

INFERENCE ABOUT DOMAINS DEPENDENT ON NUISANCE PARAMETERS

Milorad S. Kovacevic and Lenka Mach¹

ABSTRACT

A domain of study is a sub-population for which statistical inference is requested. Sometimes the domain definition depends on the unknown parameters which are not *per se* of interest for inference but need to be estimated from the same sample. Typical examples include the domain of low income households, the domain of underweight children, and a population after discarding the outliers using the 'three-sigma' rule. The definition of these domains is dependent on the nuisance population parameters: the median income, the average weight for a child in a given age group, and the population mean and the standard deviation, respectively. The domain parameters of interest, such as means, totals, and ratios, are then non-smooth functions dependent on the nuisance parameters. They are usually estimated using the 'plug-in' estimators after substituting the nuisance parameters by their estimates. Under some regularity conditions for a complex sampling design we show the consistency of these estimators and derive their standard errors using the estimating equations approach. We support our theoretical results by an illustration based on data from Statistics Canada Survey of Income and Labour Dynamics (SLID).

KEY WORDS: Complex sampling design; Distribution estimation; Estimating equations; Non-smooth functions; SLID; Variance estimation.

RÉSUMÉ

Le domaine d'estimation d'une enquête est une sous-population pour laquelle on souhaite faire de l'inférence statistique. Parfois, la définition des domaines dépend de paramètres inconnus qui ne sont pas d'intérêt en soi pour l'inférence mais qui doivent être estimés à partir du même échantillon. Des exemples typiques peuvent être un domaine des ménages à faible revenu, un domaine des enfants de petit poids ou encore, une population dont les valeurs aberrantes ont été éliminées à l'aide de la règle « des trois sigmas ». Pour ces exemples, la définition de ces domaines dépend respectivement des paramètres dérangeants de la population suivants : la médiane du revenu, le poids moyen des enfants d'un groupe d'âge donné, la moyenne et l'écart-type d'une population. Les paramètres d'intérêt du domaine comme les moyennes, totaux et ratios sont des fonctions non-continues qui dépendent des paramètres dérangeants. Ils sont habituellement estimés en substituant les paramètres dérangeants par leurs estimateurs à l'intérieur de ces fonctions non-continues. Pour un plan d'échantillonnage complexe, sous certaines conditions de régularité, nous démontrons la convergence de ces estimateurs et dérivons leur écart-type en utilisant une approche d'équations d'estimation. Nous accompagnons nos résultats théoriques d'un exemple basé sur les données de l'Enquête sur la Dynamique du Travail et du Revenu (EDTR) de Statistique Canada.

MOTS CLÉS: EDTR; estimation de distribution; estimation de la variance; équations d'estimation; fonction non-continue; plan d'échantillonnage complexe.

1. INTRODUCTION

A domain of study is a sub-population for which statistical inference (estimation and testing of hypotheses) is requested. The standard approach to domain estimation when the domain is well defined consists of redefining the variable of interest to be 0 outside the domain. Totals and other domain parameters are then consistently estimated by the corresponding population estimates for the redefined variable (see for example, Sarndal, Swensson and Wretman, 1992). The domain sample size is usually not fixed. Recently, Casady, Dorfman and Wang (1998) addressed the

problem of the confidence interval estimation for domain parameters when the domain sample size is not fixed. They showed that the traditional normal confidence intervals can be seriously underestimated and suggested conditional *t*-based intervals with better coverage properties.

Sometimes the definition of a domain depends on the unknown parameters. These nuisance parameters need to be eliminated in order to proceed with the inference about the domain parameters of interest (say domain means, totals, ratios). When the nuisance parameters are substituted with their estimates, the domain is not fixed anymore which further implies that the estimates of

¹

Milorad Kovacevic, Social Survey Methods Division, Statistics Canada, Ottawa, ON K1A 0T6 and Lenka Mach, Business Survey Methods Division, Statistics Canada, Ottawa, ON K1A 0T6, Canada, Milorad.Kovacevic@statcan.ca.

domain parameters and of their variances should account for the additional variability of the estimated nuisance parameters as well. Some related work was done for the variance estimation of gross flows when the transitional states were based on the estimated distribution of the survey data (see Caron and Chambaz, 1998, and Kovacevic, 2000.)

