Queen's Economics Department Working Paper No. 1505

Sample Sizes for Reliably Estimating Lower and Upper Income Shares in Income Distribution Analysis

Charles Beach

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

4-2023

Sample Sizes for Reliably Estimating Lower and Upper Income Shares in Income Distribution Analysis

By

Charles M. Beach
Department of Economics
Queen's University
Kingston, ON, Canada

April 2023

The author wishes to thank Roda Mendoza for technical assistance. All errors and limitations of this paper are solely the author's.

Abstract

This paper uses distribution-free formulas for the asymptotic variances of sample quantile income shares – as typically published by statistical agencies as measures of the distribution of income inequality – to calculate how large a survey sample must be in order to estimate a more refined quantile breakdown for a given level of confidence. The approach is applied to decile and quintile earnings data to calculate required increases in sample size to obtain tail 5 percent quantal share estimates and to test changes in income shares. Simple rules of thumb are offered for such a required increase.

1. Introduction

Since about 1980 to the early twenty-first century, income equality in many developed economics rose dramatically to historic levels (Guvenen et al., 2022). Media attention has focused on issues of "equitable growth" (Drummond, 2021), "fairness in growth" (Lohr, 2022), and "common prosperity" (The Economist, 2021a, b). In both Canada and the United States, federal governments have been focusing policies to better target low-income and middle-class families. Regional and provincial differences have required surveys to be large enough to identify significant differences in outcomes such as unemployment, education levels, and incomes. And the rapidly growing social, ethnic, and racial diversity in Canada prompts survey agencies to want to capture such outcomes for separate population groups as well (Bowlus et al., 2022; McKinney et al., 2022). One key outcome is inequality and the distribution of incomes across regions and population groups in the country.

One standard way of measuring income inequality is the share of total income (in the economy or for some population group) received by various quantile groups. For example, the share of the lowest-income quintile (or 20 percent) of income recipients is typically around 5-6 percent, while that of the top-income quintile is around 38-40 percent in Canada. These end-point (lower and upper) quantile income shares are of especial interest to inequality analysis and survey agencies. Typical breakdowns of income shares for official government statistical agencies such as Statistics Canada or the U.S. Bureau of the Census are by quintiles or deciles (eg., Statistics Canada, 2018). But for some occasions, even more refined breakdowns may be desirable (eg., Bell et al., 2022). Getting reliable estimates of income shares of smaller quantile groups, however, means having larger (and hence costlier) survey sizes. This paper examines how much survey samples must increase in order to obtain reliable estimates of lower and upper income shares as quantile breakdowns become more refined.

More specifically, this paper applies the statistical framework and techniques in Beach (2021a, b) to develop explicit formulas for (asymptotic) variances of quantile income shares. It then uses these formulas to solve for the survey sample sizes necessary to obtain a given degree of reliability (or confidence) for a given level of quantile breakdown (quintile, decile, vigintile). As it turns out, these income share (asymptotic) variances are distribution-free in the sense that they can be readily and consistently estimated from conventionally available statistics without having to know the specific underlying income distribution function itself. As a result, one can derive an explicit relationship between the required sample size and the degree of quantile refinement, and this relationship holds for all typical empirical distributions of income. In the process of this derivation, the paper corrects or adjusts for previous faulty specifications in the published statistics literature.

The paper thus makes three main contributions. First, it provides distribution-free formulas for the (asymptotic) variances of empirical income shares that can be used to evaluate the reliability or confidence of sample-based income share estimates. Second, the paper presents an approach to deriving the sample size required to estimate income shares to a given level of confidence. Third, it provides Rule of Thumb results to calculate how much larger a sample needs to be to estimate a smaller quantile end share and changes over time in income shares to obtain a given level of confidence.

The paper is organized as follows. The next section sets out the Quantile Function Approach used in this study as a basis for calculating (asymptotic) variances of income shares. Section 3 presents the basic analytical results for income shares. Section 4 shows how required sample sizes (for given quantile breakdown and confidence level) are derived from the analytical formulas of the previous section. Section 5 then presents empirical estimates and useful Rules of Thumb for calculating the required sample sizes. Some implications of these results for overall survey sample sizes are discussed in Section 6. Section 7 concludes.

2. Quantile Function Approach

To explain the Quantile Function approach (QFA) taken in this paper, consider first some formal concepts and notation. Suppose the distribution of income Y is divided into K ordered income groups, so that K=10 in the case of deciles and k=5 for quintiles. Let the dividing proportions of recipients be $p_1 < p_z < \ldots < p_{K-1}$ (with $p_o = 0$ and $p_K = 1.0$)[1]. Then in terms of the underlying (population) density of income recipients, the mean income of the i'th quantile is given by

$$\mu_i = \int_{\xi_{i-1}}^{\xi_i} y\, f(y)dy \,/\, \int_{\xi_{i-1}}^{\xi_i} f(y)dy \qquad \text{for i = 1, ..., K} \qquad (1)$$

where $f(\cdot)$ is the underlying (population) density function and the $\xi_{i's}$ are the cut off income levels corresponding to the proportions $p_1, p_2, \ldots, p_{K-1}$ (with $\xi_o = 0$ where incomes are assumed positive).[2] Similarly, the income share of the i'th income group can be expressed as

$$IS_i = \int_{\xi_{i-1}}^{\xi_i} \left(\frac{1}{\mu}\right) yf(y)dy \qquad \text{for i = 1, ..., K} \qquad (2)$$

and μ is the mean of the overall population distribution of income.

The integral expressions — what we'll refer to as quantile functions — link the quantile means μ_i and quantile income shares IS_i — to the quantile income cut-offs ξ_i, ξ_{i-1}, and the overall mean μ. A very broad theorem by C.R. Rao (1965) says that, if one knows the asymptotic distribution of the sample estimates of ξ_i, ξ_{i-1}, and μ as joint normal and if, in the population, functions of ξ_i, ξ_{i-1}, and μ are continuous and differentiable in these parameters, then sample estimates of these functions will also be asymptotically normally distributed with asymptotic means and variances (and covariances) that can be calculated in a straightforward fashion. We refer to this as Rao's linkage theorem. From both (1) and (2), it can be seen that one can use this theorem to thus establish the asymptotic distributions of sample estimates of both μ_i and IS_i.

In the case of quantile means, as a simple illustration of the usefulness of the quantile function approach, it has long been established that the sample cut-offs $\hat{\xi}_i$'s are indeed asymptotically normally distributed. More specifically, let $\hat{\xi} = (\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_{K-1})^1$ be a

[1] We assume in what follows that the data samples used are random samples. If the survey records are indeed weighted (as in the case of stratified samples, for example), the formulas can be readily adjusted by replacing sums of observations by sums of the sample weighted observations.

[2] If some incomes do take negative values (such as with capital gains losses in a year or net self-employment income that is negative in a year where illness has prevented the recipient working for a time), then simply define p_o to be the lowest income value in the sample.

vector of K-1 sample quantile cut-offs[3] from a random sample of size N drawn from a continuous population density $f(\cdot)$ such that the $\hat{\xi}_i$'s are uniquely defined and $f_i = f(\xi_i) > 0$ for all i = 1, ..., K-1. Then it can be shown (see, for example, Wilks, 1962, p. 273, or Kendall and Stuart, 1969, pp. 237-9) that the vector $\sqrt{N}$ ($\hat{\xi}$ – ξ) converges in distribution to a (K-1) – variate normal distribution with mean zero and variance – covariance matrix Λ where

$$\Lambda = \begin{bmatrix} \frac{p_1(1-p_1)}{{f_1}^2} & \cdots & \frac{p_1(1-p_{K-1})}{f_1 f_{K-1}} \\ \vdots & & \vdots \\ \frac{p_1\,(1-p_{K-1})}{f_1 f_{K-1}} & & \frac{p_{K-1}(1-p_{K-1})}{{f_{K-1}}^2} \end{bmatrix}.$$

Note how the (asymptotic) variances and covariances of the $\hat{\xi}_i$'s depend on the specific functional form of $f(\cdot)$ in the denominators.

