Journal of Official Statistics, Vol. 13, No. 1, 1997, pp. 41–58

Variance Estimation for Measures of Income Inequality and Polarization – The Estimating Equations Approach

Milorad S. Kovačević[1] *and David A. Binder*[2]

The estimating equations technique for variance estimation is demonstrated on a variety of income inequality and polarization measures when data are obtained in a complex survey. This method, based on the Taylor linearization, is computationally nonintensive and easy to implement. Six different measures are considered. An example based on data from the Canadian Survey of Consumer Finance is given.

Key words: Coefficient of variation; exponential measure; Gini index; Lorenz curve ordinates; polarization curve ordinates; polarization index.

1. Introduction

Estimates of measures of income inequality and polarization are often required in studies of income distributions. When income distributions are compared from region to region or through time one should account for their sampling variability. Lack of information on standard errors confines the role of measures of income inequality to that of descriptive devices rather than inferential tools for formal statistical inference.

The common characteristic of these measures is their complexity. They are nonlinear functions of the observations. Some of them depend on the ordered observations or quantiles. In addition, income data usually come from complex surveys (stratified, multistage, cluster samples with unequal probabilities of selection). Consequently, the variances of these measures are not expressible by simple formulae. Since they cannot be estimated by conventional variance estimation methods, one has to rely on approximate variance estimation techniques.

Here an approximate standard error estimation technique is presented. It is based on the theory of estimating equations (EE) as developed by Binder (1991), and Binder and Patak (1994). The problem of estimating income inequality measures and their standard errors using estimating equations has been addressed by Binder (1992) and the extension to survey sampling was made by Binder and Kovačević (1995). In this article, the EE method is applied to estimating some common measures of income inequality and their standard errors that were not addressed previously by Binder and Kovačević (1995). These

[1] Milorad S. Kovačević is a Senior Methodologist with the Social Survey Methods Division, Statistics Canada.
[2] David A. Binder is Director of the Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Acknowledgment: We thank H.J. Mantel for his careful reading and useful comments on previous versions of this article. We are also grateful to the associate editor and referees for suggestions improving the readability of the article.

measures are the coefficient of variation and the exponential measure of inequality. We also present the estimation of the polarization curve and the polarization index as defined in Foster and Wolfson (1992) and Wolfson (1994).

Some basic EE theory is reviewed in Section 2. The finite population versions of the measures of income inequality and polarization and their estimates based on a complex sample are introduced in Section 3. This section and the Appendix contain derivations of the standard error estimators for these measures. Finally, we apply EE methodology to data on earnings from the Canadian Survey of Consumer Finance from 1992. In the summary section we give a table with expressions for the most frequently used income inequality and polarization measures and their standard errors.

In this article we consider only the form of the estimator of the sampling variance for these complex measures of income inequality. A fuller understanding of the properties of these estimators requires the use of simulation studies. Results from a large simulation study undertaken at Statistics Canada (Kovačević, Yung, and Pandher 1995) strongly confirm the advantage of the EE method over several competing methods for the variance estimation of measures based on order statistics – such as quantiles – Lorenz curve ordinates and the polarization index. For the same measures, the study showed that the EE method and the bootstrap method performed similarly. We summarize the relevant findings of the study in Section 4.

2. The Estimating Equations Method in Survey Sampling – A Review

A general formulation of the EE approach for large sample complex surveys is given in Binder and Patak (1994). In this section we summarize the main results needed for the derivations in Section 3. Whereas Godambe and Thompson (1986) considered the optimality of estimating equations, the approach we take follows that of Binder and Patak (1994) who concentrate on the asymptotic properties for a given set of EE's.

In general, for infinite populations with a continuous distribution function $F(y;\theta)$ and differentiable density function $f(y;\theta)$, an estimate of the parameter θ may be obtained from the maximum likelihood equation $U(\theta)=\Sigma\partial\log f(y_i;\theta)/\partial\theta=0$. The optimality of this estimating equation for superpopulation models is discussed by Godambe and Thompson (1986).

In the case of the finite population, parameters are defined explicitly as functions of the y values of all population units, i.e., $\theta=g(y_1,\ldots,y_N)$. They may also be defined implicitly, for example, by maximizing the likelihood function derived by considering the finite population as a sample from an infinite parametric population, or by minimizing a certain loss function. Examples are the population mean $\mu_N=\Sigma Y_i/N$ for the first, and the regression coefficient for the second type

$$\sum(y_i-x_i'\beta)^2\rightarrow\min.$$

In both cases, the parameter θ can be regarded as the solution, θ_N, to the equation

$$U(\theta)=\sum_{i=1}^{N}u(y_i,\theta)=0 \tag{1}$$

The choice of an estimating function $u(y_i,\theta)$ for a particular parameter may not be unique in general. However, the inference obtained using our approach will not depend on the choice of estimating function, even when more than one set of estimating functions is used to define the parameter of interest.

To estimate a finite population parameter when the estimating function $u(y_i, \theta)$ is given, suppose that a random sample s from a finite population is available. The expression $U(\theta)$ in (1) can be estimated by an extension of the Horvitz-Thompson (HT) unbiased estimator, as discussed by Rao (1979), as

$$\hat{U}(\theta) = \sum_{i=1}^{N} w_i(s) u(y_i, \theta) \tag{2}$$

where the weight $w_i(s)$ is equal to 0 whenever the ith unit is not in the sample and $\Sigma_{i=1}^{N} w_i(s) = \hat{N}$.

If we use the HT estimator, the weights are the inverse of the inclusion probabilities

$$w_i(s) = \begin{cases} 1/\pi_i, & i \in s \\ 0, & i \notin s \end{cases}$$

Or, for example, if we use general regression estimation based on an auxiliary variable x

$$w_i(s) = \begin{cases} \dfrac{1}{\pi_i}\left[1 + (X - \hat{X})\dfrac{x_i}{\hat{X}'}\right], & i \in s \\ 0, & i \notin s \end{cases}$$

where X is the known population total for the variable x, and $\hat{X}$ and $\hat{X}'$ are the HT estimates of the totals for x and x^2, respectively.

The solution $\hat{\theta}$ of the equation $\hat{U}(\theta) = 0$ is the EE estimate of the finite population parameter θ_N.

The estimating function $u(y, \theta)$ provides a Gauss consistent estimate (Godambe and Kale 1991). That is, if we observe all values in the finite population then the estimate obtained using the estimating equation is equal to the parameter.

