marseille Université
CNRS, EHESS, AMSE
13205 Marseille cedex 01, France

email: **russell.davidson@mcgill.ca**

Abstract

This paper uses a stochastic quantile function approach to derive the (asymptotic) variances and covariances – and hence standard errors – of quantile means and income shares in terms of explicit formulas that are distribution-free and easily computable. The paper then develops a toolbox of quantile-based disaggregative income inequality measures, based on the means and income shares, which allow for detailed inferential analysis of income distributions in a straightforward unified framework. The analytical formulas are applied with Canadian Census public-use microdata files, and the results highlight the statistical significance of some quite distinct distributional changes in decile earnings in Canada between 2000 and 2005.

Key words: quantile function, income shares, distribution-free inference, disaggregative measures

JEL codes: C10, C42

* This paper is dedicated to the memory of Aidan Worswick, research assistant to both authors, and a truly fine person, who died recently at far too young an age.

February 2024

1. Introduction

Since about 1980 and into the current century, income inequality in many developed economies rose dramatically to historic levels (Guvenen *et al.* (2022)) driven by advances in automation, globalization and shifting production patterns, and long-run demographic forces. The resulting labour market effects were experienced quite differently by different groups of workers over different regions of the income distribution (*e.g.*, Acemoglu *et al.*, (2016); Autor, Dorn and Hanson (2013); Beach (2016); and Goos, Manning and Salomons (2014)). More specifically, so-called middle-class workers in the labour market lost out substantially to higher-skilled and top-income recipients. More recently, the advent of the covid pandemic (and resulting policy responses), inflation and higher interest rates, and wars in Ukraine and the Middle East are eliciting major economic adjustments to international trade patterns, natural resource and high-tech investments and supply chains. Again, these changes affect different groups of the workforce and different regions of the income/earnings distributions in ways that can only be analyzed in a quite disaggregated fashion. Perhaps not surprisingly, economists have been showing growing interest in distributional National Accounts within statistical agencies, greater heterogeneity of agent behaviour in macro model development, and a greater focus in the media on attaining equitable growth and common prosperity (*The Economist*, (2021a,b)).

This paper seeks to provide statistical tools that can help dealing with these concerns. More specifically, the paper (i) establishes the (asymptotic) variance-covariance structure – and hence standard errors – of quantile means and income shares in terms of explicit easily-computable formulas, and (ii) forwards a toolbox of quantile-based disaggregative income inequality measures along with their statistical properties so as to allow for detailed inferential analysis of such inequality measures. In doing so, it helps to simplify and unify income distribution analysis in a common statistical framework that involves explicit variance-covariance formulas that are distribution-free and can be directly estimated without density estimation or burdensome computational procedures.

This set of disaggregative statistical tools can be applied well beyond traditional income inequality analysis. They can be applied to all of income, earnings, wages, and wealth distributions. They can better allow for detailed comparisons of, say, male *vs* female earnings distributions and how they may have changed over time, or of age, regional, industry, or occupational earnings differences as well. Such a set of disaggregative statistical tools makes better use of the torrent of microdata bases such as large Census files that have become available over the last several decades. The paper is also written in the spirit of Cowell (2011), Lambert (2001), and Jenkins (1999) of expanding the broad set of statistical tools available to general empirical practitioners in the income distribution field.

The paper proceeds as follows. The next section uses a stochastic quantile function approach to work out formulas for (asymptotic) variances and covariances — and hence standard errors — of quantile mean statistics where a distribution is divided up into a set of K ordered quantile income groups. Corresponding results for the income shares of these quantile (e.g., decile) groups are derived in Section 3 and Appendix 1. Section 4 presents a toolbox or set of disaggregative distributional statistics on income inequality and polarization all related to the quantile means and income share statistics and sets out

their corresponding standard error formulas of each. An empirical illustration of the use of these quantile toolbox statistics based on Canadian Census earnings data appears in Section 5. The paper then concludes with several implications of the results of the analysis.

2. Quantile Function Approach for Quantile Means

In this section we develop what we may call the stochastic quantile function approach. Some of the analysis in this section is based to some extent on material in Davidson (2018).

Consider first some formal concepts and notation. Suppose the distribution of income Y is divided into K ordered income groups. Let the dividing proportions of recipients be $p_1 < p_2 < \ldots < p_{K-1}$ (with $p_0 = 0$ and $p_K = 1$). Then in terms of the underlying (population) distribution F of income recipients, the mean income of recipients with incomes between the p_{i-1} and p_{i-1} quantiles, $i = 1, \ldots, K$, is given by

$$\mu_i = \int_{\xi_{i-1}}^{\xi_i} y \, \mathrm{d}F(y) \Big/ \int_{\xi_{i-1}}^{\xi_i} \mathrm{d}F(y), \tag{1}$$

where ξ_i is the p_i quantile of the distribution F, and ξ_0 is taken to be the smallest (possibly negative) income in the support of the distribution. Note that the proportions p_i do not need to be equally spaced. They could for instance be more refined at one or both ends of the distribution.

Suppose we have a random sample, y_j, $j = 1, \ldots, N$, drawn from the population characterised by F. The empirical distribution function (EDF) of the sample is

$$\hat{F}(y) = \frac{1}{N} \sum_{j=1}^{N} \mathrm{I}(y_j \leq y),$$

where I is the indicator function, with value 1 if its argument is true, and 0 otherwise. Natural estimators of the numerator and denominator of (1) are

$$\int_{\hat{\xi}_{i-1}}^{\hat{\xi}_i} y \, \mathrm{d}\hat{F}(y) \quad \text{and} \quad \int_{\hat{\xi}_{i-1}}^{\hat{\xi}_i} \mathrm{d}\hat{F}(y), \tag{2}$$

respectively, where $\hat{\xi}_i$ is the p_i quantile of the empirical distribution. The estimator of the denominator is $\hat{F}(\hat{\xi}_i) - \hat{F}(\hat{\xi}_{i-1})$, which is just $p_i - p_{i-1}$, since $\hat{F}(\hat{\xi}_i) = p_i$ by the definition of the sample quantiles. The estimator is of course non-random, and can therefore be ignored for the purpose of deriving an asymptotic expression for the estimator of the numerator. Denote by n_i the numerator of μ_i in (1), so that $n_i = \mu_i(p_i - p_{i-1})$, and by $\hat{n}_i$ the estimator given in (2), which can be rewritten as

$$\int_{\hat{\xi}_{i-1}}^{\hat{\xi}_i} y \, \mathrm{d}\hat{F}(y) = \int_{\xi_{i-1}}^{\xi_i} y \, \mathrm{d}\hat{F}(y) + \int_{\hat{\xi}_{i-1}}^{\xi_{i-1}} y \, \mathrm{d}\hat{F}(y) + \int_{\xi_i}^{\hat{\xi}_i} y \, \mathrm{d}\hat{F}(y). \tag{3}$$

Note that this formulation captures the randomness both of the sample quantiles ξ_i and of the EDF itself, without any need for restrictive assumptions on the population distribution function F. Under the usual assumption that the $\hat{\xi}_i$ and $\hat{F}$ are root-n consistent, we see that

$$\int_{\xi_i}^{\hat{\xi}_i} y \, \mathrm{d}\hat{F}(y) = \xi_i\big(\hat{F}(\hat{\xi}_i) - \hat{F}(\xi_i)\big) = \xi_i(p_i - \hat{F}(\xi_i)) + O_p(N^{-1}),$$

and similarly

$$\int_{\hat{\xi}_{i-1}}^{\xi_{i-1}} y \, \mathrm{d}\hat{F}(y) = -\xi_{i-1}(p_{i-1} - \hat{F}(\xi_{i-1})) + O_p(N^{-1}).$$

With this, (3) becomes, to leading order asymptotically,

$$p_i\xi_i - p_{i-1}\xi_{i-1} + N^{-1}\sum_{j=1}^{N}\big[y_j\mathrm{I}(\xi_{i-1} < y_j \le \xi_i) - \xi_i\mathrm{I}(y_j \le \xi_i) + \xi_{i-1}\mathrm{I}(y_j \le \xi_{i-1})\big]. \qquad (4)$$

Let $w_{ij} = y_j\mathrm{I}(\xi_{i-1} < y_j \le \xi_i) - \xi_i\mathrm{I}(y_j \le \xi_i) + \xi_{i-1}\mathrm{I}(y_j \le \xi_{i-1})$, and denote by W_i the random variable of which the w_{ij} are IID drawings, that is,

$$W_i = Y\mathrm{I}(\xi_{i-1} < Y \le \xi_i)) - \xi_i\mathrm{I}(Y \le \xi_i) + \xi_{i-1}\mathrm{I}(Y \le \xi_{i-1}), \qquad (5)$$

where Y denotes the random variable of which the y_j are IID drawings. Then, since

$$\mathrm{E}\big(Y\mathrm{I}(\xi_{i-1} < Y \le \xi_i)\big) = \int_{\xi_{i-1}}^{\xi_i} y \, \mathrm{d}F(y), \ \mathrm{E}\big(\mathrm{I}(Y \le x_{i-1})\big) = p_{i-1}, \text{ and } \mathrm{E}\big(\mathrm{I}(Y \le \xi)\big) = p_i,$$

it follows that

$$\mathrm{E}(W_i) = n_i - p_i\xi_i + p_{i-1}\xi_{i-1}. \qquad (6)$$

Then the expression (4), which is asymptotically equal to $\hat{n}_i$, has an expectation of n_i if terms of order N^{-1} are neglected, so that (4) is a root-n consistent estimator of n_i.