Then applying a multivariate version of Rao's linkage theorem (Rao, 1965, p. 388), consider the full set of K sample quantile means $\hat{m} = \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K)^1$ corresponding to the vector of population quantile means $m = (\mu_1, \mu_2, \ldots, \mu_K)$ where $\hat{\mu}_i$ is defined in eq. (1). Then according to Rao's theorem for continuous differentiable functions, the vector $\hat{m}$ is asymptotically joint normally distributed in that $\sqrt{N}$ ($\hat{m}$ - m) converges in distribution to a joint normal with K x K (asymptotic) variances – covariance matrix V where

$$\text{Asy. var } (\hat{m}) = V = G\Lambda G^1 \qquad (3)$$

and the K x (K – 1) matrix G is

$$G = \begin{bmatrix} g_{11} & \cdots & g_{1,K-1} \\ \vdots & & \vdots \\ g_{K,1} & \cdots & g_{K,K-1} \end{bmatrix}$$

[3] To estimate the sample quantile cut-offs, order the sample of N observations by income level. Then, in the case of deciles, say, $\hat{\xi}_i$ is the income level such that $p_i N$ observations lie below it and the rest at or above. If there is no single observation meeting this condition, simply take the average of the two adjacent observations (below and above) that are closest.

$$= \left[\frac{\partial \mu_i}{\partial \xi_j}\right] \qquad \text{with i = 1, …, K rows and j = 1, …, K-1 columns.}$$

As a result,

$$Asy.var\,(\hat{\mu}_i) = \left(\frac{1}{D_{i-1}}\right)^2 p_{i-1}\,(1-p_{i-1})\xi\,{}_{i-1}{}^2 + \left(\frac{1}{D_i}\right)^2\, p_i(1-p_i)^2\xi_i{}^2 \qquad (4a)$$
$$-2\left(\frac{1}{D_{i-1}}\right)\left(\frac{1}{D_i}\right)p_{i-1}\,(1-p_i)\,\xi_{i-1}\,\xi_i$$

for i = 2, …, K-1, and

$$Asy.var\,(\hat{\mu}_1) = \left(\frac{1}{D_1}\right)^2 p_1(1-p_1)\xi_1{}^2 \qquad (4b)$$

$$Asy.var\,(\hat{\mu}_K) = \left(\frac{1}{D_K}\right)^2 p_{K-1}(1-p_{K-1})\xi_{K-1}{}^2\,, \qquad (4c)$$

where $D_i = p_i - p_{i-1}$.

Since the asymptotic variance (and covariance) formulas involve unknown population parameters, one obtains *estimated* (asymptotic) variances by replacing all the unknown parameters by their consistent estimates. So, for example,

$$Asy.\widehat{v}ar(\hat{\mu}_1) = \left(\frac{1}{D_1}\right)^2 p_1(1-p_1)\hat{\xi}_1{}^2 = \left(\frac{1-p_1}{p_1}\right)\hat{\xi}_1{}^2 \qquad (5)$$

where ξ_1 is replaced by its standard sample estimate and $D_1 = p_1$. Rao (1965, p. 355) has also shown that, if $f(\cdot)$ is strictly positive, then the $\hat{\xi}_i$'s are indeed (strongly) consistent. Standard errors are simply obtained from estimated (asymptotic) variances rescaled by the size of the estimation sample:

$$S.E.(\hat{\mu}_i) = \left[\frac{Asy.\widehat{v}ar(\hat{\mu}_i)}{N}\right]^{1/2} \qquad (6a)$$

where N is the sample size of the estimation sample. Testing for the statistical significance of $\hat{\mu}_i$ can then be undertaken by calculating the conventional "t-ratio"

$$t = \hat{\mu}_i \,/\, S.E.\,(\hat{\mu}_i) \qquad (6b)$$

and comparing it to appropriate critical values on the standard normal table (as an asymptotic test at a given level of significance or confidence).

In general, one would expect the (asymptotic) variances to depend on the specific functional form of the underlying income distribution density $f(\cdot)$. Certainly the (asymptotic) variance – covariance structure of the $\hat{\xi}_i$'s does. But — as seen in equations (4) — perhaps surprisingly, the resulting (asymptotic) variances and standard errors of the quantile means are a special case that do *not* depend on the specific functional form of $f(\cdot)$. In this sense, they are said to be distribution-free, and are very straightforward to calculate. Taking a quantile function approach thus allows one to avoid having to estimate assumed underlying population density functional forms (such as the lognormal in Beach, 2021a) or to undertake burdensome bootstrapping estimation techniques for density ordinate evaluation (as in Davidson, 2018).

3. Basic Results for Income Shares

In the case of income shares, similar reasoning goes through, though the formulas are a bit more complicated. Since in this paper, we are specifically interested in the lower and upper tail income shares, for convenience divide the distribution into only three regions – the lower p proportion of recipients, the upper (1 – q) proportion of recipients, and the rest (or middle q – p proportion of recipients). Let the lower cut-off quantile (corresponding to p) be ξ_L, and let the upper cut-off (corresponding to q) be ξ_U.

Now the income shares in eq.(2) are functions of three parameters: the income cut-offs ξ_L and ξ_U, as well as the mean, μ, of the income distribution. To take account of this, Lin, Wu, and Ahmad (1979, 1980) — hereafter LWA — established that, under general regularity conditions, $\hat{\xi}_L$, $\hat{\xi}_U$ and $\hat{\mu}$ are asymptotically joint normally distributed with (asymptotic) variance-covariance matrix

$$\Sigma = [\sigma_{ij}]$$

$$\text{where } \sigma_{11} = \frac{p(1-p)}{[f(\xi_L)]^2}, \quad \sigma_{22} = \frac{q(1-q)}{[f(\xi_U)]^2}, \quad \sigma_{33} = \sigma^2 \qquad (7)$$

$$\sigma_{12} = \frac{\rho(1-q)}{f(\xi_L)f(\xi_U)} = \sigma_{21}$$

$$\sigma_{13} = \frac{x_p - \mu(1-P)}{f(\xi_L)}$$

and $$\sigma_{23} = \left[\frac{x_q - \mu(1-q)}{f(\xi_U)}\right]$$

where σ^2 is the variance of the overall (population) distribution of income, and

$$x_P = \int_{\xi_L}^{\infty} yf(y)dy \qquad x_q = \int_{\xi_U}^{\infty} yf(y)dy.$$

(Note that the published journal version of LWA has some serious typos in the statement of their theorem 2.1, but the earlier discussion paper version presents the theorem correctly. To keep the correct version clear, I have adopted slightly different notation from LWA.) The last two terms, however, can be stated more conveniently. Note that

$$\begin{aligned} \mu &= \int_{o}^{\infty} yf(y)dy \\ &= \int_{0}^{\xi_L} yf(y)dy + x_p \\ &= p\ \mu_L + x_p \end{aligned}$$

$$\text{or } x_p = \mu - p\mu_L \tag{8a}$$

where μ_L is the lower quantile group mean income. Hence $\sigma_{13} = p\ (\mu - \mu_L) / f(\xi_L)$ which, interestingly, is strictly positive. Similarly, $x_q = (1 - q)\ \ \mu_U$ where μ_U is the upper quantile group's mean. Therefore,

$$\sigma_{23} = (1 - q)\ (\mu_U - \mu) / f(\xi_U) \tag{8b}$$

which, again, is also strictly positive.

One can now combine these results with Rao's linkage theorem. So, if $\hat{\xi}_L$, $\hat{\xi}_U$ and $\hat{\mu}$ are asymptotically joint normal with (asymptotic) variance – covariance matrix Σ above — as given in (7) and (8) — then the (asymptotic) variance of $I\hat{S}_i$ (for i = L, M, U) is given by

$$\text{Asy. var}\,(I\hat{S}_i) = \text{G}^1\,\Sigma\,\text{G} \tag{9}$$

where

$$\text{G} = \left[\frac{\partial IS_i}{\partial \xi_L}\,,\ \frac{\partial IS_i}{\partial \xi_U}\,,\ \frac{\partial IS_i}{\partial \mu}\right]^1 \qquad = [g1, g2, g3]^1.$$

So, in the case of the lower income share, (i = L):

$$g_1 = \left(\frac{1}{\mu}\right) \xi_L\, f\,(\xi_L)$$

$$g_2 = 0$$

$$g_3 = \frac{-IS_L}{\mu}$$

and

$$\text{Asy. var}\,(I\hat{S}_L) = g_1^2\,\sigma_{11} + g_3^2\sigma_{33} + 2g_1g_3\sigma_{13}$$

$$= p\,(1\text{-}p)\left(\frac{\xi_L}{\mu}\right)^2 + (IS_L)^2\left(\frac{\sigma}{\mu}\right)^2 \tag{10}$$

$$-2\left(\frac{\xi_L}{\mu}\right)(IS_L)[p - IS_L]$$

where use has been made that $IS_L = p\,(\mu_L / \mu)$ from (1) and (2).

