A Statistical Characterization of Median-Based Inequality Measures

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 considers income distributions which are divided into lower, middle and upper regions based on separating points that are scalar multiples of the median. For each of these regions or income groups, the paper works out the asymptotic distribution properties – including explicit asymptotic variance formulas and hence standard errors – of sample estimates of the proportion of the population within the group, their share of total income and the groups' mean incomes. It then applies these results to do the same for relative mean income ratios, various polarization measures and decile-mean income ratios. Since the derived formulas are not distribution-free, the study advises a density estimation technique of Comte and Genon-Catalot (2012). Major distributional shifts out of the middle of the distribution and polarization at the top are found to be highly statistically significant.

Key words: median-based inequality, income shares, population shares, income polarization

JEL codes: C10, C42, D31

March 2025

1. Introduction

Two major distributional changes have characterized many developed economies since around 1980: declining middle-class incomes and rising top incomes (Hoffman *et al.* (2020); Blanchet *et al.* (2022); and Guvenen *et al.*, (2022)). Structural factors behind these changes have been extensively reviewed in, for example, Acemoglu *et al.*, (2016); Autor, Dorn and Hanson (2013); Beach (2016); Goos, Manning and Salomons (2014);, Saez and Veall (2007); and Veall (2012). It would clearly be useful to be able to capture both of these sets of changes efficiently in a simple empirical framework that allows for a conventional statistical inference methodology, in order that one can test for the statistical significance of such changes over time.

The distributional measures that are typically used to examine these patterns of distributional change are the income shares of middle- and upper-income groups, the relative sizes of these groups, and the relative incomes of these groups. Beach (2016) demonstrated the usefulness, in examining these changes, of characterizing the income groups in terms of their relationship to the median income level. So, for example, the middle-income group (M) could be defined as including those with incomes between, say, fifty percent and two hundred percent of the median, the upper group (H) as those with incomes above twice the median, and the lower group (L) as those with incomes below half the median. This allows one to get separate estimates for group income shares (IS_i, $i = L, M, H$) and for the proportion of recipients within the group (or population share) PS_i, as well as for the group mean incomes (μ_i). This distributional framework allows a more insightful interpretation of distributional change, since one can then analyze both the size (PS_i) and the relative prosperity (μ_i/μ) of the income group separately. (Percentile- or quantile-based measures, by construction, assign the size of the income groups as a prespecified percentage such as the top decile or 10% of all income recipients.) Characterizing group size and prosperity allows one to capture the quantity dimension of a change in the group's total income separately from the income per recipient. This in turn can be used to help identify the relative strength of demand-side or supply-side driving factors behind observed distributional change (Katz and Murphy (1992)). Such insights, though, have heretofore been based on the relative magnitude of these effects, not on their statistical significance. This framework also allows for a richer and more extensive set of measures of income polarization, in terms of both quantity and relative income dimensions at the tails of the distribution.

Davidson (2018) proposes a stochastic quantile function approach to derive asymptotic covariances and variances – and hence standard errors – for sample estimates of IS_M and PS_M for middle-group income recipients within the median-based empirical framework, thus providing for formal statistical inference on these measures. The present paper extends Davidson's statistical analysis to apply to lower- and upper-income groups as well (all defined in terms of the median), so that one can examine a full set of population subsets covering an income distribution (*i.e.*, for L, M, and H subsets) jointly. The analysis shows how this approach leads to explicit formulas for asymptotic variances and standard errors, which can be easily programmed, for $\widehat{IS}_i$ and $\widehat{PS}_i$, for all of $i = L, M, H$ income groups. And the paper extends the set of distributional measures to a relative mean income statistic

$\hat{\mu}_i/\hat{\mu}$, where μ_i is the mean of group i incomes and μ is the overall population mean, and also to $\hat{\mu}_i$ itself, so that one can test for the statistical significance of growing income gaps among income groups.

The paper thus proposes a general framework for median-based income inequality analysis, based on asymptotic statistical inference. The present study serves as a complement to a separate piece by the authors (Beach and Davidson (2024)) that develops a comparable framework for inequality measures, based on quantile income shares as typically published by government statistical agencies. Together, the two papers provide the basis for a black-box set of calculations that can be readily implemented to allow standard statistical inference for frequently used statistics of disaggregated income inequality change. The paper is 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 is organized as follows. The next section outlines the stochastic quantile function approach to statistical inference. It then extends Davidson's (2018) middle-class group results for estimated income shares and population shares to corresponding lower- and upper-income groups as well and expresses the asymptotic variance results in terms of simple explicit formulas that can be estimated from available microdata. The extension of these results to group mean income measures is also presented. In Section 3 the results in Section 2 are used to obtain results for relative group mean incomes, measures of polarization, and mean-decile distribution functions. Section 4 provides an empirical application of the Section 2 theoretical results to Canadian Census earnings data. The final section summarizes the main results of the paper and notes some implications.

2. Basic Asymptotic Analysis

Let F be the population distribution of income recipients, and let Y denote a random variable of which the cumulative distribution function (CDF) is F. We make the following somewhat restrictive assumption:

Assumption

The CDF F is differentiable and strictly increasing on its compact support.

The assumption is made for convenience and in order to simplify the asymptotic analysis. If it is not satisfied, various asymptotically negligible terms appear in the estimators of group population and income shares, which complicate the analysis.

2.1 Population Shares

Let m denote the median of the distribution F. Then the population share of those recipients with income no greater than bm for some $b > 0$ is $F(bm)$. If we have a random sample from the population of size N, we can estimate the distribution by the empirical distribution function (EDF) $\hat{F}$, defined as follows:

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

where the y_i, $i = 1, \ldots, n$, are the observed incomes in the sample, and I is the indicator function, with value 1 if its argument is true, 0 if it is false. The sample median $\hat{m}$ is defined as usual:

$$\hat{m} = \begin{cases} y_{(n+1)} & \text{if } N = 2n+1 \ (N \text{ odd}) \\ (y_{(n)} + y_{(n+1)})/2 & \text{if } N = 2n \ (N \text{ even}). \end{cases}$$

The natural estimate of the population share is $\hat{F}(b\hat{m})$. We have

$$\begin{aligned} \hat{F}(b\hat{m}) - F(bm) &= \int_0^{b\hat{m}} \mathrm{d}\hat{F}(y) - \int_0^{bm} \mathrm{d}F(y) \\ &= \int_0^{bm} \mathrm{d}(\hat{F} - F)(y) + \int_{bm}^{b\hat{m}} \mathrm{d}F(y) + \int_{bm}^{b\hat{m}} \mathrm{d}(\hat{F} - F)(y). \end{aligned} \tag{1}$$

Under our Assumption, and also under less restrictive but still conventional regularity conditions, the first two terms above are of order $N^{-1/2}$, while the last, being of order N^{-1}, can be ignored for asymptotic analysis. Then, to leading order, we see that

$$N^{1/2}\big(\hat{F}(b\hat{m}) - F(bm)\big) = N^{-1/2} \sum_{i=1}^{N} \Big(\mathrm{I}(y_i \le bm) - F(bm)\Big) + bf(bm)(\hat{m} - m),$$

where $f(y) = F'(y)$ is the population density function. According to the Bahadur (1966) representation of quantiles,

$$\hat{m} - m = -\frac{1}{Nf(m)} \sum_{i=1}^{N} \big[\mathrm{I}(y_i \le m) - \tfrac{1}{2}\big] + O(N^{-3/4}(\log N)^{3/4}), \tag{2}$$

and so

$$\begin{gathered} N^{1/2}\big(\hat{F}(b\hat{m}) - F(bm)\big) = \\ N^{-1/2} \sum_{i=1}^{N} \Big[\mathrm{I}(y_i \le bm) - F(bm) - \frac{bf(bm)}{f(m)} \big[\mathrm{I}(y_i \le m) - \tfrac{1}{2}\big]\Big] + o_p(1). \end{gathered} \tag{3}$$

Let B be equal to $bf(bm)/f(m)$ and consider the random variable $U(b)$ defined as follows:

$$U(b) \equiv \mathrm{I}(Y \le bm) - B\,\mathrm{I}(Y \le m), \tag{4}$$

where Y is a variable that has the distribution F. Then clearly

$$\mathrm{E}\big(U(b)\big) = F(bm) - B/2. \tag{5}$$

The terms in the sum in (3) can be seen to be IID realizations of the random variable $U(b) - \mathrm{E}\big(U(b)\big)$, and so it follows that $N^{1/2}\big(\hat{F}(b\hat{m}) - F(bm)\big)$ is asymptotically equal

in distribution to $U(b) - \mathrm{E}\big(U(b)\big)$. Asymptotic normality follows from the central-limit theorem. The variance of the limiting distribution, which, following standard terminology, we refer to as the *asymptotic variance* of $\hat{F}(b\hat{m})$, is then just $\mathrm{Var}\big(U(b)\big)$. In order to estimate this variance, let

$$\hat{u}_i(b) = \mathrm{I}(y_i \le b\hat{m}) - \hat{B}\,\mathrm{I}(y_i \le \hat{m}), \quad i = 1, \ldots, N, \tag{6}$$

with $\hat{B} = b\hat{f}(b\hat{m})/\hat{f}(\hat{m})$, using appropriate estimates $\hat{f}$ of the density. Then, to leading order,

$$N^{1/2}\big(\hat{F}(b\hat{m}) - F(bm)\big) = N^{-1/2}\sum_{i=1}^{N}\big[\hat{u}_i(b) - \widehat{\mathrm{E}}\big(U(b)\big)\big],$$

with, from (5),

$$\widehat{\mathrm{E}}\big(U(b)\big) = \hat{F}(b\hat{m}) - \hat{B}/2,$$

Then $\mathrm{Var}\big(U(b)\big)$ can be estimated by the sample variance of the $\hat{u}_i(b)$.

A possibly better approach is simply to compute $\mathrm{Var}\big(U(b)\big)$ directly, and then estimate the result. It is easy to see from (4) that

$$U^2(b) = \mathrm{I}(Y \le bm) + B^2\mathrm{I}(Y \le m) - 2B\,\mathrm{I}\big(Y \le \min(m, bm)\big),$$

whence

$$\mathrm{E}\big(U^2(b)\big) = F(bm) + \tfrac{1}{2}B^2 - 2B\,\min\big(F(bm), \tfrac{1}{2}\big). \tag{7}$$

Next, $\mathrm{Var}\big(U(b)\big) = \mathrm{E}\big(U^2(b)\big) - \big(\mathrm{E}(U)\big)^2$, and so from (5), for $b < 1$, we have

$$\mathrm{Var}\big(U(b)\big) = F(bm)\big(1 - F(bm)\big) + \tfrac{1}{4}B^2 - BF(bm). \tag{8}$$

We see that $\mathrm{Var}\big(U(b)\big)$ can be estimated in a distribution-free manner by

$$\widehat{\mathrm{Var}}\big(U(b)\big) = \hat{F}(b\hat{m})\big(1 - \hat{F}(b\hat{m})\big) + \tfrac{1}{4}\hat{B}^2 - \hat{B}\hat{F}(b\hat{m}).$$

Let $a > b$, and make the definitions

$$U(a) = \mathrm{I}(Y \le am) - A\,\mathrm{I}(Y \le m); \quad A = af(am)/f(m), \tag{9}$$

and, for $a > 1$

$$\mathrm{Var}\big(U(a)\big) = F(am)\big(1 - F(am)\big) + \tfrac{1}{4}A^2 - A\big(1 - F(am)\big). \tag{10}$$

Then $N^{1/2}\big(\hat{F}(a\hat{m}) - F(am)\big)$ is asymptotically equal in distribution to $U(a) - \mathrm{E}\big(U(a)\big)$.

Some comments are in order concerning the "appropriate" estimates $\hat{f}(b\hat{m})$ amd $\hat{f}(\hat{m})$. In Appendix 2 we sketch an alternative to conventional kernel density estimation that works

much better with distributions that have support only on the positive real line or a subset of it. Here we follow the work of Comte and Genon-Catalot (2012).

The analysis so far developed is sufficient for estimating and providing standard errors for the population share with income less than bm or greater than bm. But in order to estimate the population share of recipients with income in some interval $bm < y \leq am$, $a, b > 0$, $b < a$, as was done in Davidson (2018), one needs not only the variances of $\hat{F}(a\hat{m})$ and $\hat{F}(b\hat{m})$, but also their covariance. The asymptotic covariance of $N^{1/2}\big(\hat{F}(a\hat{m}) - F(am)\big)$ and $N^{1/2}\big(\hat{F}(b\hat{m}) - F(bm)\big)$ is the covariance of $U(a)$ and $U(b)$.

Make the definitions $m_a = \min(am, m)$ and $m_b = \min(bm, m)$. Then

$$\begin{aligned} U(a)U(b) &= \big(\mathrm{I}(Y \leq am) - A\,\mathrm{I}(Y \leq m)\big)\big(\mathrm{I}(Y \leq bm) - B\,\mathrm{I}(Y \leq m)\big) \\ &= \mathrm{I}(Y \leq bm) - A\,\mathrm{I}(Y \leq m_b) - B\,\mathrm{I}(Y \leq m_a) + AB\,(Y \leq m), \end{aligned}$$

whence

$$\mathrm{E}\big(U(a)U(b)\big) = F(bm) - AF(m_b) - BF(m_a) + \tfrac{1}{2}AB, \tag{11}$$

whereas

$$\begin{aligned} \mathrm{E}\big(U(a)\big)\mathrm{E}\big(U(b)\big) &= \big(F(am) - \tfrac{1}{2}A\big)\big(F(bm) - \tfrac{1}{2}B\big) \\ &= F(bm)F(am) - \tfrac{1}{2}\big(BF(am) + AF(bm)\big) + \tfrac{1}{4}AB. \end{aligned}$$

From this, we see immediately that

$$\begin{aligned} &\mathrm{cov}\big(U(a), U(b)\big) = \mathrm{E}\big(U(a)U(b)\big) - \mathrm{E}\big(U(a)\big)\mathrm{E}\big(U(b)\big) \\ &= F(bm)\big(1 - F(am)\big) - A\big(F(m_b) - \tfrac{1}{2}F(bm)\big) - B\big(F(m_a) - \tfrac{1}{2}F(am)\big) + \tfrac{1}{4}AB, \end{aligned} \tag{12}$$

and this can be estimated in a distribution-free manner.