The next task is to derive the asymptotic variance of (4). But, before doing so explicitly, note that the variance can be estimated by the sample variance of the w_{ij}, $j = 1, \ldots, N$, in a distribution-free manner. However, it is not difficult to obtain an analytic expression for the asymptotic variance, which can in turn be estimated. We have

$$\begin{aligned} W_i^2 = Y^2\mathrm{I}(\xi_{i-1} < Y \le \xi_i) + \xi_i^2\mathrm{I}(Y \le \xi_i) + \xi_{i-1}^2\mathrm{I}(Y \le \xi_{i-1}) \\ -2Y\xi_i\mathrm{I}(\xi_{i-1} < Y \le \xi_i) - 2\xi_{i-1}\xi_i\mathrm{I}(Y \le \xi_{i-1}) \end{aligned},$$

whence

$$\mathrm{E}(W_i^2) = n_{2i} + p_i\xi_i^2 + p_{i-1}\xi_{i-1}^2 - 2\xi_i n_i - 2p_{i-1}\xi_{i-1}\xi_i,$$

where we make the definition $n_{2i} = \int_{\xi_{i-1}}^{\xi_i} y^2 \, \mathrm{d}F(y)$. Thus the variance of W_i is

$$\begin{aligned} \mathrm{Var}(W_i) = \mathrm{E}(W_i^2) - \big(\mathrm{E}(W_i)\big)^2 = n_{2i} - n_i^2 + \xi_i^2 p_i(1-p_i) + \xi_{i-1}^2 p_{i-1}(1-p_{i-1}) \\ -2\xi_{i-1}\xi_i p_{i-1}(1-p_i) - 2n_i\big(\xi_i(1-p_i) + \xi_{i-1}p_{i-1}\big). \end{aligned} \qquad (7)$$

This expression can be estimated in an obvious manner from the sample information. Specifically,

$$\begin{aligned}\widehat{\mathrm{Var}}(W_i) = \hat{n}_{2i} - \hat{n}_i^2 + \hat{\xi}_i^2 p_i(1-p_i) + \hat{\xi}_{i-1}^2 p_{i-1}(1-p_{i-1}) \\ - 2\hat{\xi}_{i-1}\hat{\xi}_i p_{i-1}(1-p_i) - 2\hat{n}_i\big(\hat{\xi}_i(1-p_i) + \hat{\xi}_{i-1}p_{i-1}\big),\end{aligned} \tag{8}$$

where

$$\hat{n}_i = \int_{\hat{\xi}_{i-1}}^{\hat{\xi}_i} y \, \mathrm{d}\hat{F}(y) = N^{-1} \sum_{\hat{\xi}_{i-1} < y_j \le \hat{\xi}_i} y_j \quad \text{and} \tag{9}$$

$$\hat{n}_{2i} = \int_{\hat{\xi}_{i-1}}^{\hat{\xi}_i} y^2 \, \mathrm{d}\hat{F}(y) = N^{-1} \sum_{\hat{\xi}_{i-1} < y_j \le \hat{\xi}_i} y_j^2. \tag{10}$$

Now, since $\hat{n}_i$ is asymptotically equal to expression (4), of which the variance is the variance of W_i divided by N, it follows that the variance of $\hat{n}_i$ can also be estimated by the right-hand side of (8) over N. Standard errors are then given by the square roots of the estimated variances. Since $\mu_i = n_i/(p_i - p_{i-1})$, we can estimate μ_i by $\hat{\mu}_i = \hat{n}_i/(p_i - p_{i-1})$, and the standard errors of the $\hat{\mu}_i$ are just those of the $\hat{n}_i$ divided by $p_i - p_{i-1}$.

Covariances

For some purposes we may need not only estimates of the variances of the $\hat{\mu}_i$ but also their covariances. For $j < i$, compute as follows:

$$\begin{aligned}W_i W_j &= \big[Y\mathrm{I}(\xi_{i-1} < Y \le \xi_i) - \xi_i \mathrm{I}(Y \le \xi_i) + \xi_{i-1}\mathrm{I}(Y \le \xi_{i-1})\big] \\ &\qquad \big[Y\mathrm{I}(\xi_{j-1} < Y \le \xi_j) - \xi_j \mathrm{I}(Y \le \xi_j) + \xi_{j-1}\mathrm{I}(Y \le \xi_{j-1})\big] \\ &= -Y(\xi_i - \xi_{i-1})\mathrm{I}(\xi_{j-1} < Y \le \xi_j) + \xi_j(\xi_i - \xi_{i-1})\mathrm{I}(Y \le \xi_j) - \xi_{j-1}(\xi_i - \xi_{i-1})\mathrm{I}(Y \le \xi_{j-1}) \\ &= (\xi_i - \xi_{i-1})\big[\xi_j \mathrm{I}(Y \le \xi_j) - \xi_{j-1}\mathrm{I}(Y \le \xi_{j-1}) - Y\mathrm{I}(\xi_{j-1} < Y \le \xi_j)\big].\end{aligned}$$

Then

$$\mathrm{E}(W_i W_j) = (\xi_i - \xi_{i-1})(p_j\xi_j - p_{j-1}\xi_{j-1} - n_j),$$

and, since $\mathrm{cov}(W_i, W_j) = \mathrm{E}(W_i W_j) - \mathrm{E}(W_i)\mathrm{E}(W_j)$, from this we see that, for $j < i$,

$$\begin{aligned}\mathrm{cov}(W_i, W_j) = (\xi_i - \xi_{i-1})(p_j\xi_j - p_{j-1}\xi_{j-1} - n_j) \\ - (n_j - p_j\xi_j + p_{j-1}\xi_{j-1})(n_i - p_i\xi_i + p_{i-1}\xi_{i-1}).\end{aligned} \tag{11}$$

Consequently, $\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_j)$ is asymptotically equal to the above expression divided by $N(p_i - p_{i-1})(p_j - p_{j-1})$. As with the variances, the covariances can be estimated in an obvious distribution-free manner, replacing the quantiles ξ_i in the expression (11) by their estimates $\hat{\xi}_i$ and the n_i by their estimates $\hat{n}_i$ as in (9). Appendix 2 provides some evidence from simulations that the estimated variances for the quantile means and income shares are indeed very reliable and show no apparent biases.