In the case of the upper income share (i=U):

$$g_1 = 0$$

$$g_2 = -\left(\frac{1}{\mu}\right) \xi_u\, f\,(\xi_u)$$

$$g_3 = -\frac{IS_U}{\mu}$$

and

$$\text{Asy. var}\,(I\hat{S}_U) = q\,(1\text{-}q)\left(\frac{\xi_U}{\mu}\right)^2 + (IS_U)^2\left(\frac{\sigma}{\mu}\right)^2 \tag{11}$$

$$+2\left(\frac{\xi_U}{\mu}\right)(IS_U)[\,IS_U - (1 - q)]$$

where use has been made that $IS_U = (1 - q)(\mu_U/\mu)$.

For completeness, one can also derive the asymptotic variance for the middle income share (i = M). In this case,

$g_1 = -\left(\frac{1}{\mu}\right) \xi_L f(\xi_L)$

$g_2 = \left(\frac{1}{\mu}\right) \xi_U f(\xi_U)$

$g_3 = -\frac{IS_M}{\mu}$.

Consequently,

$$\begin{aligned} \text{Asy. var } (\hat{IS}_M) &= g_1^2\sigma_{11} + g_2^2\sigma_{22} + g_3^2\sigma_{33} \\ &\quad +2g_1g_2\sigma_{12} \\ &\quad +2g_1g_3\sigma_{13} \qquad +2g_2g_3\sigma_{23} \\ = p(1-p)\left(\frac{\xi_L}{\mu}\right)^2 &+ q(1-q)\left(\frac{\xi_U}{\mu}\right)^2 + (IS_M)^2\left(\frac{\sigma}{\mu}\right)^2 \\ &-2\left(\frac{\xi_L}{\mu}\right)\left(\frac{\xi_U}{\mu}\right) p(1-q) \\ &+2\left(\frac{\xi_L}{\mu}\right)(IS_M)[p - IS_L] \\ &-2\left(\frac{\xi_U}{\mu}\right)(IS_M)[IS_U - (1-q)]. \end{aligned} \qquad (12)$$

The standard error of the i'th quantile income share is thus given by

$$\text{S.E. } (\hat{IS}_i) = \left[\frac{Asy.\widehat{v}ar(\hat{IS}_i)}{N}\right]^{1/2}.$$

Once again, the asymptotic variances and standard error formulas for income shares are also distribution-free, so all components in these expressions are known or can be estimated consistently, and hence conventional statistical inference can be undertaken in straightforward fashion.

4. Sample Sizes for Tail Income Shares: Analytical Results

As indicated in (6b), such a conventional test can be done by comparing a calculated "t-ratio"

$$t = I\hat{S}_i \ / S.E\left(I\hat{S}_i\right)$$

to a specified critical value on the standard normal distribution. This would be a test of whether the estimated income share is indeed significantly different from zero; ie., whether the income share is estimated reliably at a high degree of confidence. (For convenience of discussion, we'll work with a two-tailed test.) For a 95 percent level of confidence, the critical value is $t_{.95}$ =I.960, and for 99 percent level of confidence, the critical value is $t_{.99}$ = 2.576. This is equivalent to comparing

$$t^2 = \frac{(I\hat{S}_i)^2}{Asy.\widehat{var}(I\hat{S}_i)/N} \tag{13}$$

to critical values $(1.960)^2 = 3.842$ and $(2.576)^2 = 6.636$, respectively.

But one can turn the question around and instead ask what sample size would be required to attain a specified level of confidence for such a standard normal "t-ratio" test on an income share with given sample estimates of $I\hat{S}_i$ and $Asy\,\hat{v}ar\left(I\hat{S}_i\right)$. This can be answered by inverting the equation

$$(t_{crit})^2 = \frac{(I\hat{S}_i)^2}{Asy.\hat{v}ar(I\hat{S}_i)/N}$$

to get

$$\hat{N} = (t_{crit})^2 \cdot \left[\frac{Asy\,\hat{v}ar(I\hat{S}_i)}{(I\hat{S}_i)^2}\right] \tag{14}$$

where t_{crit} is the critical value on the standard normal distribution, and $\hat{N}$ is the required sample size to estimate IS_i to at least that level confidence given by t_{crit}.

One can indeed push this question further. $\hat{N}$ is a function of how wide an income share is. Applied to the lower and upper end quantiles, a wide share such as the bottom or top quintile income share is likely to be estimated more reliably for a given sample size than a top or bottom decile share or vigintile share. Thus one can use the relationship in (14) to ask how much longer $\widehat{N}$ would have to be to estimate IS_L or IS_U when p is .10, say, or .05 (correspondingly, q is .90 or .95). That is the question this paper addresses.

Given we know the exact formulas linking IS_L and $Asy.var\left(I\hat{S}_L\right)$ to p, the width of the lower income quantile, a natural approach to addressing the above question is to calculate the derivative $\partial\, \widehat{N}_L/\partial p$ in the case of the lower share. (For notational convenience, we work out the derivatives deleting the cap on the income share and Asy.var terms in (14).

$$\frac{\partial \widehat{N}_L}{\partial p} = (t_{crit})^2 \left\{ (IS_L)^{-2} - \frac{\partial[Asy.var(IS_L)]}{\partial p} \right.$$

$$\left. + Asy.var\left(I\hat{S}_L\right) \cdot \frac{\partial\left({IS_L}^{-2}\right)}{\partial p} \right\}$$

$$= (t_{crit})^2 \left\{ \left(\frac{1}{IS_L}\right)^2 \cdot \frac{\partial\left[Asy.var\left(I\hat{S}_L\right)\right]}{\partial p} \right.$$

$$\left. - \left(\frac{2}{IS_L}\right)\left(\frac{\widehat{N}_L}{(t_{crit})^2}\right) \cdot \frac{\partial IS_L}{\partial p} \right\}$$

where it can be shown that

$$\frac{\partial IS_L}{\partial p} = \frac{\xi_L}{\mu}. \text{ Therefore,} \tag{15}$$

$$\frac{\partial \widehat{N}_L}{\partial p} = \left(\frac{t_{crit}}{IS_L}\right)^2 \cdot \frac{\partial[Asy.var\ (I\hat{S}_L)]}{\partial p} \tag{16}$$

$$- \left(\frac{2N_L}{IS_L}\right)\left(\frac{\xi_L}{\mu}\right).$$

Now from eq (10) above,

$$Asy.var\left(I\hat{S}_L\right) = p(1-p)\left(\frac{\xi_L}{\mu}\right)^2 + (IS_L)^2\left(\frac{\sigma}{\mu}\right)^2$$

$$-2\left(\frac{\xi_L}{\mu}\right)(IS_L)[p - IS_L]$$

and it can be shown that

$$\frac{d\xi_L}{dp} = \frac{1}{f(\xi_L)}\,. \tag{17}$$

Asy. var $\left(I\hat{S}_L\right)$ above has three terms; for convenience, call them Term 1, Term 2, and Term 3. Then

$$\frac{\partial[Asy.var\ (I\hat{S}_L)]}{\partial p} = \frac{\partial[Term\ 1]}{\partial p} + \frac{\partial[Term\ 2]}{\partial p} + \frac{\partial[Term\ 3]}{\partial p}$$

$$\frac{\partial[Term\ 1]}{\partial p} = \left(\frac{1}{\mu}\right)^2 \cdot \frac{\partial\left[p\ (1-p)\xi_L{}^2\right]}{\partial p}$$

$$= \left(\frac{\xi_L}{\mu}\right)^2 \left[\, 2p(1-p)\left(\frac{1}{\xi_L f(\xi_L)}\right) + (1-2p)\right]$$

which is positive if p < .5 — which is indeed the situation for the lower income share.