To estimate the variance of $\hat{\theta}$ we proceed as follows. Equation (2) can be rewritten as

$$0 = \hat{U}(\hat{\theta}) = \sum_{i=1}^{N} [u(y_i, \hat{\theta}) - u(y_i, \theta_N)] + \sum_{i=1}^{N} w_i(s) u(y_i, \theta_N) + \sum_{i=1}^{N} [u(y_i, \hat{\theta}) - u(y_i, \theta_N)][w_i(s) - 1] \tag{3}$$

This decomposition of the estimating equation is analogous to the one given in Binder (1991). We denote the last term in (3) by R. The remainder R is generally of order $o(|\hat{\theta} - \theta_N|)$, which is asymptotically negligible as $\hat{\theta} \rightarrow \theta_N$. Expanding the function $u(y, \hat{\theta})$ around θ_N using Young's form of Taylor's Theorem (Serfling 1980, p. 45), we have

$$0 = \hat{U}(\hat{\theta}) = \sum_{i=1}^{N} (\hat{\theta} - \theta_N) \left.\frac{\partial u(y_i, \theta)}{\partial \theta}\right|_{\theta = \theta_N} + o(|\hat{\theta} - \theta_N|) + \sum_{i=1}^{N} w_i(s) u(y_i, \theta_N) + R$$

Ignoring the remainder terms, the difference $\hat{\theta} - \theta_N$ can be expressed as

$$\hat{\theta} - \theta_N \approx \sum_{i=1}^{N} w_i(s) u^*(y_i, \theta_N)$$

where

$$u^*(y_i, \theta_N) = -J_\theta^{-1} u(y_i, \theta_N) \text{ and } J_\theta = \sum_{i=1}^{N} \frac{\partial u(y_i, \theta)}{\partial \theta}\bigg|_{\theta = \theta_N}$$

Once the expression for $u^*(y, \theta_N)$ is obtained, estimation of the mean squared error of $\hat{\theta}$ becomes straightforward. Since $\hat{\theta} - \theta_N$ can be approximated by an estimator of the population total of $u^*(y_i, \theta_N)$'s, we can use the variance estimation technique for the estimate of a total, i.e.,

$$var(\hat{\theta}) = var(\hat{\theta} - \theta_N) \approx var\left(\sum_s w_i u^*(y_i, \theta_N) \right) \tag{4}$$

Note that $u^*(y_i, \theta_N)$ depends on an unknown parameter. When we substitute its estimate into $u^*(y, \theta_N)$, we obtain $u_i^* = u^*(y_i, \hat{\theta})$, and the value of $\Sigma_s w_i u_i^*$ is exactly zero. To approximate the variance of $\hat{\theta}$, we must treat the u^*'s according to the EE approach explained in Binder and Patak (1994) and replace θ_N by $\hat{\theta}$ only in the final expression of the variance (4).

Generally, the parameter θ is multidimensional. Binder (1991) and Binder and Patak (1994) considered the case where the first component of the vector θ is the parameter of interest θ_N and the others are nuisance parameters λ_N. In this case $U(\theta) = U(\theta, \lambda)$ and can be partitioned as $[U_1(\theta, \lambda), U_2(\theta, \lambda)]'$ and estimated so that the following equality holds

$$0 = \begin{bmatrix} \hat{U}_1(\hat{\theta}, \hat{\lambda}) \\ \hat{U}_2(\hat{\theta}, \hat{\lambda}) \end{bmatrix} \tag{5}$$

A decomposition similar to (3) can be applied to (5). Hence

$$0 = \begin{bmatrix} \hat{U}_1(\hat{\theta}, \hat{\lambda}) \\ \hat{U}_2(\hat{\theta}, \hat{\lambda}) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}[u_1(y_i, \hat{\theta}, \hat{\lambda}) - u_1(y_i, \theta_N, \hat{\lambda})] + \sum_{i=1}^{N}[u_1(y_i, \theta_N, \hat{\lambda}) - u_1(y_i, \theta_N, \lambda_N)] \\ \sum_{i=1}^{N}[u_2(y_i, \hat{\theta}, \hat{\lambda}) - u_2(y_i, \theta_N, \hat{\lambda})] + \sum_{i=1}^{N}[u_2(y_i, \theta_N, \hat{\lambda}) - u_2(y_i, \theta_N, \lambda_N)] \end{bmatrix}$$

$$+ \begin{bmatrix} \sum_{i=1}^{N} w_i(s) u_1(y_i, \theta_N, \lambda_N) \\ \sum_{i=1}^{N} w_i(s) u_2(y_i, \theta_N, \lambda_N) \end{bmatrix} + R$$

where

$$R = \begin{bmatrix} \sum_{i=1}^{N}[u_1(y_i, \hat{\theta}, \hat{\lambda}) - u_1(y_i, \theta_N, \lambda_N)](w_i(s) - 1) \\ \sum_{i=1}^{N}[u_2(y_i, \hat{\theta}, \hat{\lambda}) - u_2(y_i, \theta_N, \lambda_N)](w_i(s) - 1) \end{bmatrix}$$

and is negligible whenever $R = o(|\hat{\theta} - \theta|)$.

Expanding the function $\hat{U}(\hat{\theta}, \hat{\lambda})$ around (θ_N, λ_N) we obtain

$$0 \approx \begin{bmatrix} J_{1\theta} J_{1\lambda} \\ J_{2\theta} J_{2\lambda} \end{bmatrix} \begin{bmatrix} \hat{\theta} - \theta_N \\ \hat{\lambda} - \lambda_N \end{bmatrix} + \begin{bmatrix} \hat{U}_1(\theta_N, \lambda_N) \\ \hat{U}_2(\theta_N, \lambda_N) \end{bmatrix} \tag{6}$$

where

$$J_{1\theta} = \left.\frac{\partial U_1(\theta, \lambda)}{\partial \theta}\right|_{\theta=\theta_N, \lambda=\lambda_N}, \quad J_{1\lambda} = \left.\frac{\partial U_1(\theta, \lambda)}{\partial \lambda}\right|_{\theta=\theta_N, \lambda=\lambda_N}, \quad J_{2\theta} = \left.\frac{\partial U_2(\theta, \lambda)}{\partial \theta}\right|_{\theta=\theta_N, \lambda=\lambda_N}$$

and $J_{2\lambda} = \left.\frac{\partial U_2(\theta, \lambda)}{\partial \lambda}\right|_{\theta=\theta_N, \lambda=\lambda_N}$

Matrices $J_{1\theta}$, $J_{1\lambda}$, $J_{2\theta}$, and $J_{2\lambda}$ are of order 1×1, $1 \times k$, $k \times 1$, and $k \times k$, respectively, where k is the number of nuisance parameters. Solving equation (6) with respect to the difference $\hat{\theta} - \theta_N$ we obtain

$$\begin{aligned} \hat{\theta} - \theta_N \approx & - [J_{1\theta}^{-1} + J_{1\theta}^{-1} J_{1\lambda} (J_{2\lambda} - J_{2\theta} J_{1\theta}^{-1} J_{1\lambda})^{-1} J_{2\theta} J_{1\theta}^{-1}] \hat{U}_1(\theta_N, \lambda_N) \\ & + J_{1\theta}^{-1} J_{1\lambda} (J_{2\lambda} - J_{2\theta} J_{1\theta}^{-1} J_{1\lambda})^{-1} \hat{U}_2(\theta_N, \lambda_N) \end{aligned}$$