Although the results of this section so far are quite general, for most of the rest of the paper, interest will be focused on the case with $b < 1 < a$. The share of the population with income not exceeding bm, that is, $F(bm)$, will be denoted by PS_L, where 'L' stands for the group of lower-income recipients. The population share of the middle-income group is $F(am) - F(bm)$; it is denoted by PS_M. The share of the higher-income group, $1 - F(am)$, is denoted by PS_H.

It is clear from (8) that

$$\text{Asy var}(\widehat{PS}_L) = \mathrm{Var}\big(U(b)\big) = PS_L(1 - PS_L) + B^2/4 - B\,PS_L, \tag{13}$$

and from (10) that

$$\text{Asy var}(\widehat{PS}_H) = \mathrm{Var}\big(U(a)\big) = PS_H(1 - PS_H) + A^2/4 - A\,PS_H. \tag{14}$$

Note that the terms on the right-hand sides of these equations have simple intuitive interpretations. The first (product) term corresponds to the variance of random recipients

lying within the respective population share, the second (squares) term corresponds to the variance of the estimated median-based cut-off points, and the last term corresponds to the covariance or interaction between the first two components.

The population share of recipients of incomes between bm and am is $PS_M = F(am) - F(bm)$, and the limiting variance of $N^{1/2}(\hat{P}S_M - PS_M)$ is equal to $\mathrm{Var}\big(U(a) - U(b)\big) = \mathrm{Var}\big(U(a)\big) + \mathrm{Var}\big(U(b)\big) - 2\,\mathrm{cov}\big(U(a), U(b)\big)$. The covariance (12) can now be rewritten as

$$\mathrm{cov}\big(U(a), U(b)\big) = PS_L PS_H - A\,PS_L/2 - B\,PS_H/2 + AB/4, \tag{15}$$

and so the asymptotic variance of $\widehat{PS}_M$, after a little algebra based on (13), (14), and (15), can be seen to be

$$PS_M(1 - PS_M) + \tfrac{1}{4}(A - B)^2 - (A - B)(PS_H - PS_L). \tag{16}$$

The same expression results from calculating $\mathrm{E}\big((U(a) - U(b))^2\big) - \big(\mathrm{E}(U(a) - U(b))\big)^2$ directly. Let $C = A - B$. Then (16) can also be written as

$$\text{Asy var}(\widehat{PS}_M) = PS_M(1 - PS_M) + C^2/4 - C(PS_H - PS_L). \tag{17}$$

2.2 Income Shares

We begin by considering the income share of recipients of incomes no greater than bm, with $b < 1$. The average income earned by these recipients is $n(bm)$, defined as follows:

$$n(bm) = \int_0^{bm} y\,\mathrm{d}F(y), \quad \text{estimated by} \quad \hat{n}(b\hat{m}) = \int_0^{b\hat{m}} y\,\mathrm{d}\hat{F}(y), \tag{18}$$

and the income share is $n(bm)/\mu$, where $\mu \equiv \int_0^\infty y\,\mathrm{d}F(y)$ is the mean income of the population, estimated by $\int_0^\infty y\,\mathrm{d}\hat{F}(y)$. Note that $\mu = n(\infty)$ and $\hat{\mu} = \hat{n}(\infty)$. With $b < 1$, we denote $n(bm)$ and $\hat{n}(b\hat{m})$ by n_L and $\hat{n}_L$ respectively, and we denote the income share of the lower-income group by IS_L. Clearly $IS_L = n_L/\mu$.

For incomes greater than am, with $a > 1$, the average income is $\mu - n_H$ with $n_H = n(am)$ defined just as in (18), replacing b by a. The income share is $IS_H = (\mu - n_H)/\mu = 1 - n_H/\mu$. For the middle-income group, average income is $n_H - n_L$, and the income share is $IS_M = (n_H - n_L)/\mu = 1 - IS_H - IS_L$,

By analogy with (1) for population shares, we have

$$\begin{aligned}\hat{n}_L - n_L &= \int_0^{b\hat{m}} y\,\mathrm{d}\hat{F}(y) - \int_0^{bm} y\,\mathrm{d}F(y) \\ &= \int_0^{bm} y\,\mathrm{d}(\hat{F} - F)(y) + \int_{bm}^{b\hat{m}} y\,\mathrm{d}F(y) + \int_{bm}^{b\hat{m}} y\,\mathrm{d}(\hat{F} - F)(y),\end{aligned}$$

where the third term can be ignored asymptotically. With a random sample of size N, as in the preceding subsection, the first term is exactly equal to

$$N^{-1}\sum_{i=1}^{N}\big[y_i\mathrm{I}(y_i \le bm) - n(bm)\big],$$

and the second term can be approximated to leading order by $b(\hat{m}-m)bmf(bm)$, and, by (2), that approximation is to leading order equal to

$$-N^{-1}\sum_{i=1}^{N}\frac{b^2 m f(bm)}{f(m)}\big(\mathrm{I}(y_i \le m) - \tfrac{1}{2}\big).$$

This leads to

$$N^{1/2}(\hat{n}_L - n_L) = N^{-1/2}\sum_{i=1}^{N}\big[y_i\mathrm{I}(y_i \le bm) - bmB\,\mathrm{I}(y_i \le m) - n(bm) + \tfrac{1}{2}bmB\big] + o_p(1). \quad (19)$$

Next, we define the random variable $U_1(b)$ as

$$U_1(b) = Y\,\mathrm{I}(Y \le bm) - bmB\,\mathrm{I}(Y \le m), \quad (20)$$

noting that $\mathrm{E}\big(U_1(b)\big) = n_L - bmB/2$. It follows now that $N^{1/2}(\hat{n}_L - n_L)$ is asymptotically equal in distribution to $U_1(b) - \mathrm{E}\big(U_1(b)\big)$.
Similarly, for $a > 1$, we can define

$$U_1(a) = Y\,\mathrm{I}(Y \le am) - amA\,\mathrm{I}(Y \le m),$$

where $\mathrm{E}\big(U_1(a)\big) = n(am) - amA/2 = n_H - amA/2$, and $N^{1/2}(\hat{n}_H - n_H)$ is asymptotically equal in distribution to $U_1(a) - \mathrm{E}\big(U_1(a)\big)$.
For the variance of $U_1(b)$, we compute as follows:

$$U_1^2(b) = Y^2\,\mathrm{I}(Y \le bm) + (bmB)^2\,\mathrm{I}(Y \le m) - 2bmB\,Y\,\mathrm{I}\big(Y \le bm)\big), \quad (21)$$

so that

$$\mathrm{E}\big(U_1^2(b)\big) = n_{2,L} + \tfrac{1}{2}(bmB)^2 - 2bmB\,n_L, \quad (22)$$

where we define $n_{2,L} = \int_0^{bm} y^2\,\mathrm{d}F(y)$. It follows that

$$\mathrm{Var}\big(U_1(b)\big) = \mathrm{E}\big(U_1^2(b)\big) - \Big(\mathrm{E}\big(U_1(b)\big)\Big)^2 = n_{2,L} - n_L^2 + \tfrac{1}{4}(bmB)^2 - bmB\,n_L.$$

In the same way, we find that

$$\mathrm{Var}\big(U_1(a)\big) = \mathrm{E}\big(U_1^2(a)\big) - \Big(\mathrm{E}\big(U_1(a)\big)\Big)^2 = n_{2,H} - n_H^2 + \tfrac{1}{4}(amA)^2 + amA(n_H - 2n_{\mathrm{med}}),$$

where $n_{2,H} = \int_0^{am} y^2 \, \mathrm{d}F(y)$ and $n_{\mathrm{med}} = n(m) = \int_0^m y \, \mathrm{d}F(y)$. Everything here can be straightforwardly estimated in a distribution-free manner.

Alternatively, by setting

$$\hat{u}_{1i}(b) = y_i \, \mathrm{I}(y_i \leq b\hat{m}) - b\hat{m}\hat{B} \, \mathrm{I}(y_i \leq \hat{m}), \text{ and } \hat{u}_{1i}(a) = y_i \, \mathrm{I}(y_i \leq a\hat{m}) - a\hat{m}\hat{A} \, \mathrm{I}(y_i \leq m) \quad (23)$$

for $i = 1, \ldots, N$, the variance of $U_1(b)$ and that of $U_1(a)$ can be estimated by the sample variances of the $\hat{u}_{1i}(b)$ and the $\hat{u}_{1i}(a)$ respectively.

The income share of the low-income group is $IS_L = n_L/\mu$, and this income share can be estimated by $\widehat{IS}_L = \hat{n}_L/\hat{\mu}$. We have

$$\begin{aligned} N^{1/2}(\widehat{IS}_L - IS_L) &= N^{1/2}\Big[\frac{\hat{n}_L}{\hat{\mu}} - \frac{n_L}{\mu}\Big] = N^{1/2}\frac{(\mu\hat{n}_L - \hat{\mu}n_L)}{\mu\hat{\mu}} \\ &= \frac{1}{\mu\hat{\mu}}\big[\mu N^{1/2}(\hat{n}_L - n_L) - n_L N^{1/2}(\hat{\mu} - \mu)\big]. \end{aligned} \quad (24)$$

Now since $\hat{\mu} = \mu + O_p(n^{-1/2})$, for the purposes of our asymptotic analysis we can replace the denominator $\mu\hat{\mu}$ by μ^2. Given (19) and the definition (20) of the random variable $U_1(b)$, and the fact that $N^{1/2}(\hat{\mu} - \mu) = N^{-1/2}\sum_{i=1}^N (y_i - \mu)$, we are led to define the random variable $W(b) = U_1(b)/\mu - n_L Y/\mu^2$ and to conclude that (24) is asymptotically equal in distribution to $W(b) - \mathrm{E}\big(W(b)\big)$. First, note that

$$\mathrm{E}\big(W(b)\big) = n_L/\mu - \tfrac{1}{2}bmB/\mu - n_L/\mu = -\tfrac{1}{2}bmB/\mu. \quad (25)$$

For the variance, we have

$$W^2(b) = U_1^2(b)/\mu^2 + n_L^2 Y^2/\mu^4 - 2\, n_L U_1(b)\, Y/\mu^3.$$

Now

$$\begin{aligned} U_1(b)\, Y &= Y^2 \, \mathrm{I}(Y \leq bm) - bmB\, Y \, \mathrm{I}(Y \leq m) \quad \text{so that} \\ \mathrm{E}\big(U_1(b)\, Y\big) &= n_{2,L} - bmB\, n_{\mathrm{med}}. \end{aligned} \quad (26)$$

Then, from (22) and (26), we see that the asymptotic variance of $\widehat{IS}_L$ is

$$\begin{aligned} \mathrm{Var}\big(W(b)\big) = \mathrm{E}\big(W^2(b)\big) - \big(\mathrm{E}(W(b))\big)^2 &= \frac{1}{\mu^2}\big[n_{2,L} + \tfrac{1}{4}(bmB)^2 - 2bmBn_L\big] \\ &+ \frac{1}{\mu^4}n_L^2\mu_2 - \frac{2n_L}{\mu^3}\big[n_{2,L} - bmB\, n_{\mathrm{med}}\big], \end{aligned} \quad (27)$$

where $\mu_2 = \int_0^\infty y^2 \, \mathrm{d}F(y)$. Note that the term involving B^2 corresponds to the variability of the estimated median-based cut-off point about its true population cut-off value. The terms without any B in them correspond to the variability of random recipients lying within the true median-based cut-off range. And terms involving only B then correspond to the covariance or interaction between the first two components.