$$\frac{\partial[Term\ 2]}{\partial p} = \left(\frac{\sigma}{\mu}\right)^2 \cdot \frac{\partial\left(IS_L{}^2\right)}{\partial p}$$

$$= 2\left(\frac{\sigma}{\mu}\right)^2\ (IS_L)\left(\frac{\xi_L}{\mu}\right)$$

which is clearly positive. Finally,

$$\frac{\partial[Term\ 3]}{\partial p} = \left(\frac{2}{\mu}\right) \cdot \frac{\partial[\xi_L(IS_L)(p - IS_L)]}{\partial p}$$

$$= \left(\frac{2}{\mu}\right) \left\{\xi_L(IS_L)\ \left[1 - \frac{\xi_L}{\mu}\right] + \ (p - IS_L)\ \left[\xi_L\ \left(\frac{\xi_L}{\mu}\right)\ +\ \frac{IS_L}{f(\xi_L)}\right]\right]\right\}$$

which is again positive. So it turns out

$$\frac{\partial[Asy.var\ (I\hat{S}_L)]}{\partial p} > 0.$$

This going back to eq (16),

$$\frac{\partial \widehat{N}_i}{\partial p} = \left(\frac{t_{crit}}{IS_L}\right)^2 \cdot \frac{\partial[Asy.var\ (I\hat{S}_L)]}{\partial p} \quad - \quad \left(\frac{2N_L}{IS_L}\right)\left(\frac{\xi_L}{\mu}\right), \tag{18}$$

has a positive first term minus another positive term. Now intuitively, one would perhaps expect that estimating a wider income share — for a given level of confidence — could be done with a smaller overall sample size; ie., the above derivative would be negative. So long as the second term dominates the first, this will indeed be so. This will occur when

$$\frac{\partial[Asy.var\ (I\hat{S}_L)]}{\partial p} < \frac{2\widehat{N}_L(IS_L)}{(t_{crit})^2}\left(\frac{\xi_L}{\mu}\right). \tag{19}$$

Since $Asy.var\ (I\hat{S}_L)$ is essentially independent of sample size, the right-hand term in (19) is almost certainly likely to dominate for typically large cross-sectional sample surveys and national censuses. Nonetheless, the actual size of $\partial\widehat{N}_L/\partial p$ depends upon the actual shape of the income distributions over its lower range — since $\partial\left[Asy.var\ \left(I\hat{S}_L\right)\right]/\partial p$ involves $f(\xi_L)$ — and the distribution's actual values of all the components in eq.(18). So to proceed further on how much sample sizes need to increase in order to estimate narrower income shares (ie., for smaller values of p) to a given level of confidence or reliability, one must turn to empirical estimation and evaluation of the parameters and expressions involved.

For the upper income share, the derivation is similar, but the conclusion is somewhat different. From eq. (11) above,

$$Asy.var\ \left(I\hat{S}_U\right) = q(1-q)\left(\frac{\xi_U}{\mu}\right)^2 + (IS_U)^2\left(\frac{\sigma}{\mu}\right)^2$$

$$+2\left(\frac{\xi_U}{\mu}\right)(IS_U)\ [IS_U - (1-q)].$$

Note also that

$$\frac{d\xi_U}{dq} = \frac{1}{f(\xi_U)} \tag{20}$$

and $\frac{\partial IS_U}{\partial q} = \frac{-\xi_U}{\mu}$. (21)

Again, one can divide the right-hand side of eq.(11) into three terms.

$$\frac{\partial[Term\ 1]}{\partial q} = \left(\frac{1}{\mu}\right)^2 \cdot \frac{\partial[q\ (1-q)\xi_U{}^2]}{\partial q}$$

$$= \left(\frac{\xi_U}{\mu}\right)^2 \left[2q(1-q)\left(\frac{1}{\xi_U f(\xi_U)}\right) + \ (1-2q)\right]$$

$$\frac{\partial[Term\ 2]}{\partial q} = \left(\frac{\sigma}{\mu}\right)^2 \cdot \frac{\partial\left(IS_U{}^2\right)}{\partial q}$$

$$= \ -2\ \left(\frac{\sigma}{\mu}\right)^2 (IS_U)\left(\frac{\xi_U}{\mu}\right)$$

$$\frac{\partial[Term\ 3]}{\partial q} = \frac{2}{\mu} \cdot \frac{\partial[\xi_U(IS_U)(IS_U-(1-q))]}{\partial q}$$

$$= \left(\frac{2}{\mu}\right)\left\{\ \xi_U(IS_U)\left[1-\frac{\xi_U}{\mu}\right] + \ [IS_U-(1-q)]\left[\frac{IS_U}{f(\xi_U)} - \ \xi_U\left(\frac{\xi_U}{\mu}\right)\right]\right\}.$$

The derivative of the second term is clearly negative. But, since q > 0.5 in our analysis, the derivatives of the first and third terms can't be signed a priori, so neither can the derivative of $Asy.var\ (I\hat{S}_U)$ as a whole. As q gets larger, the (upper) income share gets smaller. But smaller scale random variables are typically associated with smaller variances (a scale effect). But also as q gets larger, the upper (1-q) income share becomes less reliably estimated for a given sample size as fewer data points fall into its range (a reliability effect). In the case of increasing q, these two effects lead to an inconclusive sign on the derivative of $Asy.var\ (I\hat{S}_U)$.

The effect of an increase in q on the required sample size then is:

$$\frac{\partial\widehat{N}_U}{\partial q} = \ (t_{crit})^2\ \left\{(IS_U)^{-2}\ \cdot \frac{\partial[Asy.var(I\hat{S}_U)]}{\partial q}\right.$$

$$\left. + \ Asy.var\ \left(I\hat{S}_U\right)\ \cdot \frac{\partial\left(IS_U{}^{-2}\right)}{\partial q}\right\}$$

$$= \ (t_{crit})^2\ \left\{\left(\frac{1}{IS_U}\right)^2 \cdot \frac{\partial[Asy.var\ (I\hat{S}_U)]}{\partial q}\right. \qquad (22)$$

$$+ \left(\frac{2}{IS_U}\right)\left(\frac{\hat{N}_U}{(t_{crit})^2}\right)\left(\frac{\xi_U}{\mu}\right)\Big\}.$$

Since the second term in (22) is clearly positive, this works to attenuate the inconclusive sign effect of the first term. So again, the actual outcome for evaluating (22) requires that we turn to empirics to evaluate the size of the effect of larger q (ie., smaller 1-q) on required sample size.

5. Sample Sizes for Tail Income Shares: Empirical Estimates

Empirical estimates are based on parameter estimates in appendix Table A taken from Beach (2021b) calculated from the May 2015 Canadian Labour Force Survey (LFS). The income measure used is usual weekly earnings, and results are broken down separately for men and women in the Canadian labour market. So, for example, the earnings share of the lowest ten percent for men is 1.97 percent and for women 1.82 percent. The shares of the top ten percent are 22.7 and 23.5 percent, respectively.

Using the formulas in equations (10), (11), and (14) above, one can directly estimate the required sample sizes for lower and upper earnings shares for different values of p and q and for 95 and 99 percent levels of confidence. These results are presented in the first and third columns of Tables 1 (for the lower shares) and 2 (for upper shares). So, for example, to estimate the lowest decile earnings share for women with at least a 99 percent level of confidence requires at least 130 sample observations to be used in the calculation. Rows in both tables are organized from wider share intervals to narrower more refined intervals (ie., p goes from 0.40 to 0.20 to 0.10 in the case of lower shares, and (1-q) goes from 0.40 to 0.20 to 0.10 for upper shares).

In both Tables 1 and 2, required sample sizes are seen to substantially rise as one moves from wider to more refined or detailed earnings intervals, and to be larger for higher levels of confidence — exactly as one would expect. But interestingly, the required sample sizes (which depend on underlying parameter estimates) are not always quite the same between male and female earners, so a gender breakdown in empirical analysis makes sense. And required sample sizes for upper earnings shares (in Table 2) are not as large as for lower earnings shares (in Table 1) — not surprising since the tests refer to differences from zero. So, if one is especially interested in analyzing inequality at the lower end of the distribution, considerably larger sample sizes are required for a given level of confidence or reliability of inference.

For completeness, similar sets of calculations and results are provided for the middle-income group (M) in Appendix B. One could also examine empirically how the $Asy.var\left(I\hat{S}_i\right)$'s vary with sizes of p and q. This is provided in appendix Table A2 in the first and third columns. For both lower and upper earnings shares, the asymptotic variances (AVs) indeed decrease as the shares become narrower. However, if one calculates (asymptotic) coefficients of variation (ACV) of the earnings shares (ie., $(Asy.\widehat{v}ar)^{1/2}\ /\ I\hat{S}_i\big)$, these can be seen in columns two and four to increase as shares become more refined. The ACV's, interestingly, turn out to be virtually the same for male and female earners.