In most cases of practical importance we find that $U_2(\theta, \lambda)$ does not depend on θ, so that $J_{2\theta} = 0$. Also, we assume that the first derivatives of the functions $u_1(y, \theta, \lambda)$ and $u_2(y, \theta, \lambda)$ with respect to θ are independent of λ. Taking these assumptions into account we reduce the above expression to

$$\begin{aligned} \hat{\theta} - \theta_N &\approx [-\hat{U}_1(\theta_N, \lambda_N) + J_{1\lambda} J_{2\lambda}^{-1} \hat{U}_2(\theta_N, \lambda_N)] J_{1\theta}^{-1} \\ &= \sum_{i=1}^{N} w_i(s) [-u_1(y_i, \theta_N, \lambda_N) + J_{1\lambda} J_{2\lambda}^{-1} u_2(y_i, \theta_N, \lambda_N)] J_{1\theta}^{-1} \\ &= \sum_{i=1}^{N} w_i(s) u^*(y_i, \theta_N, \lambda_N) \end{aligned} \tag{7}$$

Using this expression we can estimate the variances of various complex statistics as the variances of the estimated totals. As mentioned earlier, to approximate the variance of $\hat{\theta}$, we replace θ_N, λ_N, and possibly N, by $\hat{\theta}$, $\hat{\lambda}$ and $\Sigma_s w_i$, respectively, only in the final expression of the variance. It should be noted that the derivation of this expression in Binder and Patak (1994) was based directly on confidence interval construction.

Income data are usually collected on a stratified, multistage sample with many strata, where a few primary sampling units (clusters), $n_h (\geq 2)$, are sampled from each stratum without replacement. However, to simplify calculation for variance estimation we assume that the clusters are selected with replacement. Such an approximation leads to a conservative estimate of the variance with a small relative bias when the number of primary sampling units is small in each stratum. Let w_{hci} be the weight attached to the ith ultimate unit in the cth cluster of the hth stratum such that the appropriate estimator of the population total for some characteristic, say x, is

$$\hat{T} = \sum_s w_{hci} x_{hci}$$

Then its variance can be estimated by

$$var(\hat{T}) = \sum_h \frac{n_h}{n_h - 1} \sum_c \left(\sum_i w_{hci} x_{hci} - \frac{\sum_c \sum_i w_{hci} x_{hci}}{n_h} \right)^2$$

Accordingly,

$$var(\hat{\theta}) = var(\hat{\theta} - \theta_N) \approx var\left(\sum_s w_{hci} u^*(y_{hci}, \theta_N, \lambda_N) \right)$$

$$\approx \sum_h \frac{n_h}{n_h - 1} \sum_c \left(\sum_i w_{hci} u^*_{hci} - \frac{\sum_c \sum_i w_{hci} u^*_{hci}}{n_h} \right)^2$$

where $u^*_{hci} = u^*(y_{hci}, \hat{\theta}, \hat{\lambda})$.

3. Income Inequality and Polarization Measures and Their Standard Errors

We now consider several measures of inequality and polarization. Usually these measures are computed from grouped data. Here we provide their estimates based on a probability sample from a finite population. Also, we derive the estimates of their standard errors. Since our goal is not to study these measures and their properties in depth, we summarize some general notions about measuring inequality using the selected measures.

Determining whether the values of x are ''more equal'' than the values of y essentially coincides with measuring the dispersion of the corresponding distributions. Depending on whether we are interested in inequality in a particular segment or in the whole distribution, the relevant measures may be categorized as extreme, average or summary measures. Extreme measures focus on inequality in the tails of the income distribution. In this article we include one such measure, the exponential measure (Wolfson 1986), which is low-income sensitive in the sense that the main contribution to its value comes from small or negative incomes. A smaller value of the exponential measure indicates greater equality among the poor.

Although the coefficient of variation belongs formally to the category of average inequality measures, because of its sensitivity to high incomes it is often used to assess inequality among high-income earners. While the coefficient of variation measures deviation from a central value (the mean income), the Gini index accounts for the deviations of all the values in the population among themselves. In this way the Gini index is an average measure of inequality, and very robust to inequalities in the tails.

The primary summary measure is the Lorenz curve. For a given distribution it plots the cumulative percentage of the population (displayed from the poorest to the richest) against its total income share. The area between the Lorenz curve and the 45-degree line is known as the Lorenz area. The Gini index is equal to twice the Lorenz area. A population with the Lorenz curve closer to the 45-degree line has a more equal distribution of income. If all incomes are the same, the Lorenz curve degenerates to the 45-degree line. However, if Lorenz curves for two or more income distributions intersect, only a partial ranking of the distributions is possible. In such cases, the use of different measures may rank the distributions differently. A small set of inequality measures that are sensitive to different magnitudes of income could help interpretation of such ‘‘conflicting’’ findings.

Since the early 1980s there have been discussions about the disappearing middle class phenomenon which is essentially different from the notion of inequality in income (or wealth) distribution and therefore needs a different quantification. The usual term for the shrinking of the middle class is polarization, indicating that the middle is moving toward the tails. Here we present two summary polarization measures, the polarization curve and the polarization index, as defined by Foster and Wolfson (1992).

In this section we present six measures of income inequality and polarization along with their complex sample estimators and the u^* variates needed for variance estimation via the EE method. To simplify notation we drop the subscript N from the finite population parameters.

3.1. The coefficient of variation

The coefficient of variation (squared) is defined as $CV^2 = V/\mu^2$ where V is the variance and μ is the mean income of the population. As a measure of income inequality it belongs to the family of measures that are high-income sensitive and relatively robust to the low income part of the population. It is easy to calculate and has a familiar interpretation. Its disadvantage is, however, a high sampling variability. The CV^2 can be obtained as a solution to the system of equations

$$\begin{cases} U_1(CV^2, \mu) = \sum_U [(y_i/\mu - 1)^2 - CV^2] = 0 \\ U_2(CV^2, \mu) = \sum_U (y_i - \mu) = 0 \end{cases}$$

Note that the parameter of interest is CV^2 and the nuisance parameter is μ.

Then the estimate of the coefficient of variation can be obtained as a solution to the following estimating equations

$$\begin{cases} \hat{U}_1(CV^2, \mu) = \sum_s w_i[(y_i/\mu - 1)^2 - CV^2] = 0 \\ \hat{U}_2(CV^2, \mu) = \sum_s w_i(y_i - \mu) = 0 \end{cases}$$

and takes on the familiar form

$$\hat{CV}^2 = \frac{1}{\hat{N}} \sum_s w_i(y_i/\hat{\mu} - 1)^2$$

where $\hat{\mu} = \Sigma_s w_i y_i / \hat{N}$.