Since the income share of the high-income group is $IS_H = 1 - n_H/\mu$, similarly to (24) we see that

$$N^{1/2}(\widehat{IS}_H - IS_H) = -\frac{1}{\mu^2}\big[\mu N^{1/2}(\hat{n}_H - n_H) - n_H N^{1/2}(\hat{\mu} - \mu)\big] + o_p(1).$$

Make the definition $W(a) = U_1(a)/\mu - n_H Y/\mu^2$; the asymptotic variance of $\widehat{IS}_H$ is then $\mathrm{Var}\big(W(a)\big)$, and after some algebra we see that this is

$$\mathrm{Var}\big(W(a)\big) = \frac{1}{\mu^2}\big[n_{2,H} + \tfrac{1}{4}(amA)^2 - 2amAn_{\mathrm{med}}\big] + \frac{1}{\mu^4}n_H^2\mu_2 - \frac{2n_H}{\mu^3}\big[n_{2,H} - amA\, n_{\mathrm{med}}\big]. \tag{28}$$

For the middle-income group, we have $IS_M = (n_H - n_L)/\mu$, and so, again similarly to (24), we find that

$$N^{1/2}(\widehat{IS}_M - IS_M) = \frac{1}{\mu^2}\big[\mu N^{1/2}(\hat{n}_H - \hat{n}_L - n_H + n_L) - (n_H - n_L)N^{1/2}(\hat{\mu} - \mu)\big] + o_p(1).$$

Define the random variable

$$W(a,b) = U_1(a) - U_1(b) - IS_M\, Y.$$

It is easy to check that the asymptotic variance of $\widehat{IS}_M$ is $\mathrm{Var}\big(W(a,b)/\mu\big)$. First,

$$\mathrm{E}\big(W(a,b)\big) = n_H - n_L - m(aA - bB)/2 - IS_M\,\mu = -m(aA - bB)/2.$$

Then

$$W^2(a,b) = \big(U_1(a) - U_1(b)\big)^2 + IS_M^2\, Y^2 - 2IS_M\, Y\big(U_1(a) - U_1(b)\big). \tag{29}$$

Since

$$U_1(a) - U_1(b) = Y\,\mathrm{I}(bm < Y \le am) - m(aA - bB)\mathrm{I}(Y \le m),$$

it follows that

$$\mathrm{E}\big(U_1(a) - U_1(b)\big) = n_H - n_L - m(aA - bB)/2, \tag{30}$$

and since

$$\begin{aligned}\big(U_1(a) - U_1(b)\big)^2 = Y^2\,\mathrm{I}(bm < Y \le am) + m^2(aA - bB)^2\,\mathrm{I}(Y \le m)\\ -2m(aA - bB)Y\,\mathrm{I}(bm < y \le m),\end{aligned}$$

it follows that

$$\mathrm{E}\big[\big(U_1(a) - U_1(b)\big)^2\big] = n_{2,H} - n_{2,L} + m^2(aA - bB)^2/2 - 2m(aA - bB)(n_{\mathrm{med}} - n_L). \tag{31}$$

From (29), (30), and (31) we conclude after a bit of algebra that

$$\begin{aligned}\mathrm{Var}\big(W(a,b)/\mu\big) = \Big[(n_{2,H} - n_{2,L})(1 - 2IS_M) + m^2(aA - bB)^2/4 + \mu_2 IS_M^2\\ +2m(aA - bB)\big((IS_M - 1)n_{\mathrm{med}} + n_L\big)\Big]/\mu^2.\end{aligned} \tag{32}$$

Another way to estimate $\mathrm{Var}\big(W(b)\big)$ and $\mathrm{Var}\big(W(a)\big)$ is to define

$$w_i(b) = \hat{\mu}^{-1}\hat{u}_{1i}(b) - \hat{\mu}^{-2}y_i\hat{n}_L, \text{ and}$$
$$w_i(a) = \hat{\mu}^{-1}\hat{u}_{1i}(a) - \hat{\mu}^{-2}y_i\hat{n}_H$$

for $i = 1, \ldots, N$, and use the sample variances of the $w_i(b)$ and $w_i(a)$ as $\widehat{\mathrm{Var}}\big(W(b)\big)$ and $\widehat{\mathrm{Var}}\big(W(a)\big)$ respectively; recall the definitions (23). Further, if we define

$$w_i(a,b) = \hat{\mu}^{-1}\big[\hat{u}_{1i}(a) - \hat{u}_{1i}(b) - \widehat{IS}_M y_i\big],$$

the sample variance of the $\hat{w}_i(a,b)$ estimates $\mathrm{Var}\big(W(a,b)/\mu\big)$.

2.3 Income Group Means

The mean income of recipients with income no greater than bm is denoted μ_L and is equal to

$$\mu_L \equiv \mathrm{E}\big(Y \mid Y \le bm\big) = \int_0^{bm} y\,\mathrm{d}F(y) \Big/ \int_0^{bm} \mathrm{d}F(y) = n_L/PS_L,$$

estimated by $\hat{\mu}_L \equiv \hat{n}_L/\widehat{PS}_L$. From this, we have to leading order,

$$N^{1/2}(\hat{\mu}_L - \mu_L)) = \frac{1}{PS_L}\Big[N^{1/2}(\hat{n}_L - n_L) - \mu_L N^{1/2}(\widehat{PS}_L - PS_L)\Big].$$

This suggests the definition of a new random variable $X(b)$, as follows:

$$X(b) = \frac{1}{PS_L}\big(U_1(b) - \mu_L U(b)\big) \tag{33}$$

recall the definitions (4) and (20). Then $N^{1/2}(\hat{\mu}_L - \mu_L)$ is asymptotically equal in distribution to $X(b) - \mathrm{E}\big(X(b)\big)$. Details of the calculation of the variance of $X(b)$ are in Appendix 1a, although it is also possible to make the definition for $i = 1, \ldots, N$

$$\hat{x}_i(b) = \frac{1}{\widehat{PS}_L}\big(\hat{u}_{1i}(b) - \hat{\mu}_L\hat{u}_i(b)\big), \tag{34}$$

with $\hat{u}_i(b)$ and $\hat{u}_{1i}(b)$ defined respectively by (6) and (23), and use the sample variance of the $\hat{x}_i(b)$ as an estimate of $\mathrm{Var}(X)$. The calculation in Appendix 1a leads to a rather simple expression for $\mathrm{Var}\big(X(b)\big)$, as follows:

$$\mathrm{Var}\big(X(b)\big) = \frac{1}{PS_L^2}\Big[n_{2,L} - PS_L\mu_L^2 + \tfrac{1}{4}B^2(bm - \mu_L)^2\Big]. \tag{35}$$

Note that

$$\mathrm{Var}(Y \mid Y \le bm) = \frac{\int_0^{bm} y^2\,\mathrm{d}F(y)}{\int_0^{bm}\mathrm{d}F(y)} - \left(\frac{\int_0^{bm} y\,\mathrm{d}F(y)}{\int_0^{bm}\mathrm{d}F(y)}\right)^2 = n_{2,L}/PS_L - \mu_L^2,$$

and so, writing $\sigma_L^2 = \mathrm{Var}(Y \mid Y \le bm)$, we can reformulate (35) as

$$\text{Asy var}(\hat{\mu}_L) = \frac{1}{PS_L^2}(\mu_L - bm)^2 B^2/4 + \frac{1}{PS_L}\sigma_L^2. \tag{36}$$

Note once more that the second term in this expression corresponds to the variance of $\hat{\mu}_L$ based on the true bm cut-off value, while the first term corresponds to the variability associated with the randomness of the cut-off $b\hat{m}$ about its population value bm.

The mean of incomes greater than am is

$$\mu_H \equiv \mathrm{E}(Y \mid Y > am) = \int_{am}^{\infty} y \,\mathrm{d}F(y) \Big/ \int_{am}^{\infty} \mathrm{d}F(y) = \frac{\mu - n_H}{PS_H}, \tag{37}$$

estimated by $\hat{\mu}_H = (\hat{\mu} - \hat{n}_H)/\widehat{PS}_H$. Then $N^{1/2}(\hat{\mu}_H - \mu_H)$ is asymptotically equal in distribution to the random variable

$$X(a) = \frac{1}{PS_H}\Big[Y - U_1(a) - \mu_H\big(1 - U(a)\big)\Big]. \tag{38}$$

minus its expectation. Note that

$$\begin{aligned} 1 - U(a) &= \mathrm{I}(Y > am) + A\,\mathrm{I}(Y \le m), \quad \mathrm{E}\big(1 - U(a)\big) = PS_H + A/2; \\ Y - U_1(a) &= Y\,\mathrm{I}(Y > am) + amA\,\mathrm{I}(Y \le m), \quad \mathrm{E}\big(Y - U_1(a)\big) = \mu - n_H + amA/2, \end{aligned} \tag{39}$$

so that

$$\mathrm{E}\big(X(a)\big) = -\frac{A}{2PS_H}(\mu_H - am). \tag{40}$$

The variance of $X(a)$, derived in detail in Appendix 1b, is

$$\mathrm{Var}\big(X(a)\big) = \frac{1}{PS_H^2}\Big[\mu_2 - n_{2,H} - PS_H\mu_H^2 + \tfrac{1}{4}A^2(\mu_H - am)^2\Big]. \tag{41}$$

Now, if we define the conditional variance

$$\mathrm{Var}(Y \mid Y > am) = n_{2,H}/PS_H - \mu_H^2 \equiv \sigma_H^2,$$

then (41) can also be expressed as

$$\text{Asy var}(\hat{\mu}_H) = \frac{1}{PS_H^2}(\mu_H - am)^2 A^2/4 + \frac{1}{PS_H}\sigma_H^2. \tag{42}$$

Alternatively, for $i = 1, \ldots, N$ make the definition

$$\hat{x}_i(a) = \frac{\big(y_i - \hat{u}_{1i}(a)\big) - \hat{\mu}_H\big(1 - \hat{u}_i(a)\big)}{\widehat{PS}_H}.$$

The variance of the limiting distribution of $N^{1/2}(\hat{\mu}_H - \mu_H))$ can then be estimated by the sample variance of the $\hat{x}_i(a)$. (Recall definitions (6) and (23) for $\hat{u}_i()$ and $\hat{u}_{1i}()$.)
The mean of the incomes between bm and am is

$$\mu_M \equiv \mathrm{E}(Y \mid bm < Y \leq am) = \int_{bm}^{am} y \, \mathrm{d}F(y) \Big/ \int_{bm}^{am} \mathrm{d}F(y) = \frac{n_H - n_L}{PS_M}, \tag{43}$$

estimated by $\hat{\mu}_M = (\hat{n}_H - \hat{n}_L)/\widehat{PS}_M$. Thus $N^{1/2}(\hat{\mu}_M - \mu_M)$ is asymptotically equal in distribution to the random variable

$$\begin{aligned} X(b,a) &= \frac{PS_M\big(U_1(a) - U_1(b)\big) - (n_H - n_L)\big(U(a) - U(b)\big)}{PS_M^2} \\ &= \frac{1}{PS_M}\Big[\big(U_1(a) - U_1(b)\big) - \mu_M\big(U(a) - U(b)\big)\Big], \end{aligned} \tag{44}$$

minus its expectation.
Note that μ_M is not a function of μ_H and μ_L alone. Estimating it poses no problem, but a new calculation is needed to find an expression for its asymptotic variance. The variance of $X(b,a)$ is derived in Appendix 1c. It is

$$\begin{aligned} &\mathrm{Var}\big(X(b,a)\big) = \mathrm{E}\big[W^2(b,a)\big] - \big[\mathrm{E}\big(X(b,a)\big)\big]^2 \\ &= \frac{1}{PS_M^2}\Big[n_{2,H} - n_{2,L} - \mu_M^2 PS_M + D^2/4 + D\big[2(n_{\mathrm{med}} - n_L) + 2\mu_M PS_L - \mu_M\big]\Big], \end{aligned} \tag{45}$$

where we have made the definition

$$D = \mu_M(A - B) - m(aA - bB). \tag{46}$$

Another conditional variance:

$$\mathrm{Var}(Y \mid bm < Y \leq am) = (n_{2,H} - n_{2,L})/PS_M - \mu_H^2 \equiv \sigma_M^2,$$

so that (45) reformulated becomes

$$\text{Asy var}(\hat{\mu}_M) = \frac{1}{PS_M^2}\big[D^2/4 + D\big(2(n_{\mathrm{med}} - n_L) + 2\mu_M PS_L - \mu_M\big)\big] + \frac{1}{PS_M}\sigma_M^2. \tag{47}$$

In order to estimate the variance of the limiting distribution of $N^{1/2}(\hat{\mu}_M - \mu_M)$, another way to proceed is to make the definition, for $i = 1, \ldots, N$,

$$\hat{x}_i(b,a) = \frac{\hat{u}_{1i}(a) - \hat{u}_{1i}(b) - \hat{\mu}_M\big(\hat{u}_i(a) - \hat{u}_i(b)\big)}{\widehat{PS}_M}$$

and use the sample variance of the $\hat{x}_i(b,a)$ to estimate the desired variance.