3. Quantile Function Approach for Income Shares

The income share that accrues to recipients with incomes between ξ_{i-1} and ξ_i is

$$IS_i = \frac{1}{\mu}\int_{\xi_{i-1}}^{\xi_i} y\,\mathrm{d}F(y) = (p_i - p_{i-1})\mu_i/\mu = n_i/\mu, \tag{12}$$

where $\mu = \int_0^\infty y\,\mathrm{d}F(y)$ is the mean income of the population characterised by the distribution F. The natural estimator of IS_i is

$$\widehat{IS_i} = \frac{1}{\hat\mu}\int_{\hat\xi_{i-1}}^{\hat\xi_i} y\,\mathrm{d}\hat F(y).$$

By an asymptotic argument like those used above, it can be seen that

$$\widehat{IS_i} - IS_i = \frac{1}{\mu^2}\Big[\mu\int_{\hat\xi_{i-1}}^{\hat\xi_i} y\,\mathrm{d}\hat F(y) - \hat\mu\int_{\xi_{i-1}}^{\xi_i} y\,\mathrm{d}F(y)\Big] + O_p(N^{-1}). \tag{13}$$

and so, asymptotically,

$$\widehat{IS_i} - IS_i = \frac{1}{\mu}\hat n_i - \frac{\hat\mu}{\mu^2}n_i, \tag{14}$$

It is immediate that the expectation of the right-hand side above is zero, thereby showing that $\widehat{IS_i}$ is a root-n consistent estimator of IS_i. By use of (4), we see that the right-hand side of (14) is

$$\begin{aligned}&\frac{1}{\mu}\Big(p_i\xi_i - p_{i-1}\xi_{i-1} + N^{-1}\sum_{j=1}^N w_{ij} - \frac{n_i}{\mu}N^{-1}\sum_{j=1}^N y_j\Big)\\ &= \frac{1}{\mu}\Big(p_i\xi_i - p_{i-1}\xi_{i-1} + N^{-1}\sum_{j=1}^N\big(w_{ij} - n_i y_j/\mu\big)\Big),\end{aligned}$$

and its variance is the variance of the random variable $\mu^{-1}(W_i - n_iY/\mu)$, divided by N. Make the definitions

$$m_i = \int_{\xi_0}^{\xi_i} y\;\mathrm{d}F(y) = \sum_{k=1}^{i} n_k \quad\text{and}\quad m_{2i} = \int_{\xi_0}^{\xi_i} y^2\;\mathrm{d}F(y) = \sum_{k=1}^{i} n_{2k}. \tag{15}$$

Note that $\mu = m_K$. The variance of W_i was given in (7). The variance of Y is

$$\mathrm{Var}(Y) \equiv \sigma^2 = \int_{\xi_0}^{\xi_K}(y-\mu)^2\;\mathrm{d}F(y) = m_{2K} - \mu^2 = m_{2K} - (m_K)^2, \tag{16}$$

and the covariance of W_i and Y is $\mathrm{E}\big(W_i(Y-\mu)\big) = \mathrm{E}(W_iY) - \mu\mathrm{E}(W_i)$. Now, from (5),

$$\begin{aligned}\mathrm{E}(W_iY) &= \mathrm{E}\big(Y^2\mathrm{I}(\xi_{i-1} < Y \le \xi_i)\big) - \xi_i\mathrm{E}\big(Y\mathrm{I}(Y \le \xi_i)\big) + \xi_{i-1}\mathrm{E}\big(Y\mathrm{I}(Y \le \xi_{i-1})\big) \\ &= n_{2i} - \xi_i m_i + \xi_{i-1}m_{i-1},\end{aligned}$$

whence

$$\mathrm{cov}(W_i, Y) = n_{2i} - \mu n_i - \xi_i(m_i - p_i\mu) + \xi_{i-1}(m_{i-1} - p_{i-1}\mu). \tag{17}$$

As before, distribution-free estimates are readily obtained for $\mathrm{Var}(W_i)$, $\mathrm{Var}(Y)$, and $\mathrm{cov}(W_i, Y)$. Finally, the asymptotic variance of $\widehat{IS}_i$ is

$$N^{-1}\,\mathrm{Var}\Big(\mu^{-1}(W_i - n_iY/\mu)\Big) = \frac{1}{N\mu^4}\Big(\mu^2\,\mathrm{Var}(W_i) + n_i^2\,\mathrm{Var}(Y) - 2\mu n_i\,\mathrm{cov}(W_i, Y)\Big), \tag{18}$$

and it is estimated by use of the estimates of the variances and the covariance. Details of this and also the covariances of the $\widehat{IS}_i$ are given in Appendix 1.

4. A Quantile Toolbox of Distributional Measures

The above analytical development suggests a statistical toolbox of distributional statistics that are quantile-based and hence share the property of having an (asymptotic) variance-covariance structure that is distribution-free. Indeed, a number of disaggregative inequality measures can be obtained from the quantile means and income share statistics. These statistics provide the basis for useful computer programs to describe and assess the statistical reliability of detailed distributional change over time or distributional differences between population groups.

In order to estimate the variance-covariance results of the previous sections, one should initially calculate some basic preliminary distributional statistics. These include standard estimates of the summary parameters μ, σ, and σ/μ, and the quantile cut-off income levels $\xi_1, \ldots, \xi_{K-1}$ for the K ordered income groups.[1] The quantile toolbox then consists of three sets of quantile-based statistics and their corresponding standard errors: statistics about income shares, those about quantile means, and those about income gap and income polarization measures.

[1] Since the asymptotic variances and covariances of the $\hat{\xi}_i$ depend on the density $f(\xi_i)$ and hence are not distribution-free (Wilks (1962) p. 273), these variance-covariance estimates are not included in the quantile toolbox.

Income Share-Related Statistics

Formulas for estimating the (asymptotic) variances and covariances – and implied standard errors – of the income shares $\widehat{IS}_i$ are presented in Section 3 above and in Appendix 1. But these can also form the basis for deriving corresponding expressions for some related frequently used statistics. The first is Lorenz curve ordinates. A frequently applied criterion for income inequality dominance is based on Lorenz curve comparisons (see, for example, Maasoumi (1998); Lambert (2001); and Aaberge (2000, 2001)). Atkinson, in his famous 1970 paper, stated and proved what has come to be known as the Lorenz curve dominance theorem. According to this theorem, for any summary inequality measures (such as the Gini coefficient) satisfying the distributional properties of symmetry, mean independence, population homogeneity, and the classic principle of transfers, if the Lorenz curve for distribution A lies everywhere above the Lorenz curve for distribution B, then all summary inequality measures satisfying these four properties will indicate that overall inequality in A is less than in B.

A Lorenz curve can be empirically represented by its set of quantile ordinates, which in turn are cumulated income shares. Thus the above formulas for the (asymptotic) variance-covariance structure of a set of sample income shares can be easily applied to a corresponding set of sample Lorenz curve ordinates, and these allow a formal statistical test for overall inequality dominance or ranking. One could obviously cumulate the income shares quantile-by-quantile and work out their corresponding (asymptotic) variances and standard errors. But the most direct approach is to simply define the i^{th} quantile Lorenz curve ordinate as that of the income share over the interval from ξ_0 up to ξ_i, and then plug the terms estimated from the sample into the formulas already presented.

Another concept frequently cited in the media and literature is the relative mean income (RMI) or ratio of a quantile mean to the overall mean of an income distribution:

$$\text{RMI}_i = \mu_i/\mu, \quad i = 1, \ldots, K.$$

The RMIs intuitively display the size of the gap or distance of quantile group mean incomes to the overall mean in proportional or percentage terms. So while the dollar gap values can change substantially over time, the underlying proportional gaps may change very little. Also, analogous to Atkinson's (1970) decomposition of an empirical measure of overall social welfare into efficiency and equity components (as represented by the mean and one minus the Atkinson inequality index, respectively), the RMI can serve as a way – at a disaggregated quantile level – to decompose individual quantile means into efficiency and equity components: $n_i = \mu\mu_i$. Growth in the individual n_i can be attributed to growth of incomes generally, as well as to the shifting of relative economic well-being among quantile groups. For example, one of the characteristic features in recent decades has been the dramatic rise in top incomes in many economies – a rising tide evidently does not raise all boats evenly.

From (12) it can be seen that

$$IS_i = (p_i - p_{i-1})\text{RMI}_i.$$

Consequently, RMI_i is simply a scalar transform of IS_i, so that the standard error of $\widehat{\text{RMI}}_i$ is the standard error of $\widehat{IS}_i$ divided by $p_i - p_{i-1}$. The RMIs thus provide a simple link or bridge between the primary concepts of the income shares and corresponding quantile means.

Quantile Mean-Related Measures

Quantile means are of interest not just as simple descriptive tools. A comparison of respective quantile means between two population groups, such as women and men, can serve as a basis for estimating the extent of potential earnings discrimination between the groups (Jenkins 1994; del Rio *et al.*, 2011). Salas *et al.* (2018) also show that a comparison of cumulative mean incomes up to a given quantile cut-off can serve as a basis for the measurement of second-order earnings discrimination between two population groups, which arises when the earnings distribution for one population group second-order stochastically dominates that for another population group. But the analytical results in Section 2 also provide the basis for statistical inference for several other related inequality measures.