In order to estimate the variance of $\hat{CV}^2$, the u^* variates are needed. Their derivation is essentially a three-step procedure:

(i) First we determine the derivatives

$$J_1, CV^2 = -N, J_{1,\mu} = -2NCV^2/\mu, \text{ and } J_{2,\mu} = -N$$

(ii) Next, we substitute these derivatives into equation (7) and obtain

$$u^*(y_i, CV^2, \mu) = [(y_i/\mu - 1)^2 - (2y_i/\mu - 1)CV^2]/N$$

(iii) Finally, $var(\hat{CV}^2) = var(\Sigma_s w_i u^*(y_i, CV^2, \mu))$.

In the case of multistage sampling the estimated variance takes the form

$$var(\hat{CV}^2) = \sum_h \frac{n_h}{n_h - 1} \sum_c (u^*_{hc} - \bar{u}^*_h)^2$$

where $u^*_{hc} = \Sigma_i w_{hci} u^*_{hci}$, $u^*_{hci} = u^*(y_{hci}, \hat{CV}^2, \hat{\mu})$ and $\bar{u}^*_h = \Sigma_c u^*_{hc}/n_k$.

The remainder term R in the case of $\hat{CV}^2$ is

$$R = \begin{bmatrix} \sum_U [(y_i/\hat{\mu} - 1)^2 - \hat{CV}^2 - (y_i/\mu - 1)^2 + CV^2](w_i - 1) \\ \sum_U [(y_i - \hat{\mu}) - (y_i - \mu)](w_i - 1) \end{bmatrix}$$

The first component, after some simplification, becomes asymptotically equivalent to the product $o(|\hat{CV}^2 - CV^2|) \cdot o(|\hat{\mu} - \mu|) \cdot o(|\hat{N} - N|)$ as $\hat{CV}^2 \rightarrow CV^2$, $\hat{\mu} \rightarrow \mu$ and $\hat{N} \rightarrow N$. Similarly, the second component is equivalent to $o(|\hat{\mu} - \mu|) \cdot o(|\hat{N} - N|)$.

For the *coefficient of variation (unsquared)* a different expression for u^*_i is obtained

$$u^*_i = \frac{1}{2}[(y_i/\hat{\mu} - 1)^2/\hat{CV} - (2y_i/\hat{\mu} - 1)\hat{CV}]/\hat{N}$$

3.2. The exponential measure

The exponential measure is defined as the population mean of the exponentially transformed income (see Wolfson 1986)

$$EX = \frac{1}{N} \sum_U \exp(-y_i/\mu)$$

This measure is sensitive to low income values but takes a reasonable finite value when income is in the neighbourhood of zero. Also, it is well defined for negative incomes.

It can be obtained as a solution to the equations

$$\begin{cases} U_1(EX, \mu) = \sum_U \{\exp(-y_i/\mu) - EX\} = 0 \\ U_2(\mu) = \sum_U (y_i - \mu) = 0 \end{cases} \tag{8}$$

The derivation of u^* variates is similar to the previous case of the coefficient of variation and is given in detail in the Appendix. Here we present just the final form

$$u^*(y_i, \hat{EX}, \hat{\mu}) = [\exp(-y_i/\hat{\mu}) - \hat{EX} + (y_i - \hat{\mu})\hat{J}_{1,\mu}/\hat{N}]/\hat{N}$$

where $\hat{J}_{1,\mu} = (1/\hat{\mu}^2)\Sigma_s w_i y_i \exp(-y_i/\hat{\mu})$

3.3. The Lorenz curve

Earlier we described the Lorenz curve as a powerful descriptive and analytic tool for ranking income distributions. It simply depicts the cumulative income against the population share. The formal finite population definition of the Lorenz curve can be stated as

$$L(p) = \frac{1}{N\mu}\sum_U y_i I\{y_i \le \xi_p\} \quad 0 \le p \le 1$$

where $I\{.\}$ denotes an indicator function and ξ_p is the pth population income quantile. To use the EE method, the Lorenz curve ordinates can be expressed as the solution to the system of equations

$$\begin{cases} \sum_U [I\{y_i \le \xi_p\} - L(p)]y_i & = 0 \\ \sum_U [I\{y_i \le \xi_p\} - p] & = 0 \qquad \text{for } 0 \le p \le 1 \end{cases}$$

The second equation defines the finite population quantile. The resulting sample estimate is

$$\hat{L}(p) = \frac{1}{\hat{N}\hat{\mu}}\sum_s w_i y_i I\{y_i \le \hat{\xi}_p\}$$

where $\hat{\xi}_p$ is the pth sample quantile, $\hat{\xi}_p = \inf\{y_i \in s|\hat{F}(y_i) \ge p\}$, formally obtained as a solution to the equation

$$\sum_s w_i[I\{y_i \le \xi_p\} - p] = 0, \text{ or } \hat{F}(\xi_p) = p$$

where $\hat{F}(y) = \Sigma w_i(s)I\{y_i \le y\}/\hat{N}$ is an estimate of the finite population cumulative distribution function.

For variance estimation of the Lorenz curve ordinates we use the values of u_i^*

$$u_i^* = \frac{1}{\hat{N}\hat{\mu}}[(y_i - \hat{\xi}_p)I\{y_i \le \hat{\xi}_p\} + p\hat{\xi}_p - y_i\hat{L}(p)]$$

and formula (4). A detailed derivation of the above expression is given in Binder and Kovačević (1995).

3.4. The Gini index

One of the most popular measures of income inequality, the Gini index, is defined as the standardized Lorenz area, i.e., the ratio between the actual and the largest possible Lorenz area (which is 1/2). Hence, it takes values in [0,1]. The finite population form is given as (see Glasser 1962)

$$G = \frac{1}{N}\sum_U (2F_i - 1)y_i/\mu$$

where $F_i = F(y_i) = (1/N)\Sigma_{j\in U}I\{y_j \le y_i\}$ is the value of the finite population distribution function at y_i. The Gini index is sensitive to income values in the middle of the distribution. Its disadvantage is that it is not defined for negative incomes. It can be defined as a solution to the equation

$$U_1(G, \{F_i\}_{i\in U}, \mu) = \sum_U [(2F_i - 1)y_i/\mu - G] = 0$$

where the nuisance parameter $\lambda = \{\{F_i\}_{i\in U}, \mu\}$ is the solution to the system of equations

$$\begin{cases} \left\{ \sum_{j\in U} [I\{y_j \leq y_i\} - F_i] \right\}_{i\in U} = 0 \\ \sum_U (y_i - \mu) = 0 \end{cases} \tag{9}$$

There are N unknown parameters (since one of the F_i's is equal to 1). With N population values of y, we are able to solve the system.

The estimate of the Gini index comes as the solution to the estimated first equation

$$\hat{G} = \frac{1}{\hat{N}\hat{\mu}} \sum_s w_i(2\hat{F}_i - 1)y_i \tag{10}$$

where $\hat{F}_i$ and $\hat{\mu}$ are the solutions to the system of estimated equations (9).