Typical examples of such domains include: (i) A sub-population of the underweight children where a child is underweight if his body weight is less than 2/3 of the average body weight in his age group; (ii) the 'low income domain' of the population where the low income person (or family) is defined as one with the income below half the median income for that type of family; (iii) a third example is a 'trimmed' population with values considered only within the boundaries $\mu \pm k\sigma$.

In all three examples inference about the respective domains has to be done in the presence of nuisance parameters that determine the domains.

In the general statistical inference, different approaches are recommended for the elimination of the nuisance parameters (Basu, 1977). These include conditioning, marginalization, substitution, etc. The approach taken in this paper is a method of substitution where the unknown nuisance parameter is replaced by its estimate. The variance estimate of the estimated mean of a domain of this type is obtained by the estimating equations approach.

The outline of the paper is as follows. The notation to be used throughout the paper, the problem formulation and the examples are given in Section 2. Section 3 contains a brief summary of the estimating equations approach and its application to the domain estimation. Some problems related to some specific domain-defining parameters are discussed too. Section 4 addresses the variance estimation for the three possible cases of the nuisance parameter estimation. A numerical illustration based on the Canadian Survey of Labour Income and Dynamics (SLID) is given in Section 5 along with some concluding remarks.

2. PROBLEM FORMULATION, NOTATION AND EXAMPLES

Let $D(\lambda_x)$ denote a domain of interest where λ_x is a finite population parameter that determines the domain. The subscript x indicates that the domain is defined according to the distribution of a variable x . Without loss of generality we assume that the parameter of interest is the domain mean of a variable y , $\mu_y^D = \mu_y(\lambda_x)$. Note that the variables y and x are not necessarily different, since we may be interested in the average

weight of the underweight children, or a mean income of the low-income families, etc.

In the examples given in the introduction, the parameter λ_x represents: (i) two thirds of the average body weight μ_x for a child in a given age group, (ii) a half of the median income m_x for the population, and in (iii) it is $\mu_x \pm k\sigma_x$ where μ_x and σ_x are the population mean and standard deviation, respectively.

The parameter of interest is the domain mean of a variable y defined as

$$\mu_y^D = \mu_y(\lambda_x) = \bar{Y}_D = \sum_{i \in P} y_i I\{i \in D(\lambda_x)\} / \sum_{i \in P} I\{i \in D(\lambda_x)\} \quad (1)$$

where $I\{a\}$ indicates a true event. For the sake of simplicity, let $I\{i \in D(\lambda_x)\} = I\{x_i \leq \lambda_x\} = I_i\{\lambda_x\}$.

In order to estimate μ_y^D , a sample of size n is drawn from the finite population of size N according to a sampling design $p(s)$, so that the sample inclusion probabilities are $\pi_i = \sum_{s \in S: s \ni i} p(s)$, $i=1, \dots, N$. The corresponding sampling weights are defined as $w_i = 1/\pi_i$, if $i \in s$ and 0 otherwise. In a typical survey, the sampling design includes stratification and selection in two stages, i.e. we have a sample of primary sampling units (PSU) of size m_h from the h -th stratum ($h=1, \dots, L$) and a sample of n_{hi} ultimate units from the i th PSU, yielding a sample of total size $n = \sum_{h=1}^L \sum_{i=1}^{m_h} n_{hi}$.

If λ_x is known, the domain $D(\lambda_x)$ is a 'regular' domain with the fixed size that may not be known. In such a case μ_y^D is a usual domain mean. An example of such domain is a domain of four member families. The domain mean μ_y^D is then estimated by:

$$\hat{\mu}_y(\lambda_x) = \frac{\sum_{i \in s} w_i y_i I_i\{\lambda_x\}}{\sum_{i \in s} w_i I_i\{\lambda_x\}} = \frac{\sum_{i \in s} w_i y_i I_i\{\lambda_x\} / \hat{N}}{\sum_{i \in s} w_i I_i\{\lambda_x\} / \hat{N}} \quad (2)$$

$$= \frac{\sum_{i \in s} w_i^* y_i I_i\{\lambda_x\}}{\hat{F}_x(\lambda_x)}$$

Here w_i^* denotes the standardized sampling weight $w_i^* = w_i / \hat{N}$ so that $\sum_{i \in s} w_i^* = 1$. The cumulative distribution function for variable x estimated at λ_x is denoted by $\hat{F}_x(\lambda_x)$. Estimator (2) has the regular properties of an estimator of the domain mean: consistency ($P_d\{|\hat{\mu}_y(\lambda_x) - \mu_y(\lambda_x)| < \varepsilon\} \rightarrow 1$, as $n_D, N_D \rightarrow \infty$, $n_D/N_D \rightarrow 1$), and the asymptotic normality $\frac{\hat{\mu}_y(\lambda_x) - \mu_y(\lambda_x)}{\sqrt{\hat{V}(\hat{\mu}_y(\lambda_x))}} \rightarrow N(0,1)$,