One aspect of concern about rising income inequality is the implied growing economic and social distance between income groups, economic inclusion, and the potential political fracturing that this may bring about. The literature and media have focused on the widening gap between top incomes and the rest of the distribution, and the increasing difficulty of lower-income workers to pull ahead into stable middle-income status – the sense of belonging to the middle class may be weakening. One way to measure the growing economic distances between different income groups across a distribution may be in terms of a "distributional distance function" of mean differences between adjacent quantiles, $\hat{\mu}_i - \hat{\mu}_{i-1}$. These differences or income gaps can be viewed as successive steps on a ladder as one moves up a distribution. The wider the gaps, the greater the steps needed to advance up the distribution. Now, for inference we have

$$\text{Var}(\hat{\mu}_i - \hat{\mu}_{i-1}) = \text{Var}(\hat{\mu}_i) + \text{Var}(\hat{\mu}_{i-1}) - 2\,\text{cov}(\hat{\mu}_i, \hat{\mu}_{i-1}), \tag{19}$$

and asymptotic approximations to the three terms on the right-hand side of (19) are provided in Section 2. The standard error of the income gap then comes from these estimates.

A second related measure is the set of quantile income gaps relative to the overall mean, $|\hat{\mu}_i - \hat{\mu}|$. Here this measure can be viewed as representing how far away the incomes in each quantile group are from average incomes in the distribution. In this case, we have that

$$\text{Var}(\hat{\mu}_i - \hat{\mu}) = \text{Var}(\hat{\mu}_i) + \text{Var}(\hat{\mu}) - 2\,\text{cov}(\hat{\mu}_i, \hat{\mu}),$$

with $\text{Var}(\hat{\mu}) = \sigma^2$. Now $\hat{\mu} = \hat{m}_K = \sum_{j=1}^{K} \hat{n}_j$, and $\hat{\mu}_i = \hat{n}_i/(p_i - p_{i-1})$, so that

$$(p_i - p_{i-1})\,\text{cov}(\hat{\mu}_i, \hat{\mu}) = \sum_{j=1}^{K} \text{cov}(\hat{n}_i, \hat{n}_j) = \text{Var}(\hat{n}_i) + \sum_{j \neq i} \text{cov}(\hat{n}_i, \hat{n}_j). \tag{20}$$

Now recall from (8) and (11) that $\text{Var}(\hat{n}_i) = \text{Var}(W_i)$ and $\text{cov}(\hat{n}_i, \hat{n}_j) = \text{cov}(W_i, W_j)$. It follows that an asymptotic approximation to (20) can be computed using (8) and (11). Observe that the formula (11) is for $i < j$. Thus if $j < i$, the indices i and j must be interchanged in the formula. Again the standard error is just the square root of the estimated variance.

A further application of results for quantile means reflects growing concerns with equitable growth and common prosperity. The Economist (2021c), for example, refers to so-called Piketty lines of different growth rates of quantile means across the various regions of the income distribution. That is, one can look at

$$\hat{g}_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i0}}{\hat{\mu}_{i0}} = \frac{\hat{\mu}_{i1}}{\hat{\mu}_{i0}} - 1$$

for time periods 0 and 1. Then by the delta method (see Rao (1965)), we have approximately

$$\text{Var}(\hat{g}_i) = \text{Var}(\hat{\mu}_{i1})/\mu_{i0}^2 + \text{Var}(\hat{\mu}_{i0})\frac{\mu_{i1}^2}{\mu_{i0}^4},$$

under the assumption that $\hat{\mu}_{i0}$ and $\hat{\mu}_{i1}$ are uncorrelated. Again everything above can be estimated from the data in a distribution-free manner. Standard errors follow accordingly.

Income Polarization Statistics

An alternative way of characterizing the quantile income gap separating lower or higher incomes from middle incomes could serve as a measure of the degree of polarization or pulling apart in an income distribution. The intuitive concept of polarization can be viewed as having two quite distinct dimensions or aspects. One is the size dimension or the relative concentration of income recipients at the two ends of the distribution. This could be labelled tail frequency polarization. It could be captured, for example, by the proportion of recipients in the lower or higher income groups (Wolfson, 1994), say below half the median or above twice the median. The other is the distance dimension or the size of the income gap separating lower or upper incomes and middle-class incomes. This could be referred to as income polarization, and could be captured by, say, the gaps $\hat{\mu}_K - \hat{\mu}_M$ and $\hat{\mu}_M - \hat{\mu}_1$ where $\hat{\mu}_M$ is some measure of the middle income level. Both provide useful insights. However, the statistical properties of these two polarization measures are quite different. Tail frequency statistics are generally distribution-dependent (Davidson, 2018), while quantile-based mean income statistics are as we have seen distribution-free and hence their (asymptotic) variance-covariance structure is easy to estimate by direct formulas. For this reason, we focus on income polarization measures only.

It is useful, then, to consider what may be called an "income polarization function" $|\hat{\mu}_i - \hat{\mu}_M|$ over different quantile values of i and where $\hat{\mu}_M$ is the sample mean income in the middle quantile. This formulation has clear similarities to the Foster and Wolfson (1992) concept of a polarization curve presented in Kovacevic and Binder (1997. p. 50) as $B(p) = |\xi_p - \xi_M|/\xi_M$ where ξ_M is the median income and ξ_p is the income cut-off level for

quantile group p. In the case of deciles, say, let $\hat{\mu}_M = (0.5)(\hat{\mu}_5 + \hat{\mu}_6)$, and for vigintiles use $\hat{\mu}_M = (0.5)(\hat{\mu}_{10} + \hat{\mu}_{11})$. For quintiles, simply use $\hat{\mu}_M = \hat{\mu}_3$. Then in the case of deciles,

$$\mathrm{Var}(\hat{\mu}_i - \hat{\mu}_M) = \mathrm{Var}(\hat{\mu}_i) + \mathrm{Var}(\hat{\mu}_M) - 2\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_M) \text{ for } i = 1, \ldots, 4; 7, \ldots, 10,$$

where

$$\mathrm{Var}(\hat{\mu}_M) = 0.25\,\mathrm{Var}(\hat{\mu}_5) + 0.25\,\mathrm{Var}(\hat{\mu}_6) + 0.5\,\mathrm{cov}(\hat{\mu}_5, \hat{\mu}_6)$$

and

$$\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_M) = 0.5\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_5) + 0.5\,\mathrm{cov}(\hat{\mu}_i, \hat{\mu}_6).$$

Again, standard errors of $(\hat{\mu}_i - \hat{\mu}_M)$ are based on sample estimates of the above expressions.

The quantile toolbox of distributional statistics, then, consists of:

1. the basic preliminary statistics;
2. the estimated quantile means and income shares along with their estimated variances and covariances (*i.e.* the estimated asymptotic variances and covariances divided by N, the size of the sample used for estimation) and standard errors; and
3. the distributional statistics detailed in the subsections of this section along with their estimated standard errors. Standard "t-ratios" can also be reported corresponding to the standard errors of the various statistics. All of these statistics allow for distribution-free statistical inference procedures.

5. Empirical Illustration

In this section, we present results obtained using data from the Canadian Census Public Use Microdata Files (PUMF) for Individuals for 2000 and 2005, as recorded in the 2001 and 2006 censuses. Beach (2016) used data from the PUMF for several censuses since 1971, along with data from other sources, for his comprehensive account of the evolving fate of the Canadian middle class.

It is of interest to separate data for men and women, as their wages and labour-market participation rates are quite different. Accordingly, for each census year, two samples, one for each sex, are extracted from the census data files and are treated separately. In both cases, individuals younger than 15 years of age are dropped from the sample, as well as individuals who did not work in that year, or for whom the information on weeks worked is missing. In these files, the term earnings refers to annual earnings. Although income is split into wage income and income from self-employment, we simply combine them to yield the earnings variable. In many cases, incomes have been rounded to an integer multiple of \$1000, leading to a potential complication in the analysis, mentioned later. In all the results given in this section, earnings are expressed in thousands of 2005 (Canadian) dollars.