The variance of the Gini index is estimated by expression (4) where the u^* variates are equal to

$$u_i^* = \frac{2}{\hat{N}\hat{\mu}} \left[\hat{A}(y_i)y_i + \hat{B}(y_i) - \frac{\hat{\mu}}{2}(\hat{G}+1) \right] \tag{11}$$

where $\hat{A}(y) = \hat{F}(y) - (\hat{G}+1)/2$ and $\hat{B}(y) = \Sigma_s w_i y_i I\{y_i \geq y\}/\hat{N}$. See the Appendix for details.

3.5. The polarization curve

In order to formalize the concept of polarization, Foster and Wolfson (1992) constructed a curve which shows, for any population percentile, how far its income is from the median. A larger value of the polarization curve ordinate implies a larger 'spread' of the distribution from the middle, indicating a smaller middle class.

For a variable y with a given distribution $F(y)$ Foster and Wolfson (1992) defined the polarization curve ordinate by

$$B(p) = \left| \int_{0.5}^{p} \frac{F^{-1}(q) - m}{m} dq \right|$$

which can also be written in a simpler form

$$B(p) = \left| 0.5 - p + \frac{\mu}{m}[L(p) - L(0.5)] \right| \qquad 0 \leq p \leq 1 \tag{12}$$

where $L(p)$ and μ were previously defined as the Lorenz curve ordinate and the mean, respectively, and m is the median, $m = \inf\{y_i \in U | F(y_i) \geq 0.5\}$. In terms of the EE method, the polarization curve can be defined as a solution to the equation

$$U_1(B(p), m, \xi_p) = \sum_U \left[\frac{y_i}{m}(I\{y_i \leq \xi_p\} - I\{y_i \leq m\}) - B(p) + 0.5 - p \right] = 0 \tag{13}$$

where the nuisance parameters ξ_p and m are solutions to the system of equations

$$\begin{cases} \sum_U [I\{y_i \le \xi_p\} - p] = 0 \qquad 0 \le p \le 1 \\ \sum_U [I\{y_i \le m\} - 0.5] = 0 \end{cases}$$

The estimate of $B(p)$ based on a complex sample is of the form given by (12) with parameters $L(p)$, μ, and m replaced by their estimates.

An outline of the derivation for the polarization curve is presented in the Appendix. Here we give the final form of the u^* variate

$$\begin{aligned} u_i^* = &\frac{1}{\hat{N}\hat{m}}[(\hat{m} - y_i)I\{y_i \le \hat{m}\} + (y_i - \hat{\xi}_p)I\{y_i \le \hat{\xi}_p\} \\ &+ \hat{\xi}_p p - \frac{\hat{m}}{2} - \frac{\hat{B}(p) - 0.5 + p}{\hat{f}(\hat{m})}(0.5 - I\{y_i \le \hat{m}\} + \hat{m}\hat{f}(\hat{m}))\Bigg] \end{aligned} \qquad (14)$$

where $\hat{f}(\hat{m})$ is an estimate of the density function at the median. Estimation of the density function at estimated quantiles was discussed by Binder and Kovačević (1995).

3.6. The polarization index

Using the analogy with the Lorenz curve and the Gini index, Foster and Wolfson (1992) considered the area below the polarization curve as a summary measure of the polarization and named it the polarization index. A perfectly polarized population is divided into equal halves, each having just one of two possible values of income, the minimum or the maximum. If median income in this case is defined as the middle point between minimum and maximum (a slight departure from the usual definition of the median), the polarization curve is then a horizontal line with intercept 1/2 and the point of discontinuity at the 50th percentile, giving the value of the polarization index of 1/2. If there is no polarization, everyone has the same income. The polarization curve is the [0,1] segment of the horizontal axis and the corresponding polarization index is zero. Therefore, the polarization index takes values between 0 and 1/2, and its standardized version is the previous one multiplied by 2.

The standardized polarization index, as introduced in Foster and Wolfson (1992), is

$$P = \left(\frac{T}{2} - G\right)\frac{\mu}{m}$$

where $T = (\mu^U - \mu^L)/\mu$ with μ, μ^L, μ^U equal to the population mean income, the mean income for the population below the median, and the mean income for the population above the median income, respectively. As before, G is the Gini index and m denotes the median income.

Since $\mu = (\mu^U + \mu^L)/2$, the polarization index can be written as

$$\begin{aligned} P &= \frac{1}{m}(\mu - \mu^L - \mu G) \\ &= \frac{1}{Nm}\sum_U y_i[1 - 2I\{y_i \le m\} - G] \end{aligned}$$

or as the solution to the equation

$$U_1(P, m, G) = \sum_U \left\{ \frac{1}{m} y_i [1 - 2I\{y_i \leq m\} - G] - P \right\} = 0 \tag{15}$$

It is estimated as

$$\hat{P} = \frac{1}{\hat{N}\hat{m}} \sum_s w_i y_i [1 - 2I\{y_i \leq \hat{m}\} - \hat{G}]$$

where $\hat{m} = \inf\{y_i \in s | \hat{F}(y_i) \geq 0.5\}$ and $\hat{G}$ is given by (10).

The variance of the polarization index is estimated by (4) with u_i^* defined as

$$\begin{aligned} u_i^* = & \left\{ \frac{2}{\hat{m}} \left[(\hat{m} - y_i) \left(I\{y_i \leq \hat{m}\} - \frac{1}{2} \right) - \left(\hat{A}(y_i) y_i + \hat{B}(y_i) - \frac{\hat{G}+1}{2} \hat{\mu} + \frac{\hat{G}}{2} y_i \right) \right] \right. \\ & \left. + \frac{\hat{P}}{\hat{m}\hat{f}(\hat{m})} \left(I\{y_i \leq \hat{m}\} - \frac{1}{2} \right) - \hat{P} \right\} / \hat{N} \end{aligned}$$

where $\hat{A}(y)$ and $\hat{B}(y)$ are defined in (11), and $\hat{f}(\hat{m})$ is an estimate of the density function at the median. The details of the derivation of u_i^* for the polarization index are given in the Appendix.

4. Illustration

The EE methodology was applied to estimate the standard errors of estimates of several income inequality measures using a file on the earnings of all effective labour force participants aged 18 to 64 in 1991. An effective labour force participant is an individual with annual labour income of at least 5% of the average wage. Data were collected by the Canadian Survey of Consumer Finance (SCF) in April 1992. The SCF is an annual supplement to the monthly Canadian Labour Force Survey, which is based on a stratified, multi-stage sample of households. Approximately 40,000 households provided detailed income information for individuals 15 years of age or older. The data set used for this illustration contained 50,701 individuals, situated in 4,201 clusters (PSU's), allocated to 1,139 strata. Attached to each record is an individual survey weight which is an adjusted sample weight; this allows us to compute the standard errors for estimates of interest.