$n_D, N_D \rightarrow \infty, n_D/N_D \rightarrow 1$. Here n_D and N_D are the sample and the domain sizes, respectively. Note that the first convergence is in probability defined by the sampling design, and the second is in law.

In this paper, our interest is in domains determined by the unknown λ_x that has to be estimated. Generally, we assume that $\hat{\lambda}_x$ has the properties of consistency and asymptotic normality. An estimator of the domain mean is obtained from (2) by replacing the unknown λ_x by its estimate. Note that λ_x can be estimated from either the same sample or from an independent one. We will distinguish these cases in the section on variance estimation.

In the first example, λ_x is estimated by $\hat{\lambda}_x = (2/3) \sum_s w_i^* x_i$.

In the example of the low income domain, the parameter λ_x is defined as half the median, $\lambda_x = m_x/2 = F_x^{-1}(0.5)/2$.

It is estimated by $\hat{\lambda}_x = \hat{m}_x/2 = \inf\{x \in S \mid \hat{F}_x(x) \geq 1/2\}/2$.

In the third example where a domain of interest is defined as a population 'trimmed' by the k -sigma rule, $D(\mu_x, \sigma_x) = \{i \in P \mid \mu_x - k\sigma_x < x_i < \mu_x + k\sigma_x\}$. The domain mean in this case takes the form

$$\mu_y(\mu_x, \sigma_x) = \frac{\sum_{i \in P} y_i [I_i\{\mu_x + k\sigma_x\} - I_i\{\mu_x - k\sigma_x\}]/N}{F_x(\mu_x + k\sigma_x) - F_x(\mu_x - k\sigma_x)}$$

Population parameters μ_x and σ_x are estimated by $\hat{\mu}_x = \sum_s w_i^* x_i$ and $\hat{\sigma}_x^2 = \sum_s w_i^* (x_i - \hat{\mu}_x)^2$.

In a broader context this problem is similar to inference using statistics which involve substituting estimates for nuisance parameters as explored by Randles (1982) and Pierce (1982). The main differences with their work are that the statistic of our interest $\hat{\mu}_y(\hat{\lambda})$ is not a differentiable function of the nuisance parameters and that the data are obtained using a complex sampling design. Some results of Shao and Rao (1993), Binder and Kovacevic (1995) and Kovacevic and Binder (1997) on standard error estimation for low income proportions are directly applicable to the problems addressed in this paper.

3. ESTIMATING EQUATIONS APPROACH

In this section we review some basics of the estimating equations (EE) approach to estimation of the finite population parameters and standard error estimation, and apply them to the domain estimation. Parameters of the finite population are defined as functions of the values taken by the population units. They can be expressed as solutions to the system of estimating equations of a

general form

$$U(\theta) = \sum_{i=1}^N u(z_i; \theta)/N = 0. \quad (3)$$

Here we assume that $\theta = (\mu, \lambda)$ is a multidimensional parameter. However, we are interested only in μ , treating the other components as nuisance parameters. The estimating function $u(\cdot)$ is assumed known. The vector $z_i = (x_i, y_i, \dots)$ contains the values of the parameter relevant variables for the i th population unit ($i = 1, \dots, N$).