Table 1

Estimates of $\widehat{N}_L$ for Lower Income Shares

Canada, 2015

	Males		Females	
	$\widehat{N}_L$	%Δ	$\widehat{N}_L$	%Δ
a) 95% Level of Confidence				
p = .40	9.26		9.83	
		2.429		2.708
p = .20	31.75		36.46	
		1.980		1.065
p = .10	94.61		75.30	
b) 99% Level of Confidence				
p = .40	16.00		16.99	
		2.427		2.708
p = .20	54.83		62.98	
		1.980		1.065
p = .10	163.41		130.07	

Source: Author’s calculations based on parameter estimates in Table A1 in the Data Appendix.

Table 2

Estimates of $\widehat{N}_U$ for Upper Income Shares

Canada, 2015

	Males		Females	
	$\widehat{N}_U$	%Δ	$\widehat{N}_U$	%Δ
a) 95% Level of Confidence				
(1-q) = .40	6.89		6.91	
		1.226		1.206
(1-q) = .20	15.34		15.25	
		0.985		1.121
(1-q) = .10	30.45		32.34	
b) 99% Level of Confidence				
(1-q) = .40	11.90		11.94	
		1.226		1.206
(1-q) = .20	26.49		26.34	
		0.985		1.121
(1-q) = .10	52.60		55.87	

Source: Author's calculations based on parameter estimates in Table A1 in the Data Appendix.

The strength or degree to which required sample sizes need to increase for more detailed end-point shares can also be examined. This could be represented by the proportional change in required sample size as one moves from wider share intervals to narrower more refined intervals. This is listed in columns two and four of Tables 1 and 2. As can be seen, these are indeed very large — over 100 percent — in almost all cases. Interestingly, they are generally much lower for upper earnings shares than for lower such shares. But their actual size depends on how large the change is in p or (1-q).

A natural way to adjust for different sizes of Δp or Δ(1-q) is simply to divide the percentage change in required sample size by the given percentage change in p (or 1-q) — what economists call the elasticity of $\widehat{N}$ with respect to p (or 1-q). The results are shown in Table 3.[4] As can be seen, the elasticities all decline from wider to narrower tail intervals. The elasticities are also generally much higher for lower than for upper tail earnings shares. So one needs to consider each of the four cases (lower and upper shares, male and female earners) separately.

The above elasticities provide direct evidence on how required sample sizes need to increase as one moves, say, from quintile to decile earnings shares (for a given level of confidence). But what if one wants to estimate how much sample sizes must increase if one wished to estimate vigintile earnings shares (i.e., p = .05, 1-q = .05) to the same degree of confidence? Prediction outside of an estimation range is always "iffy". But if one looks at the results in Table 3, it seems not unreasonable to come up with Rules of Thumb elasticity values that could be used. For upper earnings shares, the estimated elasticity values are sufficiently close for men as women that a single Rule of Thumb value of, say, 2.0 could be used for both. For lower earnings shares, however, different elasticity values for men and women would seem to be more appropriate – say, a value of 2.0 for women and 4.0 for men. These Rule of Thumb elasticity values are likely a bit on the conservative side for good measure.

[4] Since the $(t_{crit})^2$ term in eq. (14) falls out in the elasticity calculations — note the identical values between upper and lower panels in columns two and four of Tables 1 and 2 – the elasticity figures hold for all confidence levels.

Table 3

Elasticities of $\widehat{N}_L$ and $\widehat{N}_U$ wrt Income Share Width

Canada, 2015

	Males	Females
a) For Δp *and* $\widehat{N}_L$		
.40 → .20	4.857	5.414
.20 → .10	3.961	2.150
b) For $\Delta(1-q)$ *and* $\widehat{N}_U$		
.40 → .20	2.455	2.411
.20 → .10	1.971	2.241

Source: Author's calculations from results in Tables 1 and 2.

To see how these Rule of Thumb elasticity values (e) can be used to predict required sample sizes if one wishes to estimate tail vigintile income shares to a given degree of reliability, consider the following calculation. In the case of lower income shares (where we omit superscript "hats" on N for notational convenience):

$$\frac{\Delta N_L}{N_L} = e_L \cdot \left(\frac{\Delta p}{p}\right)$$

If one goes from decile to vigintile lower income shares,

$$\left(\frac{N_{.05} - N_{.10}}{N_{.10}}\right) = e_L \cdot \left(\frac{.10 - .05}{.10}\right)$$

or $$N_{.05} = \left[1 + e_L\left(\tfrac{1}{2}\right)\right] \cdot N_{.10} \qquad (23)$$

So, if $e_L = 4.0$ for men, $N_{.05} = 3N_{.10}$
and if $e_L = 2.0$ for women, $N_{.05} = 2N_{.10}$,
where $N_{.10}$ is the required sample size for decile estimation (for a given level of confidence).
Similarly for upper income shares:

$$\frac{\Delta N_U}{N_U} = e_U \cdot \left(\frac{\Delta(1-q)}{(1-q)}\right)$$

ie: $$\frac{N_{.95} - N_{.90}}{N_{.90}} = e_U \cdot \left(\frac{.95 - .90}{(1 - .90)}\right)$$

or $$N_{.95} = \left[1 + e_U\left(\tfrac{1}{2}\right)\right] \cdot N_{.90}. \qquad (24)$$

So, if $e_U = 2.0$ for both men and women in the labour market, $N_{.95} = 2N_{.90}$, where $N_{.90}$ is the required sample size for estimating the top decile income share at a given level of confidence.

As a useful Rule of Thumb, then, if one wishes to estimate vigintile income shares with the same degree of reliability as for decile shares, one needs a sample about three times as large in the case of the lowest five percent share for men, and about twice as large in the rest of the cases (lowest five percent for women and top five percent share for both men and women).

6. Sample Sizes for Changes in Tail Income Shares

A more demanding challenge for distributional inference is whether it can identify statistically significant differences or changes in income shares, say over time. Suppose, for example, one has income share estimates for two years, year 0 and year 1. Then let

$$\widehat{\Delta} \equiv I\hat{S}_i\,(1) - I\hat{S}_i(0)$$

for income group i=L, M, U. One can test (asymptotically) for the statistical significances of $\widehat{\Delta}$, again using a standard normal test "t-ratio" statistic.

$$t - ratio = \hat{\Delta}/SE\left(\hat{\Delta}\right)$$

where

$$SE\left(\hat{\Delta}\right) = \left[\left(SE\left(I\hat{S}_i(0)\right)\right)^2 + \left(SE\left(I\hat{S}_i(1)\right)\right)^2\right]^{1/2}$$

$$= \left[\frac{Asy\hat{.}\,var\left(\widehat{IS}_i(0)\right)}{N(0)} + \frac{Asy.var\left(\widehat{IS}_i(1)\right)}{N(1)}\right]^{1/2}$$

where N(0) and N(1) are the sample sizes used to estimate the shares for the two years.

To illustrate the test procedure, consider years 2015 (year 1) and 2000 (year 0). Background parameter estimates for 2000 are provided in appendix Tables A3 and A4, and complementary $\widehat{N}$ estimates for the 2000 income shares appear in Appendix C. They show generally similar results to those for 2015.

Test results for comparing lower and upper income shares between 2000 and 2015 are presented in Table 4. As can be seen, lower earnings shares have declined and upper shares have risen for both females and males over this period. But in the case of lower shares, three of the four estimated declines were not statistically significant (at conventional significance levels) — only that for the bottom 20 percent of male workers was. In the case of upper earnings shares, however, at least three of the four shares shows a highly significant increase. That is, increases in earnings shares in the upper portion of the Canadian earnings distribution were highly statistically significant, while the losses in lower income shares were largely not significant (with current LFS sample sizes). Appendix Table B3 also shows that the middle earnings shares experienced notable earnings losses that were also highly statistically significant. Evidently, the big shift in the earnings distribution that occurred over the 2000-2015 period was middle earners losing out to upper earners who were the big winners.

But what can be said about sample sizes that would be needed to generally establish statistical significance of income share changes? We will use the (asymptotic) variance estimates already obtained for years 2000 (designated year 0) and 2015 (year 1) to illustrate the reasoning. Let the sample size in year 0 be N and that in year 1 to be kN for some given k since the actual sample sizes in the two years may not be the same. From the results in appendix Tables A1 and A3, one can see that for these two years, k = 2.026 for male earners and k = 2.160 for female earners.