The point estimates of the income inequality and polarization measures, their standard errors and the corresponding coefficients of variation are presented in Table 1. An analysis of the unweighted data reveals the heavy right skewness and the extreme kurtosis of the data distribution. This may explain the large standard errors of $\hat{C}V$ and $\hat{C}V^2$ which are sensitive to large income values. Also, the polarization index exhibits sensitivity to the data spread to the right. On the other hand, most of the variability of the exponential measure comes from the low income values which are concentrated in a relatively small range. The Gini index is robust to extreme observations and depends primarily on the variability in the middle of the distribution. This may explain the small standard errors of the latter two measures. The Lorenz curve ordinates were found to have smaller coefficients of variation than the polarization curve ordinates. This difference can be partly attributed to the contribution of the estimated density function at the median used in the variance estimation of the polarization curve ordinates and the polarization index, and partly to the sensitivity of these measures to large observations.

As mentioned in the introduction, the main goal of this article is to provide estimators for the sampling variance of different measures of income inequality using the EE method. However, further insight into the properties of these variance estimators can be obtained

Table 1. Canadian SCF 1991: Earnings of all effective labour force participants

Measure	Estimate	Standard error and *CV*%	
Mean	26,297	184.10	0.70
Median*	22,392	226.78	1.01
CV^2	0.7608	0.0429428	5.64
CV	0.8722	0.0246164	2.82
Gini index	0.4122	0.0027141	0.66
Exponential measure	0.4603	0.0011879	0.26
Polarization index	0.1966	0.0021153	1.07

p	Lorenz curve ordinates			Polarization curve ordinates		
	$\hat{L}(p)$	Standard error and *CV*%		$\hat{B}(p)$	Standard error and *CV*%	
0.050	0.0044	0.0000537	1.22	0.2083	0.0014237	0.68
0.100	0.0117	0.0001615	1.38	0.1669	0.0013351	0.79
0.150	0.0226	0.0002809	1.24	0.1297	0.0012150	0.93
0.200	0.0372	0.0004283	1.15	0.0968	0.0010588	1.09
0.250	0.0550	0.0005876	1.06	0.0677	0.0008773	1.29
0.300	0.0772	0.0007777	1.00	0.0438	0.0006725	1.53
0.350	0.1060	0.0009556	0.90	0.0276	0.0004672	1.69
0.400	0.1344	0.0011496	0.85	0.0110	0.0002898	2.63
0.450	0.1740	0.0013513	0.77	0.0075	0.0001106	1.47
0.500	0.2102	0.0015297	0.72	0.0000	0.0000000	–
0.550	0.2601	0.0017332	0.66	0.0086	0.0002297	2.67
0.600	0.3049	0.0018900	0.61	0.0112	0.0004579	4.08
0.650	0.3673	0.0020772	0.56	0.0345	0.0009332	2.70
0.700	0.4187	0.0022240	0.53	0.0449	0.0012915	2.87
0.750	0.4869	0.0023889	0.49	0.0750	0.0018808	2.50
0.800	0.5634	0.0025559	0.45	0.1148	0.0025855	2.25
0.850	0.6423	0.0026927	0.41	0.1575	0.0033394	2.12
0.900	0.7279	0.0028099	0.38	0.2080	0.0042041	2.02
0.950	0.8342	0.0028583	0.34	0.2828	0.0053746	1.90
0.960	0.8608	0.0028575	0.33	0.3041	0.0056776	1.86
0.970	0.8843	0.0027994	0.31	0.3217	0.0059496	1.84
0.980	0.9126	0.0026378	0.28	0.3449	0.0063331	1.83
0.990	0.9449	0.0024158	0.25	0.3729	0.0068331	1.83
0.995	0.9647	0.0022467	0.23	0.3911	0.0071469	1.82
0.997	0.9740	0.0020830	0.21	0.4000	0.0073278	1.83
0.998	0.9794	0.0018379	0.18	0.4054	0.0074922	1.84
0.999	0.9871	0.0011442	0.11	0.4134	0.0078741	1.90
1.000	1.0000	0.0000000	–	0.4276	0.0083695	1.95

*The standard error of the median is also obtained by the EE method (Binder and Kovačević 1995).

through an empirical comparison with some other estimators commonly used. A simulation study designed to compare several resampling methods with the EE method for variance estimation of income inequality measures was conducted at Statistics Canada. The results are reported in Kovačević, Yung, and Pandher (1995). The study focused on income inequality measures that are functions of the quantiles, and did not cover the coefficient of variation, the exponential measure, or the ordinates of the polarization curve. In the following we summarize some of the relevant findings in the simulation study.

Five different methods for variance estimation were compared: jackknife ''delete-one-cluster,'' the grouped balanced half-sample method, repeatedly grouped balanced half-samples, the bootstrap and the Taylor linearization via the EE approach. The underlying population was the microdata file from the Canadian Survey of Consumer Finance in 1988. Ten thousand samples were drawn from the micropopulation using a cluster sample design with the selection probabilities proportional to size. The accuracy and the precision of the considered methods were evaluated by their relative biases and relative stability.

For the Lorenz curve ordinates the EE method showed very small negative relative bias, in the range between -0.4% for the quantile $p = 0.6$ and -5.2% for $p = 0.95$. For the same p-values, the relative bias of the jackknife estimator was 20.49% and 39.02%, respectively. However, the bootstrap estimator exhibited the smallest relative bias at these points, 0.3% and -1.91%. Concerning stability, the EE method along with the bootstrap performed the best.

Similar results were obtained for the polarization index. The relative bias of the EE estimator was computed as 4.2%, whereas for other methods it varied between 2.9% (for the bootstrap) and 95.4% for the jackknife. In terms of stability, the EE and the bootstrap estimator performed similarly and outperformed other methods.

For the variance estimation of the estimate of the Gini index all methods considered performed similarly: all showed a small negative relative bias, in the range of -0.7% and -2.2% and had stability in the range 87.0% to 99.2%. The EE method had a relative bias of -1.5% and a stability of 87.0%.

Results of the study confirm the advantage of using the EE method over most resampling methods considered for variance estimation of the Lorenz curve ordinates, the polarization and the Gini index. The exception was the bootstrap method which performed slightly better. Although, the polarization curve was not studied empirically, its similarity to the Lorenz curve implies that the performance of the EE method should be acceptably good.

5. Summary

Variance estimation of complex statistics such as measures of income inequality can be done by the method of estimating equations. The advantage of this approach is that it can be used under a wide class of sampling designs and does not require intensive computations, which most of the resampling alternatives require. In order to estimate measures of income inequality, one must first compute the u_i variates (given in the second column of Table 2) and then compute their total values after multiplying by the corresponding weights. To estimate the variance of such estimates one needs to compute the u_i^* values (summarized in the third column of Table 2) and substitute them into (4).