The estimates of the unknown parameters are obtained in a two-step estimating procedure, first by estimating system (3) as a set of population means, and then by solving the estimated equations:

$$\hat{U}(\hat{\theta}) = \sum_{i \in S} w_i^* u(z_i; \hat{\theta}) = 0 \quad (4)$$

For the purpose of domain estimation we are extending a formulation of the estimating equations approach for complex sampling designs as given in Binder and Patak (1994). The domain mean μ_y^D and the parameters λ_x are defined as roots of system (3) and their estimates are the roots of the system of estimated EEs (4):

$$\begin{cases} \hat{U}_1(\hat{\mu}_y^D, \hat{\lambda}_x) = \sum_{i \in S_1} w_i^* u_1(x_i, y_i; \hat{\mu}_y^D, \hat{\lambda}_x) = 0 \\ \hat{U}_2(\hat{\lambda}_x) = \sum_{i \in S_2} w_i^* u_2(x_i; \hat{\lambda}_x) = 0 \end{cases} \quad (5)$$

The estimating function $u_1(\cdot)$ for the domain mean estimation has the general form

$$u_1(x, y; \hat{\mu}_y^D, \hat{\lambda}_x) = (y_i - \hat{\mu}_y^D) I_i(\hat{\lambda}_x). \quad (6)$$

Obviously, $u_1(x, y; \hat{\mu}_y^D, \hat{\lambda}_x)$ is not differentiable in $\hat{\lambda}_x$. Functions $u_2(\cdot)$ have different forms depending on λ_x . In our examples they are: $u_2(x; \hat{\lambda}_x) = x_i - \hat{\lambda}_x/\alpha$ with $\alpha = 2/3$; $u_2(x; \hat{\lambda}_x) = I\{x_i \leq \hat{\lambda}_x/\alpha\} - 1/2$ with $\alpha = 1/2$.

Remark 1: In the example of a 'trimmed' population, the function $u_1(\cdot)$ in (6) takes the form

$$u_1(x, y; \hat{\mu}_y^D, \hat{\mu}_x, \hat{\sigma}_x) = (y_i - \hat{\mu}_y^D) [I_i\{\hat{\mu}_x + k\hat{\sigma}_x\} - I_i\{\hat{\mu}_x - k\hat{\sigma}_x\}]$$

and for $\lambda_x = \mu_x \pm k\sigma_x$ there are two functions

$$u_2(x_i; \hat{\sigma}_x, \hat{\mu}_x) = (x_i - \hat{\mu}_x)^2 - \hat{\sigma}_x^2,$$

$$u_3(x_i; \hat{\mu}_x) = x_i - \hat{\mu}_x.$$

In (5), samples s_1 and s_2 can be the same or can be two independent samples.

3.1. Approximation of $\hat{\mu}_y^D - \mu_y^D$

System (4) can be decomposed in the following way (see Binder and Patak (1994) for a single parameter, and Kovacevic and Binder (1997) for a multi dimensional parameter):

$$\begin{aligned} \mathbf{0} &= \hat{U}(\hat{\theta}) = \hat{U}(\hat{\mu}_y^D, \hat{\lambda}_x) \\ &= [U(\hat{\mu}_y^D, \hat{\lambda}_x) - U(\mu_y^D, \lambda_x)] + [U(\mu_y^D, \hat{\lambda}_x) - U(\mu_y^D, \lambda_x)] \\ &\quad + \hat{U}(\mu_y^D, \lambda_x) + \mathbf{R} \end{aligned} \quad (7)$$

The reminder part is

$$\begin{aligned} \mathbf{R} &= \hat{U}(\hat{\theta}) - U(\hat{\theta}) - [U(\hat{\theta}) - U(\theta)] \\ &= \sum_{i=1}^N (w_i^*(s) - 1/N) [u(z_i; \hat{\theta}) - u(z_i; \theta)] \end{aligned} \quad (8)$$

where $w_i^*(s) = w_i^*$, if $i \in s$, and $w_i^*(s) = 0$, if $i \notin s$. It can be proved that the reminder is asymptotically equivalent to the product of $o(|\hat{\mu}_y^D - \mu_y^D|)$ and $o(|\hat{\lambda}_x - \lambda_x|)$, as $\hat{\mu}_y^D - \mu_y^D$ and $\hat{\lambda}_x - \lambda_x$.