2.4 Summary of Main Results

a Population shares

From the results (8), (10), and (17), we obtain directly that

$$\begin{aligned}
\text{Asy var}(\widehat{PS}_L) &= PS_L(1-PS_L) + B^2/4 - PS_L B, \\
&\quad \text{where } B = bf(bm)/f(m); \\
\text{Asy var}(\widehat{PS}_H) &= PS_H(1-PS_H) + A^2/4 - PS_H A, \\
&\quad \text{where } A = af(am)/f(m); \\
\text{Asy var}(\widehat{PS}_M) &= PS_M(1-PS_M) + C^2/4 - (PS_H - PS_L)C, \\
&\quad \text{where } C = A - B.
\end{aligned}$$

b Income shares

From the results (27), (28), and (32), we obtain

$$\text{Asy var}(\widehat{IS}_L) = \frac{1}{\mu^2}\left[n_{2,L} + \tfrac{1}{4}(bmB)^2 - 2bmBn_L\right] + \frac{1}{\mu^4}n_L^2\mu_2 - \frac{2n_L}{\mu^3}\left[n_{2,L} - bmB\, n_{\text{med}}\right];$$

$$\text{Asy var}(\widehat{IS}_H) = \frac{1}{\mu^2}\left[n_{2,H} + \tfrac{1}{4}(amA)^2 - 2amAn_{\text{med}}\right] + \frac{1}{\mu^4}n_H^2\mu_2 - \frac{2n_H}{\mu^3}\left[n_{2,H} - amA\, n_{\text{med}}\right];$$

$$\begin{aligned}
\text{Asy var}(\widehat{IS}_M) = \frac{1}{\mu^2}\Big[&(n_{2,H} - n_{2,L})(1 - 2IS_M) + m^2(aA - bB)^2/4 + \mu_2 IS_M^2 \\
&+2m(aA - bB)((IS_M - 1)n_{\text{med}} + n_L)\Big].
\end{aligned}$$

c Income group means

From the results (35), (41), and (45), we obtain

$$\begin{aligned}
\text{Asy var}(\hat{\mu}_L) &= \frac{1}{PS_L^2}\left[n_{2,L} - PS_L\mu_L^2 + \tfrac{1}{4}B^2(bm - \mu_L)^2\right]; \\
\text{Asy var}(\hat{\mu}_H) &= \frac{1}{PS_H^2}\left[\mu_2 - n_{2,H} - PS_H\mu_H^2 + \tfrac{1}{4}A^2(\mu_H - am)^2\right]; \\
\text{Asy var}(\hat{\mu}_M) &= \frac{1}{PS_M^2}\Big[n_{2,H} - n_{2,L} - \mu_M^2 PS_M + D^2/4 + D\big(2(n_{\text{med}} - n_L) \\
&\qquad + 2\mu_M PS_L - \mu_M\big)\Big], \text{ where } D = \mu_M(A - B) - m(aA - bB).
\end{aligned}$$

Expressed somewhat differently, these results also follow from (36), (42), and (47):

$$\begin{aligned}
\text{Asy var}(\hat{\mu}_L) &= \frac{1}{PS_L^2}(\mu_L - bm)^2B^2/4 + \frac{1}{PS_L}\sigma_L^2; \\
\text{Asy var}(\hat{\mu}_H) &= \frac{1}{PS_H^2}(\mu_H - am)^2A^2/4 + \frac{1}{PS_H}\sigma_H^2; \\
\text{Asy var}(\hat{\mu}_M) &= \frac{1}{PS_M^2}\left[D^2/4 + D\big(2(n_{\text{med}} - n_L) + 2\mu_M PS_L - \mu_M\big)\right] + \frac{1}{PS_M}\sigma_M^2.
\end{aligned}$$

Note that the general framework of this paper allows for more and for more refined income groups than just the three employed here – so long as the cut-off points between income groups are expressed in terms of multiples of the median.

3. Inference on Related Distributional Statistics

This section considers three sets of distributional statistics that involve applications of the analytical results developed in the previous section. As there, we restrict attention to the case in which $b < 1 < a$, thus defining three income groups: the lower group L, for incomes less than or equal to bm; the middle group M, with incomes between bm and am; and the higher group H, with incomes greater than am.

3.1 Relative Mean Income Ratios

The relative mean income for each income group is the ratio of the group's mean income to the overall mean income of the distribution:

$$\mathrm{RMI}_i = \mu_i/\mu \quad \text{for } i = L, M, H. \tag{48}$$

It shows the size of the discrepancy or distance of group mean incomes to the overall mean in proportional terms. So, for example, in recent decades for many countries, the lower-income ratio $\hat{\mu}_L/\hat{\mu}$ has not changed much, while the upper-income ratio $\hat{\mu}_H/\hat{\mu}$ has gone up very substantially. It would be useful to know if the changes in both ratios are statistically significant, or only the latter.

The relative mean income ratio can be estimated directly as

$$\widehat{\mathrm{RMI}}_i = \hat{\mu}_i/\hat{\mu}.$$

But, from the definitions of μ_L, μ_H, and μ_M, we have $\mu_L/\mu = n_L/(\mu\, PS_L) = IS_L/PS_L$, $\mu_H/\mu = (\mu - n_H)/(\mu\, PS_H) = IS_H/PS_H$, and $\mu_M/\mu = (n_H - n_L)/(\mu\, PS_M) = IS_M/PS_M$, and so for $i = L, M, H$, $\mathrm{RMI}_i = IS_i/PS_i$. Thus to leading order

$$N^{1/2}(\widehat{\mathrm{RMI}}_i - \mathrm{RMI}_i) = \frac{1}{PS_i}\big[N^{1/2}(\widehat{IS}_i - IS_i) - \mathrm{RMI}_i\, N^{1/2}(\widehat{PS}_i - PS_i)\big]. \tag{49}$$

Consider first $\widehat{\mathrm{RMI}}_L$. With $i = L$, (49) suggests the random variable

$$R(b) = W(b) - \mathrm{RMI}_L\, U(b) = 1/\mu\, U_1(b) - n_L Y/\mu^2 - \mathrm{RMI}_L\, U(b).$$

The asymptotic variance of $\widehat{\mathrm{RMI}}_L$ is then $\mathrm{Var}\big(R(b)/PS_L\big)$. An easy calculation shows that

$$\mathrm{E}\big(R(b)\big) = \tfrac{1}{2}B(\mathrm{RMI}_L - bm/\mu) - IS_L. \tag{50}$$

Then

$$R^2(b) = W^2(b) + \mathrm{RMI}_L^2\, U^2(b) - 2\mathrm{RMI}_L\, U(b)W(b).$$

The expectation of $W^2(b)$ follows from (27) and that of $U^2(b)$ is given by (7). We have

$$U(b)W(b) = \frac{1}{\mu}U(b)U_1(b) - \frac{n_L}{\mu^2}U(b)\,Y.$$

It is easy to show that

$$\begin{aligned} U(b)\,U_1(b) &= Y\,\mathrm{I}(Y \le bm)(1-B) - bmB\,\mathrm{I}(Y \le bm) + bmB^2\,\mathrm{I}(Y \le m) \text{ and} \\ U(b\,)Y &= Y\,\mathrm{I}(Y \le bm) - BY\,\mathrm{I}(Y \le m), \end{aligned}$$

so that

$$\mathrm{E}\big(U(b)\,U_1(b)\big) = n_L(1-B) - bmB\,PS_L + bmB^2/2 \text{ and} \tag{51}$$

$$\mathrm{E}\big(U(b)\,Y\big) = n_L - B\,n_{\mathrm{med}}. \tag{52}$$

Thus we have

$$\begin{aligned} \mathrm{E}\big(U(b)W(b)\big) &= \frac{1}{\mu}\big[n_L(1-B) - bmB\,PS_L + bmB^2/2\big] - \frac{1}{\mu^2}n_L(n_L - B\,n_{\mathrm{med}}) \\ &= PS_L(1-PS_L) - B\,IS_L(1 - n_{\mathrm{med}}/\mu) - bmB\,PS_L + \tfrac{1}{2}bmB^2. \end{aligned} \tag{53}$$

Some algebra lets us calculate $\mathrm{Var}(\mathrm{RMI}_L)$ from (27), (7), (50), and (53). The result is

$$\begin{aligned} \text{Asy var}(\widehat{\mathrm{RMI}}_L) &= \mathrm{Var}\big(R(b)/PS_L\big) = \\ &\frac{1}{PS_L^2}\Big[(1-2IS_L)(n_{2,L}/\mu^2 - IS_L\,\mathrm{RMI}_L) + IS_L^2\mu_2/\mu^2 + \tfrac{1}{4}B^2(\mathrm{RMI}_L - bm/\mu)^2 \\ &\quad - IS_L^2 - IS_L B(\mathrm{RMI}_L - bm/\mu)(2n_{\mathrm{med}}/\mu - 1)\Big], \end{aligned} \tag{54}$$

Similarly, for $\widehat{\mathrm{RMI}}_H$, we consider the random variable

$$R(a) = W(a) - RMI_H U(a) = 1/\mu U_1(a) - n_H Y/\mu^2 - RMI_H\,U(a),$$

and

$$R^2(a) = W^2(a) + RMI_H^2\,U^2(a) - 2RMI_H U(a)W(a). \tag{55}$$

First,

$$\mathrm{E}\big(R(a)\big) = \tfrac{1}{2}A(RMI_H - am/\mu) + IS_H - RMI_H.$$

From (28), we can deduce that

$$\begin{aligned} \mathrm{E}\big(W^2(a)\big) &= \frac{1}{\mu^2}\big[n_{2,H} + \tfrac{1}{2}(amA)^2 - 2amAn_{\mathrm{med}})\big] + \frac{n_H^2\mu_2}{\mu^4} - \frac{2n_H}{\mu^3}\big[n_{2,H} - amAn_{\mathrm{med}}\big] \\ &= \frac{1}{\mu^2}\big[n_{2,H}(2IS_H - 1) + \tfrac{1}{2}(amA)^2 - 2amA\,IS_H\,n_{\mathrm{med}}\big]. \end{aligned} \tag{56}$$

It is immediate from (9) that

$$\mathrm{E}\big(U^2(a)\big) = 1 - PS_H + A^2/2 - A. \tag{57}$$

Analogously to (51) and (52), we find that

$$\mathrm{E}\big(U(a)U_1(a)\big) = n_H - An_{\mathrm{med}} - amA/2 + amA^2/2 \text{ and} \tag{58}$$

$$\mathrm{E}\big(U(a)\,Y\big) = n_H - An_{\mathrm{med}}. \tag{59}$$

Now $U(a)W(a) = U(a)U_1(a)/\mu - n_H U(a)\,Y/\mu^2$, and so, from (58) and (59), we see that

$$\begin{aligned}\mathrm{E}\big(U(a)W(a)\big) &= \frac{1}{\mu}(n_H - An_{\mathrm{med}} - \tfrac{1}{2}amA + \tfrac{1}{2}amA^2) - \frac{n_H}{\mu^2}(n_H - An_{\mathrm{med}}) \\ &= IS_H(1 - IS_H) - A\,IS_H\,n_{\mathrm{med}}/\mu - amA(1-A)/(2\mu).\end{aligned} \tag{60}$$

And so, from (55), (56), (57), and (60), we obtain the result

$$\begin{aligned}&\text{Asy var}(\widehat{\mathrm{RMI}}_H) = \mathrm{Var}\big(R(a)/PS_H\big) = \\ &\frac{1}{PS_H^2}\Big[(2IS_H - 1)n_{2,H}/\mu^2 + n_H^2\mu_2/\mu^4 + \tfrac{1}{4}A^2(\mathrm{RMI}_H - am/\mu)^2 \\ &- IS_H^2 + IS_H^2/PS_H - 2\,IS_H^2/PS_H^2 + IS_H\,A(\mathrm{RMI}_H - am/\mu)(2n_{\mathrm{med}}/\mu - 1)\Big].\end{aligned} \tag{61}$$

Although we can derive the asymptotic variance of $\widehat{\mathrm{RMI}}_M$ along similar lines as above for $\widehat{\mathrm{RMI}}_L$ and $\widehat{\mathrm{RMI}}_H$, this leads to expressions that are neither simple nor intuitive. A simpler procedure is to note from (49) that the asymptotic variance of $\widehat{\mathrm{RMI}}_M$ is equal to

$$\frac{1}{PS_M^2}\big[\text{Asy var}(\widehat{IS}_M) + \mathrm{RMI}_M^2\,\text{Asy var}(\widehat{PS}_M) - 2\,\mathrm{RMI}_M\,\text{Asy cov}(\widehat{IS}_M, \widehat{PS}_M)\big].$$

The asymptotic variances of $\widehat{IS}_M$ and $\widehat{PS}_M$ are given by (32) and (17) respectively; see also the summary of results.

The asymptotic covariance of $\widehat{PS}_M$ and $\widehat{IS}_M$ is the covariance of $U(a) - U(b)$ and $W(a) - W(b)$. The details of the calculation of the covariance are relegated to Appendix 1d. The result is

$$\begin{aligned}\text{Asy cov}(\widehat{PS}_M, \widehat{IS}_M) &= IS_M(1 - IS_M) - \frac{1}{2\mu}m(aA - bB)(PS_H - PS_L) \\ &+ \frac{C}{\mu}\big[IS_M n_{\mathrm{med}} - n_{\mathrm{med}} + n_L + \tfrac{1}{4}m(aA - bB)\big],\end{aligned} \tag{62}$$

with $C = A - B$.

3.2 Polarization Measures

The rise of upper incomes, resulting in a growing separation between high-income recipients and middle-class workers, has led to concern about the degree of polarization in income distributions. 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 either or both 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)) – what we are referring to here as PS_L and PS_H. Such measures then are $\widehat{PS}_L$, $\widehat{PS}_H$, and $\widehat{PS}_L + \widehat{PS}_H$. Asymptotic variances for the first two have already been obtained in Section 2 above. For $\widehat{PS}_L + \widehat{PS}_H$, note that the sum of the three population shares is one, and so the asymptotic variance of $\widehat{PS}_L + \widehat{PS}_H$ is simply that of the middle group, $\widehat{PS}_M$, which again we already have in (17).

The other aspect of polarization 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-gap polarization, and could be captured by $\hat{\mu}_H - \hat{\mu}_M$, $\hat{\mu}_M - \hat{\mu}_L$, or $\hat{\mu}_H - \hat{\mu}_L$. Both sets of measures provide useful insights, and both can be implemented in our analytical framework. In the case of the income-gap polarization measures, again the asymptotic variances of $\hat{\mu}_H$, $\hat{\mu}_M$, and $\hat{\mu}_L$ have been established in Section 2. For the differences in income group means, recall that

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

for $i \neq j$. The three required covariances are provided in Appendix 1e. Thus, again, standard errors of the income-gap polarization measures can be computed in the usual fashion.

One could also posit a set of compound polarization measures which capture both of these dimensions together: $CP_H \equiv PS_H(\mu_H - \mu_M))$, $CP_L \equiv PS_L(\mu_M - \mu_L)$, and $CP \equiv (PS_H + PS_L)(\mu_H - \mu_L) = (1 - PS_M)(\mu_H - \mu_L)$.