Basic information

In all cases, results are obtained for deciles. Table 1 below shows the basic results for women in 2005. Number of observations = 218250. The mean income of the sample is $\hat{\mu} = 30.077067$, sample variance $\hat{\sigma}^2 = 935.883906$, all measured in thousands of dollars.

p_i	$\hat{\xi}_i$	$\hat{\mu}_i$	$\widehat{\mathrm{RMI}}_i$	$\widehat{IS}_i$
0.1	4.000	1.884768	0.062665	0.006266
		(0.015773)	(0.000500)	(0.000050)
0.2	8.000	5.855212	0.194674	0.019467
		(0.029013)	(0.000880)	(0.000088)
0.3	12.000	10.035693	0.333666	0.033367
		(0.035601)	(0.001040)	(0.000104)
0.4	18.000	14.934158	0.496530	0.049653
		(0.058761)	(0.001600)	(0.000160)
0.5	24.000	20.456312	0.680130	0.068013
		(0.061880)	(0.00160))	(0.000160)
0.6	30.000	26.699885	0.887716	0.088772
		(0.061860)	(0.001610)	(0.000161)
0.7	37.000	33.145475	1.102018	0.110202
		(0.068618)	(0.001780)	(0.000178)
0.8	46.000	40.994777	1.362991	0.136299
		(0.079067)	(0.002060)	(0.000206)
0.9	61.000	52.641833	1.750232	0.175023
		(0.105270)	(0.002590)	(0.000259)
1.0		94.122554	3.129379	0.312938
		(0.397366)	(0.008700)	(0.000870)

Table 1: Women in 2005 (asymptotic standard errors in brackets)

The complication mentioned earlier arises on account of the fact that, with incomes rounded to integer multiples of \$100, \$500, or \$1000, there are many obervations with identical incomes. This in no way invalidates the results in sections 2 and 3, as the distribution-free calculations work equally well for continuous or discrete variables. But it does invalidate equations (9) and (10), at least as they are written. We want the number of observations in the quantile group between ξ_{i-1} and ξ_i to be equal to $N(p_i - p_{i-1})$, but the number of observations y_j such that $\hat{\xi}_{i-1} < y_j \leq \hat{\xi}_i$ can be very considerably different from $N(p_i - p_{i-1})$, precisely because there may be several observations equal to any given $\hat{\xi}_i$. Instead, the incomes should be sorted in increasing order. Then, for $i = 1, \ldots, K-1$, let $k_i = \lceil N(p_i - p_{i-1} \rceil$. The observations in the quantile group labelled i are the order

statistics from $y_{(k_{i-1}+1)}$ to $y_{(k_i)}$ inclusive, and $\hat{\xi}_i$ is set equal to $y_{(k_i)}$. In some cases, we found that use of (9) and (10) as written led to negative estimated variances. When the quantile groups are constructed as above, this cannot happen.

Results for men in 2005 are given in Table 2. Number of observations = 238350. Mean income of sample is $\hat{\mu} = 47.553074$; sample variance = $\hat{\sigma}^2 = 4803.637463$.

p_i	$\hat{\xi}_i$	$\hat{\mu}_i$	$\widehat{\mathrm{RMI}}_i$	$\widehat{IS}_i$
0.1	6.000	2.658557	0.055907	0.005591
		(0.023209)	(0.000480)	(0.000048)
0.2	12.000	8.691252	0.182770	0.018277
		(0.041749)	(0.000890)	(0.000089)
0.3	20.000	15.610111	0.328267	0.0328270
		(0.068173)	(0.001420)	(0.000142)
0.4	28.000	23.537151	0.494966	0.049497
		(0.075639)	(0.001680)	(0.000168)
0.5	35.000	31.483533	0.662071	0.066207
		(0.069189)	(0.001860)	(0.000186)
0.6	44.000	39.572981	0.832186	0.083219
		(0.088359)	(0.002270)	(0.000227)
0.7	53.000	48.485337	1.019605	0.101960
		(0.084788)	(0.002620)	(0.000262)
0.8	66.000	59.433732	1.249840	0.124984
		(0.109780)	(0.003120)	(0.000312)
0.9	87.000	75.171177	1.580785	0.158078
		(0.140133)	(0.003750)	(0.000375)
1.0		170.886913	3.593604	0.359360
		(1.171779)	(0.015240)	(0.001524)

Table 2: Men in 2005 (asymptotic standard errors in brackets)

Comparisons across sexes and across time

Because we will make comparisons across time, we have chosen to express all dollar amounts in thousands of constant 2005 Canadian dollars. This means simply that amounts for the year 2000 are multiplied by 1.11937, based on Statistics Canada's CPI series v41690973. Results analogous to those in the preceding subsection, but for the year 2000, are reported in Table A3 for women and in Table A4 for men, both in Appendix 3, using data from the 2001 Census. These results, combined with those in Tables 1 and 2, make it possible to conduct formal tests of a number of interesting hypotheses, all at a disaggregated quantile-group level.

First, a comparison of the outcomes for men and women. Tables 3 (for 2000) and 4 (for 2005) show, for each decile, the difference in quantile means, $\hat{\mu}_i$ for men minus $\hat{\mu}_i$ for women, along with the asymptotic t statistics for the hypotheses that these differences are zero, and also the same numbers for differences in income shares, IS_i.

p_i	diffce in $\hat{\mu}_i$	t-ratio for $\hat{\mu}_i$	$\widehat{IS}_i$	t-ratio for $\widehat{IS}_i$
0.1	0.729934	29.910532	-0.000104	-1.573142
0.2	3.161883	56.554601	0.000981	7.013191
0.3	6.435799	76.734036	0.003153	16.523288
0.4	9.802160	108.406476	0.004990	26.016292
0.5	11.886115	123.468649	0.002984	15.329323
0.6	13.534368	124.035908	-0.000450	-2.155213
0.7	15.206590	149.144450	-0.004331	-22.984861
0.8	18.265594	143.996406	-0.005826	-25.654844
0.9	22.022688	144.282527	-0.009859	-34.553494
1.0	43.959026	106.158246	0.008462	10.875604

Table 3: Differences in outcomes men *vs* women in 2000

Remarks:

Note that the orders of magnitude for the μ_i and the IS_i are different, since the μ_i are measured in thousands of dollars, while the income shares must add up to one. The t-ratios are dimensionless of course, but reveal important differences between the comparisons of the $\hat{\mu}_i$ and those of the $\widehat{IS}_i$. Unsurprisingly, men uniformly have higher decile mean incomes than women, with very significant differences even for the bottom decile. But the story is quite different for the income shares. In 2005, women have significantly larger income shares across all nine lower deciles, but for the top decile a dramatic reversal occurs with men having a significant larger income share.

p_i	diffce in $\hat{\mu}_i$	t-ratio for $\hat{\mu}_i$	$\widehat{IS}_i$	t-ratio for $\widehat{IS}_i$
0.1	0.773789	27.574851	-0.000676	-9.661025
0.2	2.836040	55.783648	-0.001190	-9.474489
0.3	5.574418	72.481036	-0.000540	-3.064506
0.4	8.602993	89.818780	-0.000156	-0.675446
0.5	11.027221	118.797680	-0.001806	-7.360500
0.6	12.873095	119.349066	-0.005553	-19.938694
0.7	15.339861	140.635727	-0.008241	-26.006552
0.8	18.438955	136.292613	-0.011315	-30.248492
0.9	22.529344	128.541567	-0.016945	-37.143363
1.0	76.764359	62.040705	0.046422	26.444959

Table 4: Differences in outcomes men *vs* women in 2005

The next two tables give the results of comparisons across time, from 2000 to 2005, for women in Table 5 and for men in Table 6.

p_i	diffce in $\hat{\mu}_i$	t-ratio for $\hat{\mu}_i$	$\widehat{IS}_i$	t-ratio for $\widehat{IS}_i$
0.1	0.422538	20.298319	0.000933	13.508549
0.2	0.654866	15.642643	0.000498	3.784233
0.3	0.366390	6.646151	-0.001903	-11.701750
0.4	0.356495	4.620082	-0.003521	-16.835039
0.5	0.348522	4.045508	-0.005333	-24.537802
0.6	0.754702	7.968999	-0.005867	-25.753603
0.7	0.962875	10.508515	-0.007188	-32.382247
0.8	1.843891	16.519533	-0.006509	-24.661344
0.9	3.191459	22.868528	-0.005353	-16.009639
1.0	17.717747	40.033925	0.034242	33.950916

Table 5: Differences in outcomes 2000 *vs* 2005 for women

p_i	diffce in $\hat{\mu}_i$	t-ratio for $\hat{\mu}_i$	$\widehat{IS}_i$	t-ratio for $\widehat{IS}_i$
0.1	0.466392	15.134383	0.000361	5.395841
0.2	0.329023	5.229941	-0.001673	-12.466862
0.3	-0.494991	-4.972270	-0.005596	-27.641097
0.4	-0.842672	-7.893728	-0.008668	-40.126566
0.5	-0.510372	-4.989776	-0.010122	-44.887654
0.6	0.093430	0.773989	-0.010970	-41.687486
0.7	1.096146	9.298306	-0.011098	-37.701233
0.8	2.017252	13.620515	-0.011997	-34.369073
0.9	3.698115	19.898222	-0.012438	-29.504212
1.0	50.523080	41.161870	0.072202	44.189154

Table 6: Differences in outcomes 2000 *vs* 2005 for men

Remarks:

Women clearly made gains in decile mean incomes in all deciles over the five-year period, whereas changes for men are mixed with losses over the lower-mid range and big gains at the top end. Quantile mean incomes show gains for all deciles above the median and for the bottom two deciles. However, income shares for all but the top and bottom two deciles fell. The deterioration in the fate of the middle deciles relative to the higher ones for men is quite evident in the results of table 6. Income shares fell for all but the bottom and top deciles. Lower income polarization measures (based on the sample covariances in appendix table A5) changed relatively little over the period for both women and men. But upper income polarization estimates rose quite considerably, especially among men.