So, more formally, we want "*N*" such that the income share change between the two years is statistically significant at some designated level of confidence — whose critical value on the standard normal table is indicated by the "t-ratio" of t_{crit}. At a 95 percent level of confidence, $t_{crit} = 1.960$. So we want a sample size *N* such that

$$\frac{\hat{\Delta}}{SE(\hat{\Delta})} = t_{crit}$$

or $$\left[SE(\widehat{\Delta})\right]^2 = \left(\frac{(\hat{\Delta})}{t_{crit}}\right)^2$$

ie: $$\frac{Asy.\hat{v}ar\left(\widehat{IS}_i(0)\right)}{N} + \frac{Asy.\hat{v}ar\left(\widehat{IS}_i(1)\right)}{kN} = \left(\frac{\hat{\Delta}}{t_{crit}}\right)^2$$

or $$\widehat{N} = \left(\frac{t_{crit}}{\Delta}\right)^2 \cdot \left[Asy.\hat{v}ar\left(\widehat{IS}_i(0)\right) + \frac{Asy.var\left(\widehat{IS}_i(1)\right)}{k}\right].$$

Table 4

Tests of Changes in Lower and Upper Income Shares

Canada, 2000-2015

a) Lower Shares

	Males		Females	
	p = .20	p = .10	p = .20	p = .10
$\widehat{IS}_{2000}$	.075696	.020215	.058430	.019016
$\widehat{IS}_{2015}$	.064466	.019675	.056208	.018214
$\hat{\Delta}$	-.011230	-.000540	-.002222	-.000802
$SE\left(\widehat{IS}_{2000}\right)$	.0012507	.0006548	.0011274	.0005428
$SE\left(\widehat{IS}_{2015}\right)$	.00081519	.0004295	.00076187	.0003556
$SE\left(\hat{\Delta}\right)$	.0014929	.0007831	.0013607	.0006489
$t-ratio\left(\hat{\Delta}\right)$	-7.522	-0.690	-1.633	-1.236

b) Upper Shares

	q = .80	q = .90	q = .80	q = .90
$\widehat{IS}_{2000}$	.364474	.215057	.380780	.225539
$\widehat{IS}_{2015}$	.383075	.226621	.401294	.234761
$\hat{\Delta}$	+.018601	+.011564	+.020514	+.009222
$SE\left(\widehat{IS}_{2000}\right)$	.0044683	.0036961	.0050274	.0041802
$SE\left(\widehat{IS}_{2015}\right)$	.0033669	.0028066	.0035178	.0029969
$SE\left(\hat{\Delta}\right)$	.0055948	.0046409	.0061359	.0051435
$t-ratio\left(\hat{\Delta}\right)$	+3.325	+2.492	+3.343	+1.793

Source: Author’s calculations based on results in Tables A1-A4.

So it is immediately clear that, to attain a greater level of confidence (ie., higher t_{crit} value), one needs a larger sample size; and similarly a smaller $\widehat{\Delta}$ value also implies a larger required $\widehat{N}$.

Making use of the above equation, one can calculate $\widehat{N}$ for various lower and upper income shares for a given level of confidence and specified $\widehat{\Delta}$ value (expressed as a proportion of $I\hat{S}_i(0)$). For illustrative purposes, use a 95 percent level of confidence and three possible $\widehat{\Delta}$ sizes: 20% difference, 10% difference and 5% difference. Results are tabulated in Table 5 for lower and upper income shares (results for middle income shares appear in appendix Table B4).

Two main findings are evident. First, for a given share width (quintile or decile shares), the required sample sizes are about twice as large for lower income shares than for upper income shares — similar to what was previously found in Tables 1 and 2. Also, required sample sizes for differences in middle income shares are very similar to those differences in upper income shares. Second and most notably, the required sample sizes for *differences* in income shares are dramatically larger than for the income share themselves — by about *two orders of magnitude*. For example, $\widehat{N}_L$ for the lower quintile (p = .20) share for males is 32 (Table 1) vs. the $\widehat{N}_L$ required for a 10 percent difference to be significant of 3812 (Table 5), and the $\widehat{N}_U$ for the upper decile (q = .90) share for females which is also 32 (Table 2) vs. the $\widehat{N}_U$ required for a 10 percent difference to be significant of 4779 (Table 5). Clearly, establishing the statistical significance of *changes* in income shares requires dramatically larger sample sizes than found in the previous section.

Table 5
Required Sample Sizes to Achieve Statistically Significant
Differences in Lower and Upper Income Shares
(at a 95% Level of Confidence)
Canada, 2000-2015

	Males		Females	
a) Lower Shares				
	p = .20	p = .10	p = .20	p = .10
20% Difference	953	3,677	1,246	2,675
10% Difference	3,812	14,707	4,983	10,700
5% Difference	15,248	58,827	19,931	42,800
b) Upper Shares				
	q = .80	q = .90	q = .80	q = .90
20% Difference	577	1,141	596	1,195
10% Difference	2,309	4,564	2,386	4,779
5% Difference	9,236	18,255	9,543	19,115

Source: Author's calculations based on results in Tables A1–A4.

7. Implications for Overall Survey Size

Heretofore in this paper, attention has focused on the size of the actual Estimation Sample used to calculate income shares (and their changes) for some group of income recipients. But the size number usually referred to in general descriptions of surveys is the overall survey size or Survey Sample. The former is a proper subset of the latter. So there are several further considerations to make judgements about how the overall survey may have to change in order to estimate more refined income tail shares to a given level of confidence. We briefly consider these here. (For more general and detailed sources, see such textbooks as Groves et al., 2009, and Chaudhuri and Stenger, 2020. For technical details on the Canadian Labour Force Survey, — which is the data source used in this paper — see Statistics Canada, 2016, 2017, and 2020.) In recent years, the LFS sample size has been about 56,000 households, resulting in labour market information for approximately 100,000 individuals.

Suppose one works from a nation-wide survey. Then the size of the Estimation Sample can be seen to arise from several factors. First is the overall size of the Survey Sample (N^S). Second is the fraction of our group of interest in the population from which the Survey Sample is drawn (eg., male earners with positive earnings in one group, female earners with positive earnings in another group). Let the fraction in the overall population that the group "g" constitutes be represented by F_g (where obviously $0 < F_g < 1$). The group factor F_g obviously depends on both demographics as well as economic behaviour (eg., whether one works in the labour market or not). It also depends upon the Estimation Sample restrictions used (eg., a focus on full-time workers or on some regional or provincial focus). Third is the survey response rate of members of group g. Not everyone selected to be in a survey may choose to respond (meaningfully). Youth or elderly, for example, may be relatively hard to contact, perhaps because of their activities, travels or health status. Let the response rate of members of group g be represented by fraction R_g (where again $0 \leq R_g \leq 1$). Statistics Canada addresses non-responses — which average about 10 percent of selected households — by weighting adjustments and incomplete data generally by imputation procedures. In general, response rates are considerably higher for surveys done by official government agencies such as Statistics Canada or the U.S. Bureau of the Census than done by private sector survey organizations. Response rates also differ quite markedly across countries and social/ethnic groups. The Estimation Sample for group g, then, is

$$N_g = N^s \cdot F_g \cdot R_g \,. \tag{25}$$

What does this imply for the size of N^s required to obtain a tail income share estimate for a given group g to a specified degree of confidence? Now,

$$N^s = \left(F_g\right)^{-1} \cdot \left(R_g\right)^{-1} \cdot N_g. \tag{26}$$

So, for example, if for some g, $F_g = 0.50$ and $R_g = 0.70$, then N^s would have to be $2.86N_g$ in size for a given size of N_g. Thus, for a given required $\widehat{N}_g$, the corresponding required $\widehat{N}^s = 2.86\widehat{N}_g$.

If F_g and R_g are uniform across members of group g in the analysis of the previous section, then the elasticity calculations using $\widehat{N}^s$ yield exactly the same estimated elasticity values as when using $\widehat{N}_g$ since the constant F_g and R_g factors cancel out in the numerator of the elasticities. Thus the same Rule of Thumb values hold for $\widehat{N}^s$ as for $\widehat{N}_g$. That is, in order to estimate a tail vigintile income share for group g with a given level of confidence, $\widehat{N}^s$ would have to increase by a factor of 2 or 3 relative to its size used to estimate a decile tail income share with the same level of confidence.