Analogously, one could further identify a compound measure to capture the evident decline in the economic situation of the Middle Class in many countries over recent decades as $PS_M \cdot \mu_M$. This would allow one, for example, to use logarithmic derivatives to estimate the relative importance of changes in the relative size of the Middle Class (ΔPS_M) versus changes in their average real incomes ($\Delta\mu_M$) in this decline.

One can use the results of Section 2 to work out the asymptotic variances of these various estimated compound measures; see Appendix 1f for details.

3.3 Mean-Decile Functions

In an environment where higher incomes have been rising dramatically relative to the rest of the distribution, one measure of interest could be an indication of skewness of the distribution, as measured by the difference between the overall mean and median of the income distribution, $\hat{\mu} - \hat{m}$ or $\hat{m}/\hat{\mu}$. However, $\hat{m}$ is simply the fifth decile of the distribution. One could, more generally, define a mean-decile function.

Choose some proportions p_i, $i = 1, \ldots, m$ with $p_i < p_j$ for $i < j$. For deciles, we would have $p_i = i/10$, $i = 1, 2, \ldots, 9$. Let ξ_i be the p_i-quantile of the distribution: the proportion of incomes less than ξ_i is p_i, and let $\hat{\xi}_i$ be the corresponding sample quantile. Possible mean-decile functions could take on values $\hat{\xi}_i - \hat{\mu}$, or alternatively $\hat{\xi}_i/\hat{\mu}$, for the i^{th} decile of the distribution as a further way of capturing growing income differences over various ranges of the distribution.

Here we can make use of the work of Lin, Wu, and Ahmad (1980) (LWA). LWA show that, under general regularity conditions, the $\hat{\xi}_i$ and $\hat{\mu}$ are asymptotically joint normally distributed. We denote the asymptotic variance-covariance matrix by $\boldsymbol{\Sigma}$: it is an $(m+1) \times (m+1)$ matrix, where the index $i = 0$ refers, not to a quantile, but to μ. Then, for $0 < i \leq j \leq m$, the elements of $\boldsymbol{\Sigma}$ are:

$$\begin{aligned}
\sigma_{ij} &= p_i(1 - p_j)/\big[f(\xi_i)f(\xi_j)\big], \\
\sigma_{00} &= \sigma^2, \\
\sigma_{0i} &= p_i(\mu - \mu_i)/f(\xi_i),
\end{aligned}$$

where $f(\xi_i)$ is the density at ξ_i, $\mu_i = \mathrm{E}(Y \mid Y \leq \xi_i) = (1/p_i) \int_0^{\xi_i} y \, \mathrm{d}F(y)$, and $\sigma^2 = \mathrm{Var}(Y)$. Thus, for the mean-decile distribution defined in levels as $\hat{\xi}_i - \hat{\mu}$, we have

$$\begin{aligned}
\text{Asy var}(\hat{\xi}_i - \hat{\mu}) &= \text{Asy var}(\hat{\xi}_i) + \text{Asy var}(\hat{\mu}) - 2\,\text{Asy cov}(\hat{\xi}_i, \hat{\mu}) \\
&= \frac{p_i(1 - p_i)}{f^2(\xi_i)} + \sigma^2 - \frac{2p_i(\mu - \mu_i)}{f(\xi_i)}.
\end{aligned}$$

In relative or proportional terms,

$$\begin{aligned}
\text{Asy var}(\hat{\xi}_i/\hat{\mu}) &= \begin{bmatrix} 1/\mu & -\xi_i/\mu^2 \end{bmatrix} \boldsymbol{\Sigma}_{0i} \begin{bmatrix} 1/\mu \\ -\xi_i/\mu^2 \end{bmatrix} \\
&= \frac{\sigma^2}{\mu^2} + \frac{\xi_i^2 p_i(1 - p_i)}{\mu^4 f^2(\xi_i)} - \frac{2\xi_i p_i(\mu - \mu_i)}{\mu^3 f(\xi_i)}.
\end{aligned}$$

Note that the density appears as such in the denominator of the above expressions rather than as a ratio $f(am)/f(m)$ or $f(bm)/f(m)$ as elsewhere in this paper. But $f(\xi_i)$ can be estimated in the same way as the other densities used; see Appendix 2. Standard errors can be calculated accordingly.

3.4 Relation with the Bootstrap

Given the fact that the bootstrap has become an almost universal tool for reliable statistical inference, it is incumbent on us to outline how the material in this paper can be used in connection with bootstrap methods. It has been suggested that the asymptotic variances and standard errors provided here are unnecessary, as they can be obtained in a finite-sample context by use of the bootstrap. However, Horowitz (2001) points out that naive bootstrap standard errors are unlikely to be any better than asymptotic ones and may

well be worse. What he and numerous other authors recommend is using an asymptotic standard error in order to construct an asymptotically pivotal quantity by studentising, that is, dividing the quantity of interest, supposed to have expectation zero, by its standard error. The studentised quantity can then be bootstrapped in order to obtain a bootstrap P value for some null hypothesis, or to construct a bootstrap confidence interval for a parameter of interest.

Our results can be applied readily to such a bootstrap exercise. For instance, a test of a hypothesis that PS_M is equal to some given value M can be based on bootstrapping $(\widehat{PS}_M - M)/se_{PSM}$, where se_{PSM} is the square root of the asymptotic variance of $\widehat{PS}_M$ given by (17). Similarly a bootstrap confidence for PS_M can be constructed by conventional means.

Another reason to exercise care in applying the bootstrap to the data used in this paper is set out in Davidson (2018). The incomes given for individuals in the census data are often, indeed usually, rounded to multiples of \$500 or \$1000. This means that the empirical distribution of the sample of incomes is not smooth, and this is known to cause problems for a conventional resampling bootstrap. We verified that this is the case with our samples. Asymptotic variances as given by the formulas of this paper, and variances derived from a conventional resampling bootstrap, were compared in the context of a simulation experiment that used samples of 200000 observations realised from a lognormal distribution. Results were comparable, as might be expected with such large samples. When the same exercise was repeated with the sample of men's incomes in 2000, the bootstrap variances were very different from the asymptotic ones.

Another point of interest for practitioners is that all the asymptotic standard errors reported in Table 3 were computed in a quarter of a second, whereas the corresponding bootstrap standard errors, with 999 bootstrap repetitions, took 80 seconds.

If, as for instance with stratified sampling, observations are not equally weighted, our analysis can then be applied if the number of actual observations N is replaced by the sum of the weights over the sample.

4. Empirical Study

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. In all the results discussed in this section, earnings are expressed in thousands of 2005 (Canadian) dollars.

Density estimates were given by the approach outlined in Appendix 2. We experimented with different values of the parameter n using samples drawn from the lognormal distribution, for which the density is known analytically. It appeared that a larger value of n gave more accurate estimates, but that numerical overflow occurred in the computation of the gamma function for values of n greater than around 170. We found that setting $n = 100$ gave satisfactory results, although other choices in the neighbourhood of 100 gave results that were not markedly different.

Results are given in Appendix 3. In Table 1, results are shown for men in 2000. The entries for $\hat{\xi}$ are the upper income cutoff for group L, and the lower income cutoff for group H. For group M the entry is the sample median. Asymptotic standard errors are in brackets.

Table 2 shows the corresponding results for women in 2000.

In Table 3 and Table 4 there are similar results for men and women respectively in 2005.

The sample sizes for these four tables of basic distributional results are quite large (202,491 and 238,356). So it should perhaps not be surprising that the asymptotic standard errors are quite small, and all the reported statistics in these basic tables are highly statistically significant. They involve averages or proportions, which seem to be robustly estimated. The estimates of A and B are also all quite sensible in that they imply that the estimated density ratio $\hat{f}(b\hat{m})/\hat{f}(\hat{m})$ is considerably larger than $\hat{f}(a\hat{m})/\hat{f}(\hat{m})$ – which is what one would expect for a right-skewed distribution such as for an earnings distribution.

Table 5 and Table 6 show the differences in outcomes between men and women for the years 2000 and 2005, with asymptotic standard errors for these differences in parentheses. A positive difference means that the relevant outcome is greater for men than for women; a negative difference the reverse. Again, all the differences are highly statistically significant. Two results are evident. In both years, men are relatively more concentrated in the middle-income group with women relatively more concentrated in the lower- and higher- income groups within each distribution. This is consistent with more part-time women workers as well as generally higher levels of education for women than for men in recent decades. Second, the earnings gap between men and women changes very little within the lower and middle income groups over 2000–2005. But in the higher income group, men's earnings shot up quite dramatically compared to women's over this period.

Table 7 and Table 8 present differences or changes over time in the distributional outcome measures between 2000 and 2005, separately for men and women. For outcomes that are greater in 2005 than in 2000, the differences are positive. Again, asymptotic standard errors are in parentheses, and again all but one of the changes are highly statistically significant. Here the changes are quite dramatic given that major distributional changes have typically been rather slow and gradual over time. For both men and women, the proportion of workers in the middle-income group fell substantially between 2000 and 2005 as did relative-mean incomes of the middle group. On the other hand, mean earnings levels in the higher-income group went up dramatically. As a result, the earnings share of the middle group of so-called middle-class earners markedly declined and was made up

by a corresponding dramatic rise in the earnings share of the higher-income group. This pattern occurred for both women and men in the Canadian labour market between 2000 and 2005, but the changes were two to three times stronger in the earnings distribution for men than for women.

Table 9 and Table 10 further pursue this significant pattern of change and show results for several measures of polarization within the earnings distributions (see section 3.2 above). Table 9 focuses on population shares or the proportion of workers towards the two ends of the distributions, while Table 10 bases alternative polarization measures on mean earnings gaps over the ends of the distributions. Again, in both sets of polarization measures, one finds broadly similar patterns of change for both men and women (though with some differences). In the case of PS-based measures (Table 9), the general polarization of workers out of the middle-class region was driven by an increased proportion of workers in the H earnings group among men, but by an increased proportion of workers in the L earnings group among women. In the case of the earnings-gap measures (Table 10), the greatly widening gaps in earnings between groups in the distributions is almost entirely driven by the widening gap between middle-class and higher earnings levels – for both men and women in the labour market. Again, the changes are about twice as strong among men than among women workers, and again the results are highly statistically significant.

Finally, Table 11 and Table 12 display estimates of and changes in the compound polarization measures (in section 3.2) that combine the population share and earnings gap dimensions. As can be seen, for both men and women, changes in the upper end of the earnings distributions over the 2000–2005 period were much greater than changes in the lower end of the distributions. For women, the changes were about twice as big, while for men it was about eight times. Clearly, the big changes have been occurring between the middle-class earnings group and the higher-earnings group. This recommends the use of separate polarization measures for the lower and upper ends of the distribution rather than one that blends or combines the two and thus potentially hides the basic structural changes that are going on over the different regions of the distribution and in the Canadian labour market. Note also that, for men, both components of CP_H contribute to the big increases in earnings polarization – both increases in PS_H as well as the rising earnings gap $(\mu_H - \mu_M)$ – while for women, the increase in CP_H is driven completely by rapidly rising upper earnings levels. Again, these polarization changes are all highly statistically significant.

5. Conclusions

This paper considers income distributions that are divided into lower, middle and upper regions based on separating points that are scalar multiples of the median. For example, the lower region (L) could consist of recipients with incomes less than half the median, the middle group (M) includes those with incomes between 50 percent and 200 percent of the median, and those with incomes above twice the median lie in the higher income group (H). Such a characterization of an income distribution is very useful in evaluating changes over time in the economic experience of the middle-class income group and in the nature of polarization in the distribution. For each of these three income groups, separate

estimates are obtained for their income shares (IS_i), group size or population shares (PS_i) and their mean income levels (μ_i). The paper derives explicit formulas for the asymptotic variances (and hence standard errors) of sample estimates of the groups' population shares, income shares, and mean incomes. It is shown that these formulas are not distribution-free, but that a density-estimation technique of Comte and Genon-Catalot (2012) is well suited to provide needed data-based density estimates in empirical income distribution analyses. The results are then applied to derive asymptotic variances for relative-mean income ratios, for each income group, for various polarization measures, and for decile-mean income ratios. This statistical framework is implemented with Canadian Census public-use microdata files in order to investigate some of the key features of changes in the Canadian earnings distribution.

The empirical findings show that, with such large microdata sets, population and income shares and income-group means can indeed be estimated with a high degree of reliability. Major patterns of distributional change that have been previously highlighted in the literature have indeed been found to be highly statistically significant. The distributional framework and statistical approach used in this paper thus allow one to move beyond descriptive analysis of distributional change to a formal framework of statistical inference and hypothesis testing.

Further, since $IS_i = PS_i \cdot \mathrm{RMI}_i$, changes in group income shares have been found to arise from changes in both population shares and relative mean incomes. Estimating these two dimensions separately allows for (i) a rich economic interpretation and testing of the driving factors behind distributional change, and (ii) an extensive characterization (and hence better understanding) of polarization as a key aspect of on-going distributional change.

The results of this paper suggest that official government statistical agencies – such as Statistics Canada and the U.S. Bureau of the Census – may wish to consider providing median-based estimates of population shares, income shares and income-group means to complement their regularly published series on decile income shares and decile means. They could also provide user information on the general reliability of these estimates.