One could also consider Piketty lines or the growth rates of earnings levels across decile groups. As table 7 shows, the growth rates of earnings between 2000 and 2005 are higher for women than for men for the nine lower deciles, but for the top decile group the growth of men's earnings is substantially larger than that for women. The asymptotic t statistics for the hypotheses that the growth rates are the same for both sexes reject these hypotheses at conventional significance levels, except for the p_i-quantile group for $i = 7$.

p_i	$\hat{g}_i$ women	$\hat{g}_i$ men	t-ratio for difference
0.1	0.288968	0.212754	3.416455
	(0.016117)	(0.015423)	
0.2	0.125927	0.039346	7.508514
	(0.008592)	(0.007690)	
0.3	0.037892	-0.030735	8.147466
	(0.005828)	(0.006081)	
0.4	0.024455	-0.034564	8.597427
	(0.005348)	(0.004304)	
0.5	0.017333	-0.015952	6.211744
	(0.004321)	(0.003169)	
0.6	0.029088	0.002367	5.554504
	(0.003711)	(0.003061)	
0.7	0.029919	0.023131	1.773544
	(0.002885)	(0.002516)	
0.8	0.047097	0.035134	3.049874
	(0.002919)	(0.002621)	
0.9	0.064539	0.051741	3.251210
	(0.002902)	(0.002659)	
1.0	0.231893	0.419753	-15.324946
	(0.006076)	(0.010647)	

Table 7: Piketty line ordinates from 2000 to 2005 for women and men

In order that we might use the estimated covariances as well as variances, we computed the income polarization functions $|\hat{\mu}_i - \hat{\mu}_M|$ for $i = 1$ and $i = 10$, that is, the bottom and top deciles. The results, along with standard errors, are shown in Table A5 for the four data sets analysed in this paper: women in 2000, men in 2000, women in 2005, and men in 2005. The table also displays the estimated covariance matrices of the four decile mean earnings estimates needed for the computation, namely $\hat{\mu}_1$, $\hat{\mu}_5$, $\hat{\mu}_6$, and $\hat{\mu}_{10}$.

Finally, we undertook four joint tests, for which the null hypotheses are that all decile differences in income shares are zero. These tests have nine degrees of freedom, not ten, since the shares sum to one in all cases. For the hypothesis that there is no difference between income shares of men and those of women in 2000, the Wald statistic is 1098.120771, allowing the hypothesis to be rejected at any conventional significance level. For the same hypothesis but for 2005, the statistic is 1395.077848. For comparisons across time, the first hypothesis is that income shares were the same in 2000 and 2005: the test statistic for men is equal to 1583.309008; for women it is 234.500725, still enough for rejection,

6. Conclusions and Implications

This paper uses a stochastic quantile function approach to derive the (asymptotic) variance-covariance structure of quantiles and income shares, and obtains explicit distribution-free formulas that do not depend on any assumptions about the distribution function. Consequently, the formulas are quite straightforward to compute using sample estimates from the available microdata. The paper then develops a toolbox of quantile-based disaggregative income inequality measures – such as Lorenz curve ordinates, relative mean income ratios, distributional distance measures and quantile income gaps, income polarization measures, and Piketty quantile growth rates – that can all be obtained from the quantile means and income share estimates, and hence also have distribution-free (asymptotic) variances and standard errors. This allows for a simple unified empirical framework of inferential analysis of distributional change. The framework is then implemented with Canadian Census public-use microdata files in order to investigate some of the earnings inequality changes that have occurred between 2000 and 2005.

The empirical findings show that, with large microdata sets, one can obtain quite strong statistically significant results (even over adjacent censuses). Three findings are highlighted:

1) Earnings differences between women and men are highly statistically significant right across the distribution with men having significantly higher earnings levels, but with women having significantly larger income shares in 2005 for all nine lower deciles, while for the tenth decile the opposite is the case.
2) Changes in decile earnings levels are also highly statistically significant right across the earnings distribution between 2000 and 2005, with gains most marked at the top of the distribution. The income shares for both women and men rose significantly at the very bottom and very top of the distributions, but fell across the broad mid ranges of the distribution.
3) The Piketty line (*i.e.*, the growth rates in decile earnings levels between 2000 and 2005) for women lies significantly above that for men for the nine lower deciles, but the growth of top decile earnings was much greater among men.

Several implications follow from the analysis of this paper. First, government statistical agencies – which already publish data series on decile means and decile income shares, such as the U.S. Bureau of the Census and Statistics Canada – should now also provide reliability measures (such as standard errors) for these statistics based on the formulas in this paper. In an era when most data series are available online, the addition of a reliability appendix, or an addendum of such information, should be quite straightforward to implement at relatively minimal cost.

Second, the approach of fairly detailed quantile analysis of income distributions, and how they have changed over time or differ across population groups, allows one to consider factors which may affect only specific regions of an income distribution (*e.g.*, a change in the minimum wage). It thus complements and serves as a bridge between simple descriptive analysis and a much more structural approach, such as quantile regressions (Firpo *et al.*, 2009).

Third, the stochastic quantile function approach is quite broadly applicable. When applied to median-based inequality measures, for example, it again leads to explicit formulas for (asymptotic) variances and covariances (Davidson, 2018). But the formulas are not distribution-free, since they depend on the population density function f evaluated at median-based points. However, one could combine this approach with computer-based algorithmic techniques for density ordinate evaluation, such as bootstrapping or kernel estimation methods to obtain an estimate $\hat{f}$, and once again have a usable formula for making statistical inferences. In this case, the $\hat{f}$ algorithm would have to be undertaken at the initial data-calculation stage rather than in a stand-alone toolbox program whose input is the output of the initial data-calculation results.

Appendix 1: Covariances of Income Shares

The estimate of the asymptotic variance of $\widehat{IS}_i$ is given by replacing all the quantities in (18) by their estimates. The $\hat{\xi}_i$ are sample quantiles, and with those, $\hat{n}_i$ and $\hat{n}_{2i}$ are defined by (9) and (10). From (15) we define the $\hat{m}_i$ and $\hat{m}_{2i}$ as the cumulative sums of the $\hat{n}_i$ and $\hat{n}_{2i}$ respectively. Next $\hat{\mu}$ is given by $\hat{m}_K$. These are enough to estimate $\mathrm{Var}(W_i)$ using (7), $\mathrm{Var}(Y)$ using (16), and $\mathrm{cov}(W_i, Y)$ using (17), and these provide an estimate of the asymptotic variance (18) of $\widehat{IS}_i$.

The asymptotic relation (14) can be written as

$$\widehat{IS}_i - IS_i = \frac{1}{\mu}(\hat{n}_i - n_i) - \frac{n_i}{\mu^2}(\hat{\mu} - \mu).$$

Now from (4) and (6), we can see that, asymptotically,

$$\hat{n}_i - n_i = N^{-1} \sum_{j=1}^{N} \big(w_{ij} - \mathrm{E}(W_i)\big),$$

while

$$\hat{\mu} - \mu = N^{-1} \sum_{j-1}^{N} (y_j - \mu),$$

so that

$$\widehat{IS}_i - IS_i = \frac{1}{N\mu^2} \sum_{j=1}^{N} \big[\mu(w_{ij} - \mathrm{E}(W_i) - n_i(y_j - \mu)\big]. \tag{21}$$

Let the random variable V_i be defined as follows:

$$V_i = \frac{1}{\mu}\big(W_i - \mathrm{E}(W_i)\big) - \frac{n_i}{\mu^2}(Y - \mu)\big). \tag{22}$$

Then the terms of the sum in (21) are IID drawings from V_i. It is easy to check that the expression (18) for the asymptotic variance of $\widehat{IS}_i$ is just the variance of V_i divided by N.