Table 2. Summary of the linearized terms for the point estimation (w_i) and variance estimation (u_i)

Measure	u_i	u_i^*
CV^2	$(y_i-\hat{\mu})^2/\hat{N}\hat{\mu}^2)$	$[(y_i/\hat{\mu}-1)^2-(2y_{hci}/\hat{\mu}-1)\hat{CV}^2]/\hat{N}$
CV		$[(y_i/\hat{\mu}-1)^2/\hat{CV}-(2y_i/\hat{\mu}-1)\hat{CV}]/(2\hat{N})$
Gini index	$[2\hat{F}(y_i)-1]\,y_i/(\hat{N}\hat{\mu})$	$2[\hat{A}(y_i)\,y_i+\hat{B}(y_i)-\hat{\mu}(\hat{G}+1)/2]/(\hat{N}\hat{\mu})$ (*)
Exponential measure	$(1/\hat{N})\exp(-y_i/\hat{\mu})$	$[(y_i-\hat{\mu})\hat{J}_{1,\hat{\mu}}/\hat{N}+\exp(-y_i/\hat{\mu})-\hat{EX}]/\hat{N}$ (**)
Polarization index	$y_i[1-2I\{y_i\leq\hat{m}\}-\hat{G}]/(\hat{N}\hat{m})$	$\{\frac{2}{\hat{m}}[(\hat{m}-y_i)(I\{y_i\leq m\}-0.5)$ $-(A(y_i)y_i+B(y_i)-(\hat{G}+1)\hat{\mu}/2+\hat{G}y_i/2)]$ $+\frac{\hat{P}}{\hat{m}\hat{f}(\hat{m})}(I\{y_i\leq\hat{m}\}-0.5)-\hat{P}\Big\}/\hat{N}$
Lorenz curve	$y_iI\{y_i\leq\hat{\xi}_p\}/(\hat{N}\hat{\mu})$	$[(y_i-\hat{\xi}_p)I\{y_i\leq\hat{\xi}_p\}+p\hat{\xi}_p-y_i\hat{L}(p)]/(\hat{N}\hat{\mu})$
Polarization curve	$\{0.5-p+y_i[I\{y_i\leq\hat{\xi}_p\}-I\{y_i\leq\hat{m}\}]/\hat{m}\}/\hat{N}$	$\frac{1}{\hat{N}\hat{m}}\{(\hat{m}-y_i)I\{y_i\leq\hat{m}\}+(y_i-\hat{\xi}_p)I\{y_i\leq\hat{\xi}_p\}+\hat{\xi}_pp-\hat{m}/2$ $-(\hat{B}(p)-0.5+p)[0.5-I\{y_i\leq\hat{m}\}+\hat{m}\hat{f}(\hat{m})]/\hat{f}(\hat{m})\}$

(*) $A(y)=\hat{F}(y)-\frac{\hat{G}+1}{2}$ and $B(y)=\Sigma_s w_jy_jI\{y_j\geq y\}/\hat{N}$
(**) $\hat{J}_{1,\hat{\mu}}=\Sigma w_iy_i\exp(-y_i/\hat{\mu})/\hat{\mu}^2$

The extension of this method to comparisons between domains and comparisons over time are straightforward since we have reformulated the problem to one of estimating variances of linear statistics in the Godambe (1955) class, allowing the use of standard methods applied to these linear statistics. Of course all the complexities arising from having overlapping units over time would have to be accounted for.

6. Appendix

Detailed derivation of the u^* variates

6.1. The exponential measure

The first derivatives of $U_1(EX,\mu)$ and $U_2(\mu)$, given by (8), are

$$J_{1,EX}=-N, J_{1,\mu}=\frac{1}{\mu^2}\sum_U y_i\exp(-y_i/\mu),\text{ and }J_{2,\mu}=-N$$

After substituting into (7) and using the estimates instead of parameters

$$u_i^*=\frac{1}{\hat{N}}[\exp(-y_i/\hat{\mu})-\hat{EX}+(y_i-\hat{\mu})\hat{J}_{1,\hat{\mu}}/\hat{N}]$$

where

$$\hat{J}_{1,\hat{\mu}}=\frac{1}{\hat{\mu}^2}\sum_s w_iy_i\exp(-y_i/\hat{\mu})$$

6.2. The Gini index

The corresponding first derivatives are $J_{1G}=-N$, $J_{1\lambda}=[\{2y_i/\mu\}_{i\in U/i_o},-G/\mu]_{1\times N}$, where i_0 is the label of the maximum y_i, $J_{2G}=0_{N\times 1}$, and $J_{2\lambda}=-NI_{N\times N}$, where I is the identity matrix.

Substituting into equation (7) and replacing parameters with their estimates, we obtain

$$u_i^* = \frac{2}{\hat{N}\hat{\mu}}\left[\hat{A}(y_i)y_i + \hat{B}(y_i) - \frac{\hat{\mu}}{2}(\hat{G}+1)\right]$$

where $\hat{A}(y) = \hat{F}(y) - (\hat{G}+1)/2$ and $\hat{B}(y) = \Sigma_s w_i y_i I\{y_i \geq y\}/\hat{N}$

6.3. Polarization index

We present the derivation of the u^* variates for the polarization index in full detail because it involves approximations that are specific for quantiles and functions of them. The two approximations for the finite population quantiles ξ_p's important for the subsequent development are

$$\begin{aligned}\hat{\xi}_p - \xi_p &\approx \frac{1}{f(\xi_p)}[p - \hat{F}(\xi_p)]\\ &= \sum_U w_i(s)\frac{1}{\hat{N}f(\xi_p)}[p - I\{y_i \leq \xi_p\}] \qquad \text{(A1)}\end{aligned}$$

where $f(\xi_p)$ is the value of the density function at ξ_p. This is an extension of the Bahadur representation (Bahadur 1966) for a quantile to the finite population case.

Continuing, let $\mu(a) = (1/N)\Sigma_U y_i I\{y_i \leq a\}$. Note that the previously introduced μ^L, the mean of the lower half of the population, is equal to $2\mu(m)$. Also, the Lorenz curve ordinate at p is equal to $\mu(\xi_p)/\mu$. The following approximation holds for quantiles

$$\begin{aligned}\mu(\hat{\xi}_p) - \mu(\xi_p) &= \frac{1}{N}\sum_U y_i(I\{y_i \leq \hat{\xi}_p\} - I\{y_i \leq \xi_p\})\\ &\approx \xi_p f(\xi_p)(\hat{\xi}_p - \xi_p)\\ &\approx \sum_U w_i(s)\xi_p(p - I\{y_i \leq \xi_p\})/\hat{N}, \quad \text{as } \hat{\xi}_p \to \xi_p \qquad \text{(A2)}\end{aligned}$$

The approximation (A2) appears in a more general form in Binder and Kovačević (1995).