Appendix 1: Detailed Calculations

a: Variance of $\hat{\mu}_L$:

Recall from (33) that

$$X(b) = \frac{1}{PS_L}\big(U_1(b) - \mu_L U(b)\big).$$

Since $\mathrm{E}(U(b)) = PS_L - B/2$, $\mathrm{E}(U_1(b)) = n_L - bmB/2$, and $\mu_L = n_L/PS_L$, it follows that

$$\mathrm{E}\big(X(b)\big) = \frac{B}{2PS_L}(\mu_L - bm). \tag{63}$$

Next,

$$X^2(b) = \frac{1}{PS_L^2}\Big(U_1^2(b) + \mu_L^2 U^2(b) - 2\mu_L U(b)U_1(b)\Big).$$

For the expectation of this, we use (7) for $\mathrm{E}(U^2(b))$, (22) for $\mathrm{E}(U_1^2(b))$, and (51) for $\mathrm{E}\big(U(b)U_1(b)\big)$. Thus

$$\mathrm{E}\big(X^2(b)\big) = \frac{1}{PS_L^2}\Big[n_{2,L} + \tfrac{1}{2}(bmB)^2 - 2bmBn_L$$
$$+\mu_L^2\big(PS_L(1-2B) + \tfrac{1}{2}B^2\big) - 2\mu_L n_L(1-B) + 2bmBn_L - \mu_L bmB^2\Big].$$

By collecting coefficients of powers of B, we see that

$$\mathrm{E}\big(X^2(b)\big) = \frac{1}{PS_L^2}\Big[n_{2,L} - n_L\mu_L + \tfrac{1}{2}B^2(bm-\mu_L)^2\Big],$$

while from (63) we have

$$\big(\mathrm{E}(X(b))\big)^2 = \frac{B^2}{4PS_L^2}(bm-\mu_L)^2,$$

and so

$$\mathrm{Var}\big(X(b)\big) = \frac{1}{PS_L^2}\Big[n_{2,L} - n_L\mu_L + \tfrac{1}{4}B^2(bm-\mu_L)^2\Big],$$

b: Variance of $\hat{\mu}_H$

From (38) we have

$$X(a) = \frac{1}{PS_H}\Big[Y - U_1(a) - \mu_H\big(1-U(a)\big)\Big],$$

and so

$$X^2(a) = \frac{1}{PS_H^2}\Big[\big(Y-U_1(a)\big)^2 + \mu_H^2\big(1-U(a)\big)^2 - 2\mu_H\big(Y-U_1(a)\big)\big(1-U(a)\big)\Big].$$

From (39), it is easy to see that

$$\begin{aligned}\mathrm{E}\big[\big(Y-U_1(a)\big)^2\big] &= \mu_2 - n_{2,H} + (amA)^2/2;\\ \mathrm{E}\big[\big(1-U(a)\big)^2\big] &= PS_H + A^2/2;\\ \mathrm{E}\big[\big(Y-U_1(a)\big)\big(1-U(a)\big)\big] &= \mu - n_H + amA^2/2.\end{aligned}$$

It then follows that $\mathrm{E}\big(X^2(a)\big)$ is

$$\begin{aligned}&\frac{1}{PS_H^2}\Big[\mu_2 - n_{2,H} + (amA)^2/2 + \mu_H^2(PS_H + A^2/2) - 2\mu_H\big[\mu - n_H + amA^2/2\Big]\\ =&\frac{1}{PS_H^2}\big(\mu_2 - n_{2,H} - PS_H\mu_H^2 + A^2(\mu_H - am)^2/2\big).\end{aligned}$$

From (40) we have

$$\big[\mathrm{E}\big(X(a)\big)\big]^2 = \frac{A^2}{4PS_H^2}\big(\mu_h - am\big)^2,$$

and so we conclude that the asymptotic variance of $\hat{\mu}_H$ is

$$\mathrm{Var}\big(X(a)\big) = \frac{1}{PS_H^2}\Big[\mu_2 - n_{2,H} - PS_H\mu_H^2 + \tfrac{1}{4}A^2(\mu_H - am)^2\Big]. \tag{64}$$

c: Variance of $\hat{\mu}_M$

Recall the definition (44):

$$X(b,a) = \frac{1}{PS_M}\Big[\big(U_1(a) - U_1(b)\big) - \mu_M\big(U(a) - U(b)\big)\Big],$$

whence

$$\begin{aligned}X^2(b,a) = \frac{1}{PS_M^2}\Big[&\big(U_1(a) - U_1(b)\big)^2 + \mu_M^2\big(U(a) - U(b)\big)^2\\ &-2\mu_M\big(U_1(a) - U_1(b)\big)\big(U(a) - U(b)\big)\Big].\end{aligned} \tag{65}$$

Note that

$$\begin{aligned}U(a) - U(b) &= \mathrm{I}(bm < Y \le am) - (A-B)\,\mathrm{I}(Y \le m) \text{ and}\\ U_1(a) - U_1(b) &= Y\,\mathrm{I}(bm < Y \le am) - m(aA - bB)\,\mathrm{I}(Y \le m),\end{aligned}$$

so that $\mathrm{E}\big(U(a-U(b))\big) = PS_M - (A-B)/2$ and $\mathrm{E}\big(U_1(a) - U_1(b)\big) = n_H - n_L - m(aA - bB)/2$. Further,

$$\big(U(a) - U(b)\big)^2 = \mathrm{I}(bm < Y \le am) + (A-B)^2\,\mathrm{I}(Y \le m) - 2(A-B)\,\mathrm{I}(bm < Y \le m),$$

so that

$$\mathrm{E}\big[\big(U(a)-U(b)\big)^2\big] = PS_M + (A-B)^2/2 - (A-B)(1-2PS_L). \tag{66}$$

Next,

$$\begin{aligned}\big(U_1(a)-U_1(b)\big)^2 = Y^2\,\mathrm{I}(bm < Y \le am) + m^2(aA-bB)^2\,\mathrm{I}(Y \le m)\\ - 2m(aA-bB)Y\,\mathrm{I}(bm < Y \le m),\end{aligned}$$

so that

$$\mathrm{E}\big[\big(U_1(a)-U_1(b)\big)^2\big] = n_{2,H} - n_{2,L} + m^2(aA-bB)^2/2 - 2m(aA-bB)(n_{\mathrm{med}} - n_L). \tag{67}$$

Then

$$\begin{aligned}\big(U(a)-U(b)\big)\big(U_1(a)-U_1(b)\big) = Y\,\mathrm{I}(bm < Y \le am) + m(A-B)(aA-bB)\,\mathrm{I}(Y \le m)\\ - \big[(A-B)Y + m(aA-bB)\big]\,\mathrm{I}(bm < Y \le m),\end{aligned}$$

so that

$$\begin{aligned}\mathrm{E}\big[\big(U(a)-U(b)\big)\big(U_1(a)-U_1(b)\big)\big] = n_H - n_L + m(A-B)(aA-bB)/2\\ - (A-B)(n_{\mathrm{med}} - n_L) - m(aA-bB)(\tfrac{1}{2} - PS_L).\end{aligned}$$

From all this, we see that

$$\begin{aligned}\mathrm{E}\big[X^2(b,a)\big] = \frac{1}{PS_M^2}\Big[n_{2,H}-n_{2,L}+\tfrac{1}{2}\big[m^2(aA-bB)^2+\mu_M^2(A-B)^2-2\mu_M(A-B)m(aA-bB)\big]\\ +2\big(\mu_M(A-B)-m(aA-bB)\big)(n_{\mathrm{med}}-n_L)-\mu_M\big(\mu_M(A-B)-m(aA-bB)\big)(1-2PS_L)\\ +\mu_M^2 PS_M - 2\mu_M(n_H-n_L)\Big].\end{aligned}$$

To ease notation in the above expression, write

$$D = \mu_M(A-B) - m(aA-bB),$$

as in (46). We get

$$\mathrm{E}\big[X^2(b,a)\big] = \frac{1}{PS_M^2}\Big[n_{2,H}-n_{2,L}-\mu_M^2 PS_M + D^2/2 + D\big(2(n_{\mathrm{med}}-n_L)+2\mu_M PS_L - \mu_M\big)\Big].$$

Now

$$\mathrm{E}\big(X(b,a)\big) = \frac{1}{2PS_M}\big(\mu_M(A-B)-m(aA-bB)\big) = \frac{D}{2PS_M}, \tag{68}$$

and so

$$\begin{aligned}&\text{Asy var}(\hat{\mu}_M) = \mathrm{Var}\big(X(b,a)\big) = \mathrm{E}\big[W^2(b,a)\big] - \big[\mathrm{E}\big(X(b,a)\big)\big]^2\\ &= \frac{1}{PS_M^2}\Big[n_{2,H}-n_{2,L}-\mu_M^2 PS_M + D^2/4 + D\big(2(n_{\mathrm{med}}-n_L)+2\mu_M PS_L-\mu_M\big)\Big],\end{aligned}$$

as stated in (45).

d: Covariance of $\widehat{PS}_M$ and $\widehat{IS}_M$:

Recall that what we need is the covariance of $U(a) - U(b)$ and $W(a) - W(b)$. With $C = A - B$, we have

$$U(a) - U(b) = \mathrm{I}(bm < Y \le am) - C\,\mathrm{I}(Y \le m) \quad \text{and}$$
$$W(a) - W(b) = \frac{1}{\mu^2}\big[Y\mu\,\mathrm{I}(bm < Y \le am) - Y(n_H - n_L) - \mu m(aA - bB)\,\mathrm{I}(Y \le m)\big].$$

from which we see that

$$\big(U(a) - U(b)\big)\big(W(a) - W(b)\big) = \frac{1}{\mu^2}\big[Y\mu\,\mathrm{I}(bm < Y \le am) - Y(n_H - n_L)\,\mathrm{I}(bm < Y \le am)$$
$$-\mu m(aA - bB)\,\mathrm{I}(bm < Y \le m) - \mu CY\,\mathrm{I}(bm < Y \le m) + CY(n_H - n_L)\,\mathrm{I}(Y \le m)$$
$$+\mu Cm(aA - bB)\,\mathrm{I}(Y \le m)\big]$$

Thus

$$\mathrm{E}\big(U(a) - U(b)\big) = PS_M - C/2 \quad \text{and}$$
$$\mathrm{E}\big(W(a) - W(b)\big) = -m(aA - bB)/(2\mu),$$

and

$$\mathrm{E}\big[\big(U(a) - U(b)\big)\big(W(a) - W(b)\big)\big] = IS_M(1 - IS_M) - \frac{1}{2\mu}m(aA - bB)(1 - 2PS_L)$$
$$+\frac{C}{\mu}\big[IS_M n_{\text{med}} - n_{\text{med}} + n_L + \tfrac{1}{2}m(aA - bB)\big].$$

Therefore

$$\mathrm{cov}(\widehat{PS}_M, \widehat{IS}_M) = \mathrm{E}\big[\big(U(a) - U(b)\big)\big(W(a) - W(b)\big)\big] - \mathrm{E}\big(U(a) - U(b)\big)\mathrm{E}\big(W(a) - W(b)\big)$$
$$= IS_M(1 - IS_M) - \frac{1}{2\mu}m(aA - bB)(PS_H - PS_L)$$
$$+\frac{C}{\mu}\big[IS_M n_{\text{med}} - n_{\text{med}} + n_L + \tfrac{1}{4}m(aA - bB)\big].$$

e: Covariances of estimates of income group means:

For the purposes of evaluating the reliability of income polarization estimates, $\hat{\mu}_H - \hat{\mu}_L$, $\hat{\mu}_H - \hat{\mu}_M$, and $\hat{\mu}_M - \hat{\mu}_L$, it is necessary to calculate the asymptotic covariances of the income group means. For the case of $\hat{\mu}_H - \hat{\mu}_L$, we use the result that

$$\text{Asy cov}(\hat{\mu}_H, \hat{\mu}_L) = \mathrm{E}\big[X(b)X(a)\big] - \mathrm{E}\big[X(b)\big]\mathrm{E}\big[X(a)\big].$$

By use of the same approach to evaluation of asymptotic variances for income group means as set out in section 2, one obtains

$$\text{Asy cov}(\hat{\mu}_H, \hat{\mu}_L) = \frac{1}{4\,PS_L \cdot PS_H}(\mu_L - bm)(am - \mu_H)AB. \tag{69}$$

Since $\mu_L < bm$ and $\mu_H > am$, it follows that this is strictly positive.
For the case of $\mu_M - \mu_L$, we have

$$\begin{aligned}\text{Asy cov}(\hat{\mu}_M, \hat{\mu}_L) &= \mathrm{E}\big[X(b)X(b,a)\big] - \mathrm{E}\big[X(b)\big]\mathrm{E}\big[X(b,a)\big] \\ &= \frac{1}{4\,PS_L \cdot PS_M}(\mu_L - bm)BC \\ &\quad + \frac{1}{2\,PS_L \cdot PS_M}(\mu_L - bm)B\big(\mu_{\text{med}} + PS_L(\mu_M - \mu_L)\big).\end{aligned} \tag{70}$$

For $\mu_H - \mu_M$, we have

$$\begin{aligned}\text{Asy cov}(\hat{\mu}_H, \hat{\mu}_M) &= \mathrm{E}\big[X(a)X(b,a)\big] - E\big[X(a)\big]\mathrm{E}\big[X(b,a)\big] \\ &= \frac{1}{4\,PS_H \cdot PS_M}(am - \mu_H)AC \\ &\quad + \frac{1}{PS_H \cdot PS_M}(am - \mu_H)A\big(PS_L(\mu_M - \mu_L) - (\mu_M - \mu_{\text{med}})/2\big).\end{aligned} \tag{71}$$

f: Compound measures:

Throughout this section, the results collected in the Table of Expectations will be freely used in the calculations.