If response rates do differ across members of the group, the reasoning is similar, but slightly different. (By construction F_g is uniform across members of g.) Indeed, it is very likely that R_g does vary across income groups. For example, many low-income individuals may include some who are ill or elderly immigrants who have difficulty responding in speech or writing, and many high-income individuals may simply choose not to respond because of the high opportunity cost of their time or their aversion to providing requested information such as income. One way to represent this may be to characterize response rates as

$$R_{gi} = R_g \cdot R_i \tag{27}$$

where R_g is the group's average response rate and R_i captures how the response rate varies in general across income classes in the population. (R_g and R_i can both be calculated from Census data for use in developing the survey methodology.) Since, in this paper, we are interested in the tail income/earnings shares, and particularly the bottom and top 5 percent shares, we can make use of the lower (R_L) and upper (R_U) response rate ratios in (27). So, one can estimate required ${N_s}^S$ as

$$\widehat{N}^s = \left(F_g\right)^{-1} \cdot \left(R_{gL}\right)^{-1} \cdot \widehat{N}_g \qquad \text{(28a)}$$

and $$\widehat{N}^s = \left(F_g\right)^{-1} \cdot \left(R_{gU}\right)^{-1} \cdot \widehat{N}_g \qquad \text{(28b)}$$

That is, if, say, $F_g = 0.50, R_g = (0.70), and\ R_i = (0.70)$, then

$$\widehat{N}^s = 4.1 \cdot \widehat{N}_g.$$

So, the figures in (23) and (24) would be about four times larger than calculated in the last section. But again, if F_g and R_{gi} are given constants exogenous to the sampling process for group g, the elasticity values calculated in the last section are unaffected, and the same Rule of Thumb values hold for $\widehat{N}^s$ as for $\widehat{N}_g$. Thus once again, in order to estimate a tail 5 percent income share for a given level of confidence, $\widehat{N}^s$ would have to increase by 2 or 3 times in size relative to that used to estimate a tail 10 percent share with the same level of confidence.

8. Findings and Conclusion

This paper addresses the question of how much survey sample size has to increase in order to estimate tail income shares (and their changes) to a given level of confidence. The focus is on estimating vigintile shares (bottom 5 percent and top 5 percent) given one already has estimates of quintile and decile income shares — as typically published by national statistical agencies such as Statistics Canada and the Bureau of the Census, and changes in quintile and decile shares over time. The paper develops formulas for the specification and estimation of (asymptotic) variances of sample income shares in a distribution-free fashion so that estimates are very straightforward to calculate without having to know the functional form of the underlying distribution of income. These formulas then serve the basis for determining the sample sizes required to estimate an income share to a given degree of reliability or level of confidence.

The paper has four main findings or conclusions. First, it provides distribution-free formulas for the (asymptotic) variances of empirical income shares that can be used to evaluate the reliability or confidence of sample-based income share estimates. Second, the paper presents an approach to deriving the sample sizes required to estimate income shares (and their changes) to a given level of confidence. Third, it provides Rule of Thumb results to calculate how much larger a sample needs to be to estimate a smaller quantile end share to obtain a given level of confidence. Specifically, to estimate vigintile tail income shares, one needs to have a sample 2-3 times larger (depending on whether recipients are men or women and on which tail income share — lower or upper — is being estimated) than that required to estimate corresponding decile shares to the same level of confidence. Fourth, the analysis shows that, in order to significantly estimate changes in income shares over time, the required sample sizes need to be vastly larger than those for simply estimating the income shares themselves — by about two orders of magnitude.

References

Beach, C.M. (2021a) "A Useful Empirical Toolbox for Distributional Analysis", Queen's University, Department of Economics Discussion Paper No. 1466, August.

— (2021b) "A Nifty Fix for Published Distribution Statistics: Simplified Distribution-Free Statistical Inference", Queen's University, Department of Economics Discussion Paper No. 1477, September.

Bell, B., N. Bloom, and J. Blundell (2022) "Income Dynamics in the United Kingdom and the Impact of the Covid-19 Recession", *Quantitative Economics* 13(4), 1849-78.

Bowlus, A., E. Gouin-Bonenfant, H. Liu, L. Lochner, and Y. Park (2022) "Four Decades of Canadian Earnings Inequality and Dynamics Across Workers and Firms", *Quantitative Economics* 13(4), 1447-91.

Chaudhuri, A., and H. Stenger (2020) *Survey Sampling: Theory and Methods*, 2nd Edition. New York, N.Y.: CRC Press, Routledge Publishing.

Davidson, R. (2018) "Statistical Inference on the Canadian Middle Class", *Econometrics*, Special Issue on Econometrics and Inequality 6(1), 14, 1-18.

Drummond, D. (2021) "Viewpoint: Canada Should Establish an Equitable Growth Institute", Centre for the Study of Living Standards, *International Productivity Monitor*, No. 38 (Spring).

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

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

Groves, R.M., F.J. Fowler, Jr., M.P. Couper, J.M. Lepkowshi, E. Singer, and R. Tourangeau (2009) *Survey Methodology*, 2nd Edition. Hoboken, N.J.: John Wiley & Sons.

Guvenen, F., L. Pistaferri, and G.L. Violante (2022) "Global Trends in Income Inequality and Income Dynamics: New Insights from GRID", *Quantitative Economics* 13(4), 1321-1360.

Kendall, M.G., and A. Stuart (1969) *The Advanced Theory of Statistics*, vol. 1. London: Charles Griffen & Co.

Lin, P.-E., K.-T. Wu, and I.A. Ahmad (1979) "Asymptotic Joint Distribution of Ratios of Sample Quantiles to Sample Mean", Florida State University, FSV Statistics Report M492, February.

— (1980) "Asymptotic Joint Distribution of Sample Quantiles and Sample Mean with Applications", *Communications in Statistics — Theory and Methods* 9(1), 51-60.

Lohr, S. (2022) "Economists Eye Tech's Influence on Inequality", *The Globe and Mail* Report on Business, Jan. 12, B4.

McKinney, K.L., J.M. Abowd, and H.P. Janicki (2022) "U.S. Long-Term Earnings Outcomes by Sex, Race, Ethnicity, and Place of Birth", *Quantitative Economics* 13(4), 1879-1945.

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

Statistics Canada (2016) "The 2015 Revisions of the Labour Force Survey (LFS)", publication 71F0031X.

— (2017) "History of the Canadian Labour Force Survey, 1945 to 2016", by Jeanine Usalcas and Mark Kinock (January 6).

Statistics Canada (2018) CANSIM Table 206-0031, "Upper Income Limit, Income Share and Average of Market, Total and After-Tax Income by Economic Family Type and Income Decile, Canada and Provinces".

— (2020) "Guide to the Labour Force Survey" (April 9).

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

Data Appendix

Table A1

Sample Estimates Canada 2015

	Males	Females
ξ_{10}	284.08	179.67
ξ_{20}	425.53	302.13
ξ_{40}	630.42	472.81
ξ_{60}	873.13	651.54
ξ_{80}	1197.79	921.36
ξ_{90}	1485.11	1197.73
IS_{10}	.019675	.018214
IS_{20}	.064466	.056208
IS_{40}	.193906	.185871
IS_{top40}	.629281	.644194
IS_{top20}	.383075	.401294
IS_{top10}	.226621	.234761
σ	501.77	404.54
μ	832.91	633.38
NOBS	51,680	51,658

Source: LFS data on usual weekly earnings from May Labour Force Surveys (figures in 2015 dollars). Results from Beach (2021b), appendix tables.

Table A2

Estimates of $Asy.var\left(\widehat{IS}_i\right)$ for Lower and Upper Income Shares

Canada, 2015

	Males		Females	
	AV	ACV	AV	ACV
a) Lower Income Shares				
p=.40	.09064	1.5526	.08843	1.6000
p=.20	.03434	2.8746	.02998	3.0805
p=.10	.00953	4.9617	.00653	4.4366
b) Upper Income Shares				
(1-q) = .40	.70995	1.3390	.74689	1.3416
(1-q) = .20	.58586	1.9981	.63928	1.9924
(1-q) = .10	.40710	2.8155	.46397	2.9015

Source: Author's calculations based on parameter estimates in Table A1.

Note: AV refers to the estimated asymptotic variance, ACV refers to the $(asymptotic^\wedge variance)^{1/2}$ / estimated income share; ie., the estimated asymptotic coefficient of variation.

Table A3

Sample Estimates – Canada 2000.

	Males	Females
ξ_{10}	279.45	158.06
ξ_{20}	421.50	258.17
ξ_{40}	628.03	405.48
ξ_{60}	810.62	569.02
ξ_{80}	1053.74	790.31
ξ_{90}	1279.77	972.81
IS_{10}	.020215	.019016
IS_{20}	.075696	.058430
IS_{40}	.202345	.182351
IS_{top40}	.609463	.635287
IS_{top20}	.364474	.380780
IS_{top10}	.215057	.225539
σ	427.01	327.11
μ	765.69	536.18
NOBS	25,511	23,917

Source: See Table A1.