For the covariance of $\widehat{IS}_i$ and $\widehat{IS}_j$ with $j < i$, we get directly from (22) that, asymptotically,

$$N\,\mathrm{cov}(\widehat{IS}_i, \widehat{IS}_j) = \frac{1}{\mu^2}\,\mathrm{cov}(W_i, W_j) + \frac{n_i n_j}{\mu^4}\,\mathrm{Var}(Y) - \frac{n_i}{\mu^3}\,\mathrm{cov}(W_j, Y) - \frac{n_j}{\mu^3}\,\mathrm{cov}(W_i, Y) \tag{23}$$

Everything in this expression can be estimated in a distribution-free manner using the results (11), (16), and (17). Note that the complete $K \times K$ covariance matrix of the $\widehat{IS}_i$ is singular, since $\sum_{i=1}^{K} IS_i = \sum_{i=1}^{K} \widehat{IS}_i = 1$.

Algorithm

Since one of our aims in this paper is to make available a straightforward procedure for disaggregative statistical analysis of samples consisting of data on incomes, wealth, or other variables of interest, we provide here a detailed algorithm for constructing estimates of quantile-based measures of these variables, along with their variance-covariance matrices and standard errors.

1. Compute $\hat{\mu}$ as the sample mean, and $\hat{\sigma}^2 = \widehat{\text{Var}}(Y)$ as the sample variance.
2. Select the number of groups K for which inference is to be carried out, and the probabilities p_i $i = 0, \ldots, K$, $0 = p_0 < p_1 < p_2 < \ldots < p_K = 1$, the corresponding quantiles of which separate the groups. The most usual choice is $p_i = i/K$, but that is not necessary for our analysis.
3. Sort the data, y_j, $j = 1, \ldots, N$, from smallest to largest, so as to find the order statistics $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(N)}$. Obtain the sample quantiles for the probabilities p_i: $\hat{\xi}_0 = y_{(1)}$, $\hat{\xi}_K = y_{(N)}$, and $\xi_i = y_{(\lceil Np_i \rceil)}$ for $i = 1, \ldots, K-1$.
4. Progressively, for $i = 1, \ldots, K$:
   (i) compute $\hat{n}_i$ and $\hat{n}_{2i}$ using the formulas (9) and (10);
   (ii) compute $\hat{m}_i$ and $\hat{m}_{2i}$ according to (15);
   (iii) compute the mean income of group i as $\hat{\mu}_i = \hat{n}_i/(p_i - p_{i-1})$;
   (iv) compute $\widehat{\text{RMI}}_i$ as $\hat{\mu}_i/\hat{\mu}$;
   (v) compute the estimated income shares $\widehat{IS}_i$ as $\hat{n}_i/\hat{\mu}$;
   (vi) obtain the estimated asymptotic variance of $\hat{n}_i$, noting that it is equal to $\widehat{\text{Var}}(W_i)$ as given by (8), divided by N;
   (vii) divide the estimate $\widehat{\text{Var}}(\hat{n}_i)$ by $(p_i - p_{i-1})^2$ to get the estimated asymptotic variance of $\hat{\mu}_i$;
   (viii) compute the estimate $\widehat{\text{cov}}(W_i, Y)$ by estimating the formula (17) using $\hat{n}_i$, $\hat{n}_{2i}$, $\hat{\mu}$, $\hat{\xi}_i$, $\hat{\xi}_{i-1}$, $\hat{m}_i$, and $\hat{m}_{i-1}$, all of which have already been calculated;
   (ix) estimate the asymptotic variance of $\widehat{IS}_i$ by use of the formula (18), using $\hat{\mu}$, $\hat{n}_i$, $\widehat{\text{Var}}(W_i)$, $\widehat{\text{Var}}(Y)$, and $\widehat{\text{cov}}(W_i, Y)$, all now available;
   (x) if $i > 1$, then, for $j = 1, \ldots, i-1$, obtain the estimated asymptotic covariances $\widehat{\text{cov}}(\hat{\mu}_i, \hat{\mu}_j)$ by estimating formula (11) for $\text{cov}(W_i, W_j)$ and dividing by $N(p_i - p_{i-1})(p_j - p_{j-1})$, with the estimates $\hat{\xi}_j$, $\hat{\xi}_{j-1}$, $\hat{\xi}_i$, $\hat{\xi}_{i-1}$. $\hat{n}_j$, $\hat{n}_i$;
   (xi) if $i > 1$, then, for $j = 1, \ldots, i-1$, obtain the estimated asymptotic covariances $\widehat{\text{cov}}(\widehat{IS}_j, \widehat{IS}_i)$ by estimating the right-hand side of (23) and dividing by N.
5. Take the square roots of all estimated variances in order to obtain corresponding standard errors.

As a check, verify that $\hat{\mu}$ is equal to $\hat{m}_K$, and that $\hat{\sigma}^2 = \widehat{\text{Var}}(Y)$ is equal to $\hat{m}_{2K} - (\hat{m}_K)^2$; see (16). Another check: verify that $\sum_{i=1}^{K} \widehat{IS}_i = 1$.

Although the above algorithm makes only one pass through the data, it may be useful to think of the calculations as occurring in two stages. The first stage operates on the raw data of the available microdata file, and includes the calculations for $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\xi}_i$, $\hat{n}_i$, $\hat{n}_{2i}$, $\hat{m}_i$, $\hat{m}_{2i}$, $\hat{\mu}_i$, $\widehat{IS}_i$, $\widehat{\mathrm{RMI}}_i$; that is, steps 1–4(v) of the algorithm. These then serve as the input to a separate stand-alone program, or second stage, that calculates all the desired inference statistics such as the estimated variances, covariances, standard errors, and "t-ratios"; steps 4(vi)–5. One can then treat the second stage calculations as a "black box" program that would allow empirical practitioners to undertake convenient quantile-based inferential analysis of income (or other) distributions.

Appendix 2: Simulation Evidence

Simulations were run in order to see to what extent the numerous estimates produced by the above algorithm do indeed approximate finite-sample properties. The simulated data were generated, using a lognormal distribution, as the exponential of drawings from the standard normal distribution $\mathrm{N}(0,1)$. The simulated samples contained $N = 1000$ IID drawings from this distribution, and were split into deciles, so that $K = 10$, $p_i = i/10$, $i = 0, 1, \ldots, 10$. The true values of all the estimated properties are readily computed for the lognormal distribution.[2]

For each of 100,000 replications, realisations were obtained for $\hat{\mu}_i$, $\widehat{IS}_i$, $\widehat{\mathrm{Var}}(\hat{\mu}_i)$,and $\widehat{\mathrm{Var}}(\widehat{IS}_i)$, $i = 1, \ldots, 10$.[3] The realisations were then averaged over the 100,000 replications, and the results compared with the true theoretical values, as shown in Tables A1 and A2 below.

[2] Since the range of the lognormal distribution is unbounded above, the 0.999999 quantile was used for ξ_{10}.

[3] The variances and their estimates were not divided by $N = 1000$, to avoid very small numbers.

It is clear that the estimates are quite reliable, with no indication of bias, either upwards or downwards.

i	μ_i	$\hat{\mu}_i$	IS_i	$\widehat{IS}_i$
1	0.185612	0.185921	0.011258	0.011292
2	0.354597	0.355120	0.021507	0.021565
3	0.510099	0.510601	0.030939	0.031005
4	0.681505	0.682043	0.041335	0.041412
5	0.883971	0.884670	0.053616	0.053710
6	1.137309	1.138167	0.068981	0.069094
7	1.476320	1.477357	0.089543	0.089674
8	1.976805	1.978381	0.119899	0.120066
9	2.865101	2.867734	0.173777	0.173992
10	6.414457	6.405586	0.389056	0.388166

Table A1: point estimates

i	$\mathrm{Var}(\hat{\mu}_i)$	$\widehat{\mathrm{Var}}(\hat{\mu}_i)$	$\mathrm{Var}(\widehat{IS}_i)$	$\widehat{\mathrm{Var}}(\widehat{IS}_i)$
1	0.001131	0.001143	0.000519	0.000522
2	0.002604	0.002625	0.001238	0.001236
3	0.004433	0.004466	0.002049	0.002036
4	0.007174	0.007247	0.003105	0.003081
5	0.011571	0.011685	0.004536	0.004491
6	0.019172	0.019363	0.006567	0.006489
7	0.033776	0.034133	0.009649	0.009525
8	0.067099	0.067829	0.014975	0.014785
9	0.173934	0.175825	0.027561	0.027307
10	2.218927	2.231962	0.256691	0.253030