Each of the compound measures in Section 3.2 involves the product of two terms, for instance

$$CP_L \equiv PS_L(\mu_M - \mu_L).$$

We see that

$$\begin{aligned}\text{Asy var}(\widehat{CP}_L) &= (\mu_M - \mu_L)^2\,\text{Asy var}(\widehat{PS}_L) \\ &\quad + PS_L^2\big(\text{Asy var}(\hat{\mu}_M) + \text{Asy var}(\hat{\mu}_L) - 2\,\text{Asy cov}(\hat{\mu}_M, \hat{\mu}_L)\big) \\ &\quad + 2\,CP_L\big(\text{Asy cov}(\widehat{PS}_L, \hat{\mu}_M) - \text{Asy cov}(\widehat{PS}_L, \hat{\mu}_L)\big).\end{aligned}$$

All of the asymptotic variances above are given in Section 2, and the covariance of $\hat{\mu}_M$ and $\hat{\mu}_L$ in equation (70). What remains is to compute the two asymptotic covariances with $\widehat{PS}_L$.

First we consider Asy cov$(\widehat{PS}_L, \hat{\mu}_L)$. It is equal to the covariance of $U(b)$ in (4) and $X(b)$ in (33):

$$\begin{aligned}\text{cov}\big(U(b),X(b)\big) &= \mathrm{E}\big(U(b)\,X(b)\big) - \mathrm{E}\big(U(b)\big)\,\mathrm{E}\big(X(b)\big) \\ &= \frac{1}{PS_L}\big[\mathrm{E}\big(U(b)\,U_1(b)\big) - \mu_L\mathrm{E}\big(U^2(b)\big)\big] + \frac{B}{2PS_L}(bm - \mu_L)(PS_L - B/2) \\ &= \frac{1}{PS_L}\big[n_L(1-B) - bmB\,PS_L + bmB^2/2 - \mu_L\,PS_L(1-2B) - \mu_L B^2/2\big] \\ &= \tfrac{1}{4}B(bm - \mu_L)\big(B/PS_L - 2\big).\end{aligned} \tag{72}$$

In similar fashion, the asymptotic covariance of $\widehat{PS}_L$ and $\hat{\mu}_M$ is the covariance of $U(b)$ and $X(b,a)$ in (44):

$$\operatorname{cov}\big(U(b), X(b,a)\big) = \mathrm{E}\big(U(b)\, X(b,a)\big) - \mathrm{E}\big(U(b)\big)\, \mathrm{E}\big(X(b,a)\big).$$

Here,

$$\mathrm{E}\big(U(b)\big)\, \mathrm{E}\big(X(b,a)\big) = \frac{D}{4\, PS_M}(2\, PS_L - B),$$

while

$$\begin{aligned}\mathrm{E}\big(U(b)X(b,a)\big) &= \frac{1}{PS_M}\mathrm{E}\big(U(b)U_1(a) - U(b)U_1(b) - \mu_M U(b)U(a) + \mu_M U^2(b)\big) \\ &= \frac{1}{PS_M}\big[B(n_L - n_{\text{med}} + \mu_M/2) + \tfrac{1}{2}D(2\, PS_L - B)\big];\end{aligned}$$

recall the definition (46) of D. Thus

$$\operatorname{cov}\big(U(b), X(b,a)\big) = \frac{1}{PS_M}\big[B(n_L - n_{\text{med}} + \mu_M/2) + \tfrac{1}{4}D(2\, PS_L - B)\big] \tag{73}$$

Consider next the case of $CP_H \equiv PS_H(\mu_H - \mu_M)$, for which we need the asymptotic covariances with $\widehat{PS}_H$ of $\hat{\mu}_H$ and $\hat{\mu}_M$. The first of these is

$$-\operatorname{Asy\ cov}(1 - \widehat{PS}_H, \hat{\mu}_H) = -\operatorname{cov}\big(U(a), X(a)\big),$$

which, after some algebra, becomes

$$\frac{1}{4}A(\mu_H - am)(2 - A/PS_H). \tag{74}$$

Similarly, the asymptotic covariance of $\widehat{PS}_H$ and $\hat{\mu}_M$ is $-\operatorname{cov}\big(U(a), X(b,a)\big)$, where

$$\operatorname{cov}\big(U(a), X(b,a)\big) = \frac{1}{PS_M}\Big[\tfrac{1}{4}D\,(2PS_H - A) - A(n_{\text{med}} - n_L) - A\, PS_L\mu_M)\Big]. \tag{75}$$

The last compound polarization measure defined in Section 3.2 is CP, which was defined as $(1 - PS_M)(\mu_H - \mu_L)$. For this, we need the covariances with $\widehat{PS}_M$ of $\hat{\mu}_H$ and $\hat{\mu}_L$. First,

$$\begin{aligned}\operatorname{Asy\ cov}(\widehat{PS}_M, \hat{\mu}_L) = \operatorname{cov}\big(U(a) - U(b), X(b)\big) &= \frac{1}{PS_L}\operatorname{cov}\big[U(a) - U(b), U_1(b) - \mu_L U(b)\big] \\ &= \frac{B}{4PS_L}(bm - \mu_L)\big(C + 2(PS_L - PS_H)\big).\end{aligned}$$

In addition,

$$\begin{aligned}\operatorname{Asy\ cov}(\widehat{PS}_M, \hat{\mu}_H) &= \operatorname{cov}\big(U(a) - U(b), X(a)\big) \\ &= \frac{1}{PS_H}\operatorname{cov}\big[U(a) - U(b), Y - U_1(a) - \mu_H\big(1 - U(a)\big)\big] \\ &= \frac{A}{4PS_H}(\mu_H - am)\big(C + 2(PS_L - PS_H)\big).\end{aligned} \tag{76}$$

Finally, consider the compound middle-class measure $CM = PS_M\,\mu_M$. Then

$$\text{Asy cov}(\widehat{CM}) = \mu_M^2\ \text{Asy var}(\widehat{PS}_M) + PS_M^2\ \text{Asy var}(\hat{\mu}_M) + 2\,CM\ \text{Asy cov}(\widehat{PS}_M, \hat{\mu}_M).$$

As before, Asy var$(\widehat{PS}_M)$ is given by (17), and Asy var$(\hat{\mu}_M)$ is given by (45). For the covariance,

$$\text{Asy cov}(\widehat{PS}_M, \hat{\mu}_M) = \text{cov}\big(U(a) - U(b), X(b,a)\big).$$

For this, cov$\big(U(a), X(b,a)\big)$ is given by (75) and cov$\big(U(b), X(b,a)\big)$ by (73).

Algorithm

Here is a detailed algorithm for the computation of estimates of the numerous measures presented in this paper and of their standard errors.

1. Select the cut-off parameters a and b needed to define the three income groups. (We used $b = 0.5$, $a = 2$.)
2. Choose a base unit of account. (Here it has been thousands of 2005 constant Canadian dollars.) Convert raw income measures in the sample to the chosen unit of account, and sort the converted data.
3. Compute the mean income $\hat{\mu}$, the mean squared income $\hat{\mu}_2$, and the variance $\hat{\sigma}^2$ of the sample.
4. Compute the sample median $\hat{m}$ and the two cut-off incomes, $b\hat{m}$ and $a\hat{m}$
5. By use of the approach described in Appendix 2, or otherwise, obtain the estimates $\hat{f}(b\hat{m})$ and $\hat{f}(a\hat{m})$ of the density at the cut-off incomes, and the estimates $\hat{A}$ and $\hat{B}$.
6. Count the number of data points with incomes in the three groups defined respectively by $[0, b\hat{m}]$, $(b\hat{m}, a\hat{m}]$, and $(a\hat{m}, \infty)$, and divide these numbers by the sample size N in order to obtain $\widehat{PS}_L$, $\widehat{PS}_M$, and $\widehat{PS}_H$.
7. Compute asymptotic standard errors for the estimated population shares using the formulas in section 2.4a.
8. Obtain estimates $\hat{n}_L$, $\hat{n}_{\text{med}}$, and $\hat{n}_H$ of the quantities $n(bm)$, $n(m)$, and $n(am)$ respectively. This can be done by averaging the incomes in the low-income group, incomes less than the median, and those in the low- and middle-income groups combined respectively. Also obtain estimates $\hat{n}_{2,L}$ and $\hat{n}_{2,H}$ by averaging squared incomes in the relevant groups.
9. Compute the estimated income shares: $\widehat{IS}_L = \hat{n}_L/\hat{\mu}$; $\widehat{IS}_M = (\hat{n}_H - \hat{n}_L)/\hat{\mu}$; $\widehat{IS}_H = 1 - \hat{n}_H/\hat{\mu}$.
10. Compute the estimated income group means: $\hat{\mu}_L = \hat{n}_L/\widehat{PS}_L$, $\hat{\mu}_M = (\hat{n}_H - \hat{n}_L)/\widehat{PS}_M$, $\hat{\mu}_H = (\hat{\mu} - \hat{n}_H)/\widehat{PS}_H$, and $\hat{\mu}_{\text{med}} = 2\hat{n}_{\text{med}}$. Also obtain $\hat{\mu}_{2,L} = \hat{n}_{2,L}/\widehat{PS}_L$, $\hat{\mu}_{2,M} = (\hat{n}_{2,H} - \hat{n}_{2,L})/\widehat{PS}_M$, $\hat{\mu}_{2,H} = (\hat{\mu}_2 - \hat{n}_{2,H}/\widehat{PS}_H$.
11. Compute the estimated relative mean income ratios using (48).

12. Obtain the estimated asymptotic variances for population shares, income shares, and group mean incomes by use of the formulas in section 2.4. For the relative mean income ratios, estimated asymptotic covariances are given by (54), (61), and (62).
13. Standard errors are found by dividing the asymptotic variances by the sample size N, and taking square roots.
14. The above computations provide all information necessary for the polarization measures introduced in section 3.2.

Table of Expectations

ref	random variable	expectation
	Y	μ
	Y^2	μ_2
(5)	$U(b)$	$PS_L - B/2$
(5)	$U(a)$	$1 - PS_H - A/2$
	$U_1(b)$	$n_L - bmB/2$
	$U_1(a)$	$n_H - amA/2$
(52)	$U(b)\,Y$	$n_L - Bn_{\text{med}}$
(59)	$U(a)\,Y$	$n_H - An_{\text{med}}$
(26)	$U_1(b)\,Y$	$n_{2,L} - bmBn_{\text{med}}$
(26)	$U_1(a)\,Y$	$n_{2,H} - amAn_{\text{med}}$
(11)	$U(a)U(b)$	$PS_L(1-A) - \frac{1}{2}B(1-A)$
	$U_1(a)U(b)$	$PS_L\,\mu_L + amAB/2 - amA\,PS_L - Bn_{\text{med}}$
	$U(a)U_1(b)$	$PS_L\,\mu_L(1-A) + bmAB/2 - bmB/2$
	$U_1(a)U_1(b)$	$PS_L\,\mu_{2,L} + m^2abAB/2 - bmB\mu_{\text{med}}/2 - amAPS_L\,\mu_L$
(7)	$U^2(b)$	$PS_L(1-2B) + B^2/2$
(7)	$U^2(a)$	$(1 - PS_H) + A^2/2 - A$
(22)	$U_1^2(b)$	$n_{2,L} + (bmB)^2/2 - 2bmBn_L$
	$U_1^2(a)$	$n_{2,H} + (amA)^2/2 - 2amAn_{\text{med}}$
(51)	$U(b)U_1(b)$	$n_L(1-B) - bmB\,PS_L + bmB^2/2$
(58)	$U(a)U_1(a)$	$n_H - A(am + 2n_{\text{med}})/2 + amA^2/2$
(25)	$W(b)$	$-bmB/(2\mu)$
	$W(a)$	$-amA/(2\mu)$
	$W^2(b)$	$\mu^{-2}[n_{2,L} + (bmB)^2/2 - 2bmB\,n_L] + \mu^{-4}n_L^2\mu_2 - 2\mu^{-3}n_L[n_{2,L} - bmB\,n_{\text{med}}]$
(63)	$X(b)$	$B(\mu_L - bm)/(2PS_L)$
(40)	$X(a)$	$-A(\mu_H - am)/(2PS_H)$
(68)	$X(b,a)$	$D/(2PS_M)$

Appendix 2: Density Estimation on the Positive Real Line

In most applications, the support of the distribution F is a subset of the positive real line. But it is known that in this case ordinary kernel density estimates are biased downwards. A possible way around this difficulty is to transform the data, by taking logarithms for instance, and getting kernel density estimates of the transformed data, which can then by multiplied by the Jacobian of the transformation to obtain estimates of the density of the positive data.

A better approach is suggested by Comte and Genon-Catalot (2012), where it is unnecessary to transform the data. Here is a brief description of their approach, roughly quoted from their paper. Instead of a Gaussian or Epanechnikov kernel defined for both positive and negative arguments, consider a density function $K(u)$ defined on the positive real line, with expectation equal to 1. Let $U_1, \ldots, U_n$ be an IID set of random variables with distribution characterised by the density K. Then the density of the mean $\bar{U} = (U_1 + \ldots + U_n)/n$ is given by $K_n(u) = nK^{*n}(nu)$, where K^{*n} is the n-fold convolution of K with itself. As $n \to \infty$, the distribution with density $K_n(u)$ converges to a point mass at 1. The proposal is to estimate the density $f(x)$ for $x > 0$ by

$$\hat{f}_n(x) = \frac{1}{Nx} \sum_{i=1}^{N} K_n(y_i/x) \tag{77}$$

using the random sample y_i, $i = 1, \ldots, N$. The motivation they give is as follows:

> In usual kernel methods, the intuition is that the estimation at x counts the number of observations X_k such that $X_k - x$ is close to 0. In our strategy, the intuition is that the estimator at x counts the number of observations X_k such that X_k/x is close to 1.