Table A4

Estimates of $Asy.var(\widehat{IS}_i)$ for Lower and Upper Income Shares

Canada, 2000

	Males		Females	
	AV	ACV	AV	ACV
a) Lower Income Shares				
p=.20	.03991	2.6391	.03040	2.9840
p=.10	.01094	5.1736	.00705	4.4147
b) Upper Income Shares				
(1-q) = .20	.50934	1.9581	.60450	2.0419
(1-q) = .10	.34852	2.7451	.41794	2.8664

Source: Author's calculations based on parameter estimates in Table A3.

Note: AV refers to the estimated asymptotic variance, ACV refers to the $(asymptotic\hat{}variance)^{1/2}/estimated\ income\ share$; ie., the estimated asymptotic coefficient of variation.

Appendix B

Estimates of Required Sample Sizes for Middle Income Shares

As indicated in Section 3 of the text, the formula for the asymptotic variance of the middle income group M is given by:

$$
\begin{aligned}
Asy.var\left(\widehat{IS}_M\right) = {} & p(1-p)\left(\frac{\xi_L}{\mu}\right)^2 + q(1-q)\left(\frac{\xi_U}{\mu}\right)^2 + (IS_M)^2\left(\frac{\sigma}{\mu}\right)^2 \\
& -2\left(\frac{\xi_L}{\mu}\right)\left(\frac{\xi_U}{\mu}\right)p(1-q) \\
& +2\left(\frac{\xi_L}{\mu}\right)(IS_M)[p - IS_L] \\
& -2\left(\frac{\xi_U}{\mu}\right)(IS_M)[IS_U - (1-q)]
\end{aligned}
$$

with the same notation as given in the text. Parameter estimate information for 2015 is provided in Table A1 and for 2000 in Table A3. In addition, for 2015 and 2000, the middle income share estimates are:

	Males		Females	
	2000	2015	2000	2015
$\widehat{IS}(.20 \rightarrow .80)$	.559830	.552459	.560789	.542498
$\widehat{IS}(.40 \rightarrow .60)$	.188193	.176812	.182362	.169934

The formula for calculating $\widehat{N}_M$, the required sample size for a given year for testing whether $\widehat{IS}_M$ is statistically significant at a critical value given by t_{crit} is:

$$\widehat{N}_M = (t_{crit})^2 \cdot \left[\frac{Asy.\widehat{v}ar(\widehat{IS}_M)}{(\widehat{IS}_M)^2}\right]$$

where t_{crit} is the critical value from a standard normal distribution corresponding to specified level of confidence (95% or 99%).

Resulting estimates for $\widehat{N}_M$ are presented in Table B1 with corresponding elasticities with respect to the interval width (q-p) provided in Table B2.

The main finding is that the required sample size to estimate a middle income share to a given level of confidence is about the same as that for an upper income share for a given interval width and about half that required to estimate a lower income share of the same interval width at the same level of confidence. That is, it requires a much larger sample size to estimate lower income shares than for other income shares for a given degree of confidence.

Table B1

Estimates of $\widehat{N}_M$ for Middle Income Shares

Canada, 2015

	Males		Females	
	$\widehat{N}_M$	%Δ	$\widehat{N}_M$	%Δ
a) 95% Level of Confidence				
$(q-p) =$ .60	2.593		2.719	
		2.452		2.306
$(q-p) =$ .40	8.952		8.991	
		0.710		0.697
$(q-p) =$ .20	15.310		15.262	
b) 99% Level of Confidence				
$(q-p) =$ .60	4.479		4.697	
		2.452		2.306
$(q-p) =$ .40	15.462		15.530	
		0.710		0.697
$(q-p) =$ .20	26.444		26.362	

Source: Author's calculations based on parameter estimates in Table A3 in the Data Appendix.

Note: The interval $(q-p) =$ .60 corresponds to p= .20 and q= .80.
The interval $(q-p) =$ .40 corresponds to p= .30 and q= .70.
The interval $(q-p) =$ .20 corresponds to p= .40 and q= .60.
The first and third of these were estimated directly from the formulas in this appendix. The second row figures were estimated by proportional interpolations from the first and third row figures.

Table B2

Elasticities of $\widehat{N}_M$ wrt Income Share Width

Canada, 2015

	Males	Females
For $q - p$: .60 → .40	7.356	6.918
For $q - p$: .40 → .20	1.420	1.394

Source: Author's calculations from results in Table B1.

Table B3

Tests of Changes in Middle Income Shares

Canada, 2000-2015

	Males		Females	
Middle Shares				
	For $q - p = .60$	For $q - p = .20$	For $q - p = .60$	For $q - p = .20$
$\widehat{IS}_{2000}$	.55983	.18819	.56079	.18236
$\widehat{IS}_{2015}$	.55246	.17681	.54250	.16993
Δ	-.00737	-.01138	-.01829	-.01243
$SE(\widehat{IS}_{2000})$	.002880	.002352	.003051	.002350
$SE(\widehat{IS}_{2015})$	.002017	.001579	.001949	.001541
$SE(\Delta)$	.003565	.002833	.003620	.002810
$t - ratio(\Delta)$	-2.067	-4.017	-5.052	-4.423

Source: Author's calculations based on results in Tables A1-A4.

Table B4

Required Sample Sizes to Achieve Statistically Significant

Differences in Middle Income Shares

(at a 95% Level of Confidence)

Canada, 2000-2015

	Males		Females	
Middle Shares				
	$q - p = .60$	$q - p = .20$	$q - p = .60$	$q - p = .20$
20% Difference	97	552	96	546
10% Difference	387	2,221	383	2,182
5% Difference	1,546	8,883	1531	8,728

Source: Author's calculations based on results in Tables A1-A4.

Appendix C

Estimates of Required Sample Sizes for Lower and Upper Income Shares

Canada, 2000

Table C1

Estimates of $\widehat{N}_L$ for Lower Income Shares

Canada, 2000

	Males		Females	
	$\widehat{N}_L$	%Δ	$\widehat{N}_L$	%Δ
a) 95% Level of Confidence				
$p = .20$	26.76		34.21	
		2.843		1.189
$p = .10$	102.83		74.88	
b) 99% Level of Confidence				
$p = .20$	46.22		59.09	
		2.843		1.189
$p = .10$	177.62		129.33	

Source: Author's calculations based on parameter estimates in Table A3 in the Data Appendix.

Table C2

Estimates of $\widehat{N}_U$ for Upper Income Shares

Canada, 2000

	Males		Females	
	$\widehat{N}_U$	%Δ	$\widehat{N}_U$	%Δ
a) 95% Level of Confidence				
$(1-q) = .20$	14.73		16.02	
		.966		.970
$(1-q) = .10$	28.95		31.57	
b) 99% Level of Confidence				
$(1-q) = .20$	25.44		27.67	
		.966		.970
$(1-q) = .10$	50.01		54.52	

Source: Author’s calculations based on parameter estimates in Table A3 in the Data Appendix.

Table C3

Estimates of $\widehat{N}_M$ for Middle Income Shares

Canada, 2000

	Males		Females	
	$\widehat{N}_M$	%Δ	$\widehat{N}_M$	%Δ
a) 95% Level of Confidence				
$(q-p) = .60$	2.59		2.72	
		2.454		2.303
$(q-p) = .40$	8.95		8.99	
		.711		.698
$(q-p) = .20$	15.31		15.26	
b) 99% Level of Confidence				
$(q-p) = .60$	4.48		4.70	
		2.454		2.303
$(q-p) = .40$	15.47		15.52	
		.711		.698
$(q-p) = .20$	26.45		26.35	

Source: Author’s calculations based on parameter estimates in Table A3 in the Data Appendix.

Note: See Note to Table B1.

Table C4

Elasticities of $\widehat{N}_L, \widehat{N}_U, and\ \widehat{N}_M$ wrt Income Share Width

Canada, 2000

	Males	Females
a) For $\Delta p\ and\ \widehat{N}_L$		
$.20 \rightarrow .10$	5.686	2.378
b) For $\Delta(1-q)\ and\ \widehat{N}_U$		
$.20 \rightarrow .10$	1.932	1.940
c) For $\Delta(q-p)\ and\ \widehat{N}_M$		
$.60 \rightarrow .40$	7.362	6.909
$.40 \rightarrow .20$	1.422	1.396

Source: Author's calculations from results in Tables C1-C3.