Table A2: variance estimates

Appendix 3: More Empirical Results

In this appendix, results are given for the basic information in 2000. First, women, in Table A3

p_i	$\hat{\xi}_i$	$\hat{\mu}_i$	$\widehat{\mathrm{RMI}}_i$	$\widehat{IS}_i$
0.1	3.259605	1.462230	0.053337	0.005334
		(0.013585)	(0.000470)	(0.000047)
0.2	7.302770	5.200345	0.189689	0.018969
		(0.030181)	(0.000980)	(0.000098)
0.3	11.894426	9.669303	0.352700	0.035270
		(0.042092)	(0.001250)	(0.000125)
0.4	16.790550	14.577663	0.531738	0.053174
		(0.050011)	(0.001350)	(0.000135)
0.5	22.387400	20.107790	0.733456	0.073346
		(0.059940)	(0.001470)	(0.000147)
0.6	29.103620	25.945183	0.946382	0.094638
		(0.071710)	(0.001610)	(0.000161)
0.7	35.044117	32.182600	1.173900	0.117390
		(0.060723)	(0.001320)	(0.000132)
0.8	43.655430	39.150886	1.428077	0.142808
		(0.078785)	(0.001650)	(0.000165)
0.9	56.058050	49.450373	1.803763	0.180376
		(0.091620)	(0.002110)	(0.000211)
1.0		76.404807	2.786959	0.278696
		(0.194850)	(0.005090)	(0.000509)

Table A3: Women in 2000 (asymptotic standard errors in brackets)

Next, men, in Table A4

p_i	$\hat{\xi}_i$	$\hat{\mu}_i$	$\widehat{\mathrm{RMI}}_i$	$\widehat{IS}_i$
0.1	5.037165	2.192164	0.052300	0.005230
		(0.020273)	(0.000460)	(0.000046)
0.2	11.753385	8.362229	0.199502	0.019950
		(0.047063)	(0.001000)	(0.000100)
0.3	20.148660	16.105102	0.384228	0.038423
		(0.072545)	(0.001440)	(0.000144)
0.4	27.984250	24.379823	0.581642	0.058164
		(0.075331)	(0.001360)	(0.000136)
0.5	35.484029	31.993904	0.763295	0.076329
		(0.075331)	(0.001270)	(0.000127)
0.6	43.655430	39.479551	0.941884	0.094188
		(0.082244)	(0.001330)	(0.000133)
0.7	52.198462	47.389190	1.130588	0.113059
		(0.081904)	(0.001350)	(0.000135)
0.8	63.804090	57.416480	1.369814	0.136981
		(0.099414)	(0.001570)	(0.000157)
0.9	81.714010	71.473062	1.705169	0.170517
		(0.122080)	(0.001920)	(0.000192)
1.0	274.245650	120.363833	2.871581	0.287158
		(0.365381)	(0.005880)	(0.000588)

Table A4: Men in 2000 (asymptotic standard errors in brackets)

$$\begin{bmatrix} 0.000185 & 0.000278 & 0.000270 & 0.000181 \\ 0.000278 & 0.003593 & 0.003706 & 0.002479 \\ 0.000270 & 0.003706 & 0.005142 & 0.003692 \\ 0.000181 & 0.002479 & 0.003692 & 0.037967 \end{bmatrix} \quad \begin{bmatrix} 0.000411 & 0.000518 & 0.000458 & 0.000483 \\ 0.000518 & 0.005675 & 0.005392 & 0.005681 \\ 0.000458 & 0.005392 & 0.006764 & 0.007640 \\ 0.000483 & 0.005681 & 0.007640 & 0.133504 \end{bmatrix}$$

$$\begin{bmatrix} 0.000249 & 0.000315 & 0.000259 & 0.000321 \\ 0.000315 & 0.003829 & 0.003370 & 0.004180 \\ 0.000259 & 0.003370 & 0.003827 & 0.005054 \\ 0.000321 & 0.004180 & 0.005054 & 0.157900 \end{bmatrix} \quad \begin{bmatrix} 0.000539 & 0.000540 & 0.000569 & 0.001176 \\ 0.000540 & 0.004787 & 0.005365 & 0.011092 \\ 0.000569 & 0.005365 & 0.007807 & 0.017396 \\ 0.001176 & 0.011092 & 0.017396 & 1.373067 \end{bmatrix}$$

	Women 2000	Men 2000	Women 2005	Men 2005
lower measure	21.564256	33.544564	21.693331	32.869700
	(0.060609)	(0.072390)	(0.057220)	(0.072536)
upper measure	53.378320	84.627105	70.544455	135.358656
	(0.189296)	(0.354948)	(0.390211)	(1.162071)

Table A5: covariance matrices and income polarization statistics
covariance matrices: left column women, right column men, first row 2000, second row 2005

References

Aaberge, R. (2000). "Characterizations of Lorenz Curves and Income Distributions", *Social Choice and Welfare* 17, 639-53.

Aaberge, R. (2001) "Axiomatic Characterization of the Gini Coefficient and Lorenz Curve Orderings", *Journal of Economic Theory* 101, 115-32.

Acemoglu, D. H., D. Autor, G. H. Hanson, and B. Price (2016) "Import Competition and the Great U.S. Employment Sag of the 2000s", *Journal of Labor Economics* 34, S141-98.

Atkinson, A. (1970) "On the Measurement of Inequality", *Journal of Economic Theory* 2, 244-63.

Autor, D. H., D. Dorn, and G. H. Hanson (2013) "The Geography of Trade and Technology Shocks in the United States", *American Economic Review* 103, 220-25.

Beach, C. M. (2016) "Changing Income Inequality: A Distributional Paradigm for Canada", *Canadian Journal of Economics* 49(4), 1229-92.

Cowell, F. A. (2011) *Measuring Inequality*, Third Edition. Oxford: Oxford University Press.

Davidson, R. (2018). "Statistical Inference on the Canadian Middle Class", *Econometrics*, 6(1), 14; https://doi.org/10.3390/econometrics6010014

del Rio, C., C. Gradin, and O. Canto (2011) "The Measurement of Gender Wage Discrimination: The Distributional Approach Revisited", *Journal of Economic Inequality* 9, 57-86.

The Economist (2021a) "Free Exchange: Fleshing out the Olive", Aug 28, 2021, 65.

The Economist (2021b) "Free Exchange: Black Cat, White Cat, Fat Cat, Thin Cat", Oct 2, 2021, 62.

The Economist (2021c) "Incomes: Piketty Lines", Oct 9, 2021, 24.

Firpo, S., N. M. Fortin, and T. Lemieux (2009) "Unconditional Quantile Regressions", *Econometrica* 77(3), 953-73.

Foster, J. E., and M. C. Wolfson (1992) "Polarization and the Decline of the Middle Class: Canada and the U.S.", unpublished manuscript.

Goos, M., A. Manning, and A. Salomons (2014) "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring", *American Economic Review* 104, 2509-26.

Guvenen, F., L. Pistaferri, and G. L. Violante (2022) "Global Trends in Income Inequality and Income Dynamics: New Insights from GRID", *Quantitative Economics* 13, 1321-60.

Jenkins, S. P. (1994) "Earnings Discrimination Measurement: A Distributional Approach", *Journal of Econometrics* 61, 81-102.

Jenkins, S. P. (1999) "Analysis of Income Distributions", Stata Technical Bulletin 48, 4-18; reprinted in Stata Technical Bulletin Reprints 8, 343-60.

Kovacevic, M. S., and D. A. Binder (1997) "Variance Estimation for Measures of Income Inequality and Polarization — The Estimating Equations Approach", *Journal of Official Statistics* 13(1), 41-58.

Lambert, P. J. (2001) *The Distribution and Redistribution of Income, Third Edition*, Manchester: Manchester University Press.

Maasoumi, E. (1998) "Empirical Analysis of Welfare and Inequality", Ch. 5 in *Handbook of Applied Econometrics, Vol II: Microeconometrics*, eds. M.H. Pesaran and P. Schmidt. Oxford: Blackwell Publishers, 187-226.

Rao, C. R. (1965) *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Salas, R., J. A. Bishop, and L. A. Zeager (2018) "Second-Order Discrimination and Generalized Lorenz Dominance", *Review of Income and Wealth* 64(3), 563-75.

Wilks, S. S. (1962) *Mathematical Statistics*, New York: John Wiley & Sons.

Wolfson, M. C. (1994)"When Inequalities Diverge", *American Economic Review* 84 (AEA Paper and Proceedings), 353-58.