They also point out that $n^{-1/2}$ plays the same role here as does the bandwidth in conventional kernel methods.

The paper provides some examples of functions K for which the corresponding K_n can be computed analytically. The easiest of these has K equal to the density of the exponential distribution, which is also the gamma distribution with parameter unity: $K(u) = \mathrm{e}^{-u}$, from which it can be shown that

$$K_n(u) = \frac{1}{\Gamma(n)} \mathrm{e}^{-nu} n^n u^{n-1}.$$

With this choice, (77) becomes

$$\hat{f}_n(x) = \frac{n^n}{xN\Gamma(n)} \sum_{i=1}^{N} \exp(-ny_i/x)(y_i/x)^{n-1}.$$

Asymptotic theory requires that $n \to \infty$ as $N \to \infty$, but the guidelines as to how fast or how slowly in Comte and Genon-Catalot are very loose:

$$n = k^2 \ : \ \log(N) \le k \le N/\log(N).$$

In section 4 we discuss how we chose n for the datasets considered in the empirical work.

Appendix 3: Empirical Results

Table 1: Men in 2000

	$\hat{\xi}$	$\widehat{PS}$	$\widehat{IS}$	$\hat{\mu}$	$\widehat{\text{RMI}}$
L	17.7420	0.2702	0.0500	7.7588	0.1851
		(0.0007)	(0.0002)	(0.0271)	(0.0006)
M	35.4840	0.5811	0.5745	41.4371	0.9886
	(0.0770)	(0.0012)	(0.0019)	(0.0937)	(0.0018)
H	70.9681	0.1487	0.3755	105.8242	2.5248
		(0.0009)	(0.0019)	(0.3020)	(0.0045)

Sample size is 227828, and the estimate $\hat{A} = 0.8603$, and $\hat{B} = 0.4362$.

Table 2: Women in 2000

	$\hat{\xi}$	$\widehat{PS}$	$\widehat{IS}$	$\hat{\mu}$	$\widehat{\text{RMI}}$
L	11.1937	0.2925	0.0564	5.2879	0.1929
		(0.0008)	(0.0002)	(0.0200)	(0.0006)
M	22.3874	0.5296	0.5205	26.9463	0.9829
	(0.0640)	(0.0013)	(0.0022)	(0.0828)	(0.0022)
H	44.7748	0.1779	0.4231	65.2021	2.3783
		(0.0011)	(0.0022)	(0.1770)	(0.0038)

Sample size is 202491, $\hat{A} = 1.1229$, $\hat{B} = 0.6276$.

Table 3: Men in 2005

	$\hat{\xi}$	$\widehat{PS}$	$\widehat{IS}$	$\hat{\mu}$	$\widehat{\text{RMI}}$
L	17.5000	0.2742	0.0466	8.0874	0.1701
		(0.0007)	(0.0002)	(0.0267)	(0.0006)
M	35.0000	0.5538	0.4762	40.8862	0.8598
	(0.0828)	(0.0012)	(0.0021)	(0.1031)	(0.0027)
H	70.0000	0.1719	0.4772	131.9640	2.7752
		(0.0009)	(0.0022)	(0.7438)	(0.0088)

Sample size is 238356, $\hat{A} = 0.9659$, $\hat{B} = 0.5133$.

Table 4: Women in 2005

	$\hat{\xi}$	$\widehat{PS}$	$\widehat{IS}$	$\hat{\mu}$	$\widehat{\mathrm{RMI}}$
L	12.0000	0.3034	0.0605	5.9934	0.1993
		(0.0007)	(0.0002)	(0.0195)	(0.0006)
M	24.0000	0.5190	0.4867	28.2055	0.9378
	(0.0670)	(0.0012)	(0.0020)	(0.0822)	(0.0023)
H	48.0000	0.1775	0.4528	76.7076	2.5504
		(0.0010)	(0.0021)	(0.2834)	(0.0056)

Sample size is 218253, $\hat{A} = 1.0437$, $\hat{B} = 0.6432$.

Table 5: Differences Men-Women in 2000

	$\Delta\widehat{PS}$	$\Delta\widehat{IS}$	$\Delta\hat{\mu}$	$\Delta\widehat{\mathrm{RMI}}$
L	-0.0224	-0.0064	2.4709	-0.0078
	(0.0011)	(0.0003)	(0.0337)	(0.0008)
M	0.0516	0.0540	14.4908	0.0057
	(0.0018)	(0.0029)	(0.1251)	(0.0029)
H	-0.0292	-0.0476	40.6221	0.1465
	(0.0014)	(0.0029)	(0.3500)	(0.0059)

Table 6: Differences Men-Women in 2005

	$\Delta\widehat{PS}$	$\Delta\widehat{IS}$	$\Delta\hat{\mu}$	$\Delta\widehat{\mathrm{RMI}}$
L	-0.0292	-0.0138	2.0940	-0.0292
	(0.0010)	(0.0003)	(0.0331)	(0.0009)
M	0.0348	-0.0106	12.6807	-0.0780
	(0.0017)	(0.0029)	(0.1319)	(0.0035)
H	-0.0056	0.0244	55.2564	0.2248
	(0.0014)	(0.0030)	(0.7949)	(0.0104)

Table 7: Differences 2000–2005 for Men

	$\Delta\widehat{PS}$	$\Delta\widehat{IS}$	$\Delta\hat{\mu}$	$\Delta\widehat{\mathrm{RMI}}$
L	0.0041	-0.0034	0.3285	-0.0150
	(0.0010)	(0.0003)	(0.0380)	(0.0008)
M	-0.0273	-0.0983	-0.5509	-0.1288
	(0.0017)	(0.0028)	(0.1394)	(0.0032)
H	0.0232	0.1017	26.1398	0.2504
	(0.0013)	(0.0029)	(0.8027)	(0.0099)

Table 8: Differences 2000–2005 for Women

	$\Delta\widehat{PS}$	$\Delta\widehat{IS}$	$\Delta\hat{\mu}$	$\Delta\widehat{\text{RMI}}$
L	0.0109	0.0040	0.7054	0.0064
	(0.0011)	(0.0003)	(0.0279)	(0.0008)
M	-0.0105	-0.0338	1.2592	-0.0451
	(0.0018)	(0.0030)	(0.1167)	(0.0032)
H	-0.0004	0.0297	11.5055	0.1721
	(0.0015)	(0.0031)	(0.3316)	(0.0067)

Table 9: Measures of Polarization I

	$\widehat{PS}_L$	$\widehat{PS}_H$	$\widehat{PS}_L + \widehat{PS}_H$
Men, 2000	0.2702	0.1487	0.4189
	(0.0007)	(0.0009)	(0.0012)
Women, 2000	0.2925	0.1779	0.4704
	(0.0008)	(0.0011)	(0.0013)
Men, 2005	0.2742	0.1719	0.4462
	(0.0007)	(0.0009)	(0.0012)
Women, 2005	0.3034	0.1775	0.4810
	(0.0007)	(0.0010)	(0.0012)

Table 10: Measures of Polarization II

	$\hat{\mu}_H - \hat{\mu}_M$	$\hat{\mu}_M - \hat{\mu}_L$	$\hat{\mu}_H - \hat{\mu}_L$
Men, 2000	64.3871	33.6783	98.0654
	(0.2785)	(0.1080)	(0.2912)
Women, 2000	38.2558	21.6584	59.9142
	(0.1659)	(0.0930)	(0.1664)
Men, 2005	91.0779	32.7988	123.8767
	(0.7251)	(0.1169)	(0.7356)
Women, 2005	48.5021	22.2121	70.7142
	(0.2687)	(0.0925)	(0.2722)

Table 11: Compound Polarization Measures

	CP_L	CP_H	CP
Men, 2000	9.0985	9.5755	41.0773
	(0.0305)	(0.1294)	(0.0995)
Women, 2000	6.3360	6.8053	28.1855
	(0.0296)	(0.0784)	(0.0705)
Men, 2005	8.9947	15.6603	55.2714
	(0.0339)	(0.1704)	(0.1504)
Women, 2005	6.7398	8.6109	34.0111
	(0.0307)	(0.0961)	(0.0824)

Table 12: Changes in Polarization Measures 2000–2005

	ΔCP_L	ΔCP_H	ΔCP
Men	-0.1039	6.0848	14.1941
	(0.0456)	(0.2140)	(0.1804)
Women	0.4038	1.8056	5.8256
	(0.0426)	(0.1240)	(0.1084)

Appendix 4: Simulation Evidence

Simulations were run in order to see to what extent the numerous estimates produced by the 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 $n = 1001$ IID drawings from this distribution. As in the empirical work, the parameters a and b are set to 2.0 and 0.5 respectively. The true values of all the estimated properties are readily computed for the lognormal distribution.

For each of 100,000 replications, realisations were obtained for $\widehat{PS}_i$, $\widehat{IS}_i$, and $\hat{\mu}_i$, for $i = L, M, H$. The variances of these realisations were computed, and then multiplied by the sample size n, since the theoretical work concerns *asymptotic* variances. The estimates of the theoretical asymptotic variances, as given in the summary of results, were also computed for each replication, and then averaged over all of them. In some cases, a second estimate of an asymptotic variance was obtained for each replication as the sample variance of quantities like the $\hat{u}_i(b)$ defined in (6). These too are averaged over the replications. In Table A1 below, the averages of the point estimates are given, and in Table A2 the averages of the variance estimates.

	PS	IS	μ
Value for low incomes	0.2441	0.0452	0.3054
Estimated value	0.2439	0.0453	0.3052
Value for middle incomes	0.5118	0.3343	1.0768
Estimated value	0.5122	0.3352	1.0777
Value for high incomes	0.2441	0.6205	4.1910
Estimated value	0.2439	0.6195	4.1926

Table A1: point estimates

	PS	IS	μ
$\mathrm{Var}_L(\alpha)$	0.1472	0.0104	0.1551
$\mathrm{Var}_L(\beta)$	0.1467	0.0103	0.1548
$\mathrm{Var}_L(\gamma)$	0.1480	0.0104	0.1546
$\mathrm{Var}_L(\delta)$	0.1483	0.0105	0.1547
$\mathrm{Var}_M(\alpha)$	0.2499	0.4282	2.4579
$\mathrm{Var}_M(\beta)$	0.2512	0.4317	2.4434
$\mathrm{Var}_M(\gamma)$	0.2519	0.4359	2.4882
$\mathrm{Var}_M(\delta)$	0.2522	0.4327	2.4956
$\mathrm{Var}_H(\alpha)$	0.1472	0.4938	52.6442
$\mathrm{Var}_H(\beta)$	0.1475	0.4934	52.6775
$\mathrm{Var}_H(\gamma)$	0.1504	0.4971	53.1611
$\mathrm{Var}_H(\delta)$	0.1504	0.4976	53.2143

Table A2: estimates of asymptotic variances

The asymptotic variances denoted $\mathrm{Var}_i(\alpha)$ for $i = L, M, H$ are the theoretical variances as described in the summary of results with the true values computed for the lognormal distribution; those denoted $\mathrm{Var}_i(\beta)$ are the variances of the sets of point estimates from all the replications; those denoted $\mathrm{Var}_i(\gamma)$ are the estimates of the theoretical variances averaged over the replications; and those denoted $\mathrm{Var}_i(\delta)$ are the sample variances of quantities like the $\hat{u}_i(b)$ in (6), again averaged over the replications.

References

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.

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.

Bahadur, R. R. (1966). "A Note on Quantiles in Large Samples", *Annals of Mathematical Statistics*, 37, 577-80.

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

Beach, C. M. and R. Davidson (2024) "Income Share Standard Errors and a Quantile Toolbox of Distributional Statistics", Working Paper.

Blanchet, T., E. Saez, and G. Zucman (2022) "Real-time Inequality", Working Paper 30229, NBER.

Comte, F and V. Genon-Catalot (2012) "Density estimation for non negative random variables", *Journal of Statistical Planning and Inference*, 142, 1698-1715.

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

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.

Hoffman, F., D. S. Lee, and T. Lemieux (2020) "Growing Income Inequality in the United States and other Advanced Economies", *Journal of Economic Perspectives*, 34, 52-78.

Horowitz, J. L. (2001) *The Bootstrap*. In J.L. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol 5, 3159-3228. Elsevier Science, B.V, Amsterdam.5

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

Katz, L. F. and K. M. Murphy (1992) "Changes in relative wages, 1963–1987: supply and demand factors", *The Quarterly Journal of Economics*, 107, 35-78; https://doi.org/10.2307/2118323

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

Lin, P.-E., K.-T. Wu, and I. A. Ahmad (1980) "Asymptotic Joint Distributions of Sample Quantiles and Sample Mean with Applications", *Communications in Statistics - Theory and Methods* 9(1), 51–60.

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

Saez, E. and M. R. Veall (2007) "The Evolution of High Incomes in Canada, 1920-2000", Chapter 6 in *Top Incomes over the Twentieth Century*, Oxford, eds. A. B. Atkinson and T. Piketty.

Veall, M. R. (2012) "Top income shares in Canada: recent trends and policy implications", *Canadian Journal of Economics*, 45, 1247-1272.

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