The estimate of the equation (15) can be expressed as

$$\begin{aligned}0 &= \hat{U}_1(\hat{P}, \hat{m}, \hat{G})\\ &= \sum_U w_i(s)\left[\frac{y_i}{\hat{m}}(1 - 2I\{y_i \leq \hat{m}\} - \hat{G}) - \hat{P}\right]\\ &\approx N(P - \hat{P}) + \sum_U\left[\frac{y_i}{\hat{m}}(1 - 2I\{y_i \leq \hat{m}\} - \hat{G}) - \frac{y_i}{m}(1 - 2I\{y_i \leq m\} - G)\right]\\ &\quad + \sum_U w_i(s)\left[\frac{y_i}{m}(1 - 2I\{y_i \leq m\} - G) - P\right]\end{aligned}$$

Approximating the function $y_i/\hat{m}$ around the median m by its first order Taylor expansion $y_i/m - y_i(\hat{m} - m)/m^2$, substituting into the expression above, and simplifying, we have the following

$$\begin{aligned}\hat{P} - P &\approx -\frac{1}{m}(\hat{\mu}^L - \mu^L) - \frac{\mu}{m}(\hat{G} - G) - \frac{\hat{m} - m}{m^2}(\mu - \hat{\mu}^L - \mu\hat{G})\\ &\quad + \sum_U w_i(s)\left[\frac{y_i}{m}(1 - 2I\{y_i \leq m\} - G) - P\right]/N \qquad \text{(A3)}\end{aligned}$$

Replacing $\hat{\mu}^L$ with $(\hat{\mu}^L - \mu^L) + \mu^L$ and $\hat{G}$ with $(\hat{G} - G) + G$ in the third term of (A3) and then removing the mixed products from approximation (A3) as higher order terms, we arrive at the following linearization

$$\begin{aligned}\hat{P} - P &\approx -\frac{1}{m}(\hat{\mu}^L - \mu^L) - \frac{\mu}{m}(\hat{G} - G) - \frac{P}{m}(\hat{m} - m)\\ &\quad + \sum_U w_i(s)\left[\frac{y_i}{m}(1 - 2I\{y_i \leq m\} - G) - P\right]/N\end{aligned}$$

Finally, we use (A1), (A2), and (11) to approximate the differences $\hat{m} - m$, $\hat{\mu}^L - \mu^L$, and $\hat{G} - G$, respectively

$$\begin{aligned}\hat{P} - P &\approx \sum_U w_i(s)u^*(y_i, P, m, G)\\ &= \sum_U w_i(s)\left\{\frac{2}{m}\left[\left(m - y_i + \frac{P}{2f(m)}\right)\left(I\{y_i \leq m\} - \frac{1}{2}\right)\right.\right.\\ &\quad \left.\left. - \left(A(y_i)y_i + B(y_i) - \frac{G+1}{2}\mu + \frac{G}{2}y_i\right)\right] - P\right\}/N\end{aligned}$$

This expression provides the final form of the u_i^* variate as given by (16).

6.4. The polarization curve

The polarization curve ordinates are given by expression (12). To estimate their variance we proceed as in the case of the polarization index, starting with the decomposition of the estimate of (13)

$$\begin{aligned}0 &= \sum_U w_i(s)\left[\frac{y_i}{\hat{m}}(I\{y_i \leq \hat{\xi}_p\} - I\{y_i \leq \hat{m}\}) - \hat{B}(p) + 0.5 - p\right]\\ &\approx N[B(p) - \hat{B}(p)] + \sum_U \left[\frac{y_i}{\hat{m}}(I\{y_i \leq \hat{\xi}_p\} - I\{y_i \leq \hat{m}\}) - \frac{y_i}{m}(I\{y_i \leq \xi_p\} - I\{y_i \leq m\})\right]\\ &\quad + \sum_U w_i(s)\left[\frac{y_i}{m}(I\{y_i \leq \xi_p\} - I\{y_i \leq m\}) - B(p) + 0.5 - p\right]\end{aligned}$$

Approximating the function $y_i/\hat{m}$ around the median m by its first order Taylor expansion $y_i/m - y_i(\hat{m} - m)/m^2$, and using similar substitutions as in the case of the polarization index, we obtain the difference $\hat{B}(p) - B(p)$ as

$$\begin{aligned}\hat{B}(p) - B(p) &\approx \sum_U w_i(s)u^*(y_i, B(p), m, \xi_p)\\ &= \sum_U w_i(s)\frac{1}{Nm}\left[(m - y_i)I\{y_i \leq m\} + (y_i - \xi_p)I\{y_i \leq \xi_p\}\right.\\ &\quad \left. + \xi_p p - \frac{m}{2} - \frac{B(p) - 0.5 + p}{f(m)}(0.5 - I\{y_i \leq m\} + mf(m))\right]\end{aligned}$$

Finally, to estimate the variance of $\hat{B}(p)$ we use formula (4) for the variance of total and u_i^* variates given by (14).

7. References

Bahadur, R.R. (1966). A Note on Quantiles in Large Samples. Annals of Mathematical Statistics, 37, 577–580.

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51, 279–292.

Binder, D.A. (1991). Use of Estimating Functions for Interval Estimation from Complex Surveys. Proceedings of the Survey Research Methods Section, American Statistical Association, 34–42.

Binder, D.A. (1992). Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equation Approach. Proceedings of the Workshop on Statistical Issues in Public Policy Analysis, Carleton University, Ottawa, May 29.

Binder, D.A. and Patak, Z. (1994). Use of Estimating Functions for Interval Estimation from Complex Surveys. Journal of the American Statistical Association, 89, 1035–1043.

Binder, D.A. and Kovačević, M.S. (1995). Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equation Approach. Survey Methodology, 21, 137–145.

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

Francisco, C.A. and Fuller, W.A. (1991). Quantile Estimation with a Complex Survey Design. Annals of Statistics, 19, 454–469.

Glasser, G.J. (1962). Variance Formulas for the Mean Difference and Coefficient of Concentration. Journal of the American Statistical Association, 57, 648–654.

Godambe, V.P. (1955). A Unified Theory of Sampling from Finite Populations. Journal of the Royal Statistical Society, Ser. B, 17, 269–278.

Godambe, V.P. and Kale, B.K. (1991). Estimating Functions: An Overview. In V.P. Godambe, (ed.) Estimating Functions, London: Oxford Statistical Science Series, 7.

Godambe, V.P. and Thompson, M.E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review, 54, 127–138.

Kovačević, M.S., Yung, W., and Pandher, G.S. (1995). Estimating the Sampling Variances of Measures of Income Inequality and Polarization – An Empirical Study. Statistics Canada, Methodology Branch Working Paper, HSMD-95-007E.

Rao, J.N.K. (1979). On Deriving Mean Square Errors and Their Non-negative Unbiased Estimators in Finite Population Sampling. Journal of the Indian Statistical Association, 17, 125–136.

Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. New York: John Wiley.

Wolfson, M.C. (1986). Stasis Amid Change – Income Inequality in Canada 1965–1983. Review of Income and Wealth, 32, 337–369.

Wolfson, M.C. (1994). When Inequalities Diverge. American Economic Review, 84, 353–358.

Received September 1994
Revised August 1996