Decomposition (7) applied to equations (5) gives the following general expression:

$$\mathbf{0} = \begin{bmatrix} J_{1\mu_y^D} & J_{1\lambda_x} \\ \mathbf{0} & J_{2\lambda_x} \end{bmatrix} \begin{pmatrix} \hat{\mu}_y^D - \mu_y^D \\ \hat{\lambda}_x - \lambda_x \end{pmatrix} + \begin{pmatrix} \hat{U}_1(\mu_y^D, \lambda_x) \\ \hat{U}_2(\lambda_x) \end{pmatrix} + \mathbf{R} \quad (9)$$

where

$$J_{1\mu_y^D} = \frac{U_1(\hat{\mu}_y^D, \hat{\lambda}_x) - U_1(\mu_y^D, \hat{\lambda}_x)}{\hat{\mu}_y^D - \mu_y^D}, \quad (10)$$

$$J_{1\lambda_x} = \frac{U_1(\mu_y^D, \hat{\lambda}_x) - U_1(\mu_y^D, \lambda_x)}{\hat{\lambda}_x - \lambda_x}, \quad (11)$$

$$\text{and } J_{2\lambda_x} = \frac{U_2(\hat{\lambda}_x) - U_2(\lambda_x)}{\hat{\lambda}_x - \lambda_x}. \quad (12)$$

From (9) we obtain a first order approximation

$$\hat{\mu}_y^D - \mu_y^D \approx -J_{1\mu_y^D}^{-1} \hat{U}_1(\mu_y^D, \lambda_x) + J_{1\mu_y^D}^{-1} J_{1\lambda_x} J_{2\lambda_x}^{-1} \hat{U}_2(\lambda_x) \quad (13)$$

which will be used in Section 5 for variance estimation of $\hat{\mu}_y^D$. It also proves the consistency of $\hat{\mu}_y^D = \hat{\mu}_y^D(\hat{\lambda}_x)$ in a sense that $\hat{\mu}_y^D - \mu_y^D \rightarrow 0$ as $n_D, N_D \rightarrow \infty$, $n_D/N_D \rightarrow 1$. Note that $\hat{\mu}_y^D = \hat{\mu}_y(\hat{\lambda}_x)$ and $\mu_y^D = \mu_y(\lambda_x)$.

Remark 2: For the third example, a ‘trimmed’ population, the \mathbf{J} matrix is 3x3 and contains additional J factors. They can be derived along the lines presented for the first two examples.

3.2 Approximation of “J” factors

By replacing $U_1(\cdot)$ in (10) and (11) with its values

defined by (3) and (6) we obtain

$$\begin{aligned} J_{1\mu_y^D} &= \frac{\sum_{i=1}^N (y_i - \hat{\mu}_y^D) I(\hat{\lambda}_x) / N - \sum_{i=1}^N (y_i - \mu_y^D) I(\hat{\lambda}_x) / N}{\hat{\mu}_y^D - \mu_y^D} \\ &\approx -F_x(\hat{\lambda}_x) + o(\hat{\lambda}_x - \lambda_x) \end{aligned} \quad (14)$$

Here we use approximation $F_x(\hat{\lambda}_x) \approx F_x(\lambda_x) + o(\hat{\lambda}_x - \lambda_x)$. The proof is given in the Appendix 1.

Similarly,

$$J_{1\lambda_x} = \frac{\sum_{i=1}^N [(y_i - \mu_y^D) I_i(\hat{\lambda}_x) - (y_i - \mu_y^D) I_i(\lambda_x)] / N}{\hat{\lambda}_x - \lambda_x} \quad (15)$$

In order to simplify equation (15) we use the following approximation:

$$\begin{aligned} \sum_{i=1}^N (y_i - \mu_y^D) [I_i(\hat{\lambda}_x) - I_i(\lambda_x)] / N \\ \approx (\hat{\lambda}_x - \lambda_x) f_x(\lambda_x) [E(y|x = \lambda_x) - \mu_y^D]. \end{aligned} \quad (16)$$

The proof is given in the Appendix 1. In expression (16) $f_x(\cdot)$ denotes the marginal density function for random variable x . The term $E(y|x = \lambda_x)$ is the conditional expectation of y given that $x = \lambda_x$. Hence

$$J_{1\lambda_x} \approx f_x(\lambda_x) [E(y|x = \lambda_x) - \mu_y^D]$$

To approximate $J_{2\lambda_x}$ we replace $U_2(\cdot)$ from (12) by its values defined after (6). Obviously $J_{2\lambda_x}$ will take different forms depending on λ_x . In our examples they are:

For $\lambda_x = \alpha \mu_x$:

$$J_{2\lambda_x} = \frac{\sum_{i=1}^N [(x_i - \hat{\lambda}_x/\alpha) - (x_i - \lambda_x/\alpha)] / N}{\hat{\lambda}_x - \lambda_x} = -\frac{1}{\alpha} \text{ with } \alpha=2/3;$$

for $\lambda_x = \alpha m_x$ (m_x is median)

$$\begin{aligned} J_{2\lambda_x} &= \frac{\sum_{i=1}^N [(I(x_i \leq \hat{\lambda}_x/\alpha) - 1/2) - (I(x_i \leq \lambda_x/\alpha) - 1/2)] / N}{\hat{\lambda}_x - \lambda_x} \\ &= \frac{F_x(\hat{\lambda}_x/\alpha) - F_x(\lambda_x/\alpha)}{\hat{\lambda}_x - \lambda_x} \approx \frac{1}{\alpha} f_x(\lambda_x/\alpha) \end{aligned}$$

with $\alpha=1/2$.

4. VARIANCE ESTIMATION

A linearized expression of the difference $\hat{\mu}_y^D - \mu_y^D$ after

substitution of $J_{1\mu_y^D}$ and $J_{1\lambda_x}$ into (13) is:

$$\hat{\mu}_y^D - \mu_y^D \approx \frac{1}{F_x(\lambda_x)} \sum_{i \in s_1} w_i^* (y_i - \mu_y^D) I_i(\lambda_x) + \frac{f_x(\lambda_x) [E(y|x=\lambda_x) - \mu_y^D]}{F_x(\lambda_x)} J_{2\lambda_x}^{-1} \sum_{i \in s_2} w_i^* u_2(x, \lambda_x). \quad (17)$$

The variance of the domain mean $\hat{\mu}_y^D$ is then estimated as the variance of an estimated total:

$$V\hat{a}r(\hat{\mu}_y^D) \approx V\hat{a}r \left(\frac{1}{F_x(\lambda_x)} \sum_{i \in s_1} w_i^* (y_i - \mu_y^D) I_i(\lambda_x) + C_{x,y}(\mu_y^D, \lambda_x) \sum_{i \in s_2} w_i^* u_2(x, \lambda_x) \right) \quad (18)$$

where $C_{x,y}(\mu_y^D, \lambda_x) = \frac{f_x(\lambda_x) [E(y|x=\lambda_x) - \mu_y^D]}{F_x(\lambda_x)} J_{2\lambda_x}^{-1}$.

The variance estimation is done according to the sampling design and the variance estimate is evaluated at $\mu_y^D = \hat{\mu}_y^D$, $F_x(\lambda_x) = \hat{F}_x(\hat{\lambda}_x)$, $f_x(\lambda_x) = \hat{f}_x(\hat{\lambda}_x)$ and $C_{x,y}(\mu_y^D, \lambda_x) = \hat{C}_{x,y}(\hat{\mu}_y^D, \hat{\lambda}_x)$. Note that for variance estimation (18) we may use any convenient method including the Taylor linearization or some of the resampling methods.

We may distinguish three cases regarding λ_x .

Case 1: Here we assume that λ_x is known. It implies that the second term in (17) is equal to zero so that the variance (18) reduces to

$$V\hat{a}r(\hat{\mu}_y^D) \approx V\hat{a}r \left(\frac{1}{F_x(\lambda_x)} \sum_{i \in s} w_i^* (y_i - \mu_y^D) I_i(\lambda_x) \right)$$

This is the variance estimator that we use when ignoring the variability of the estimated λ_x .

Case 2: λ_x is unknown and it is estimated from the same sample. In this case $s_1 = s_2$ and the two sums in (18) can be jointly written so that the variance (18) is:

$$V\hat{a}r(\hat{\mu}_y^D) \approx V\hat{a}r \left(\sum_{i \in s} w_i^* u^*(x_i, y_i; \mu_y^D, \lambda_x) \right)$$

where

$$u^*(x, y; \mu_y^D, \lambda_x) = \frac{1}{F_x(\lambda_x)} \left[(y - \mu_y^D) I(\lambda_x) + f_x(\lambda_x) [E(y|x=\lambda_x) - \mu_y^D] J_{2\lambda_x}^{-1} u_2(x, \lambda_x) \right]$$

This case is the most frequent in practice.

Case 3: λ_x is unknown but it is estimated from an

independent sample. Since $s_1 \perp s_2$, variance (18) can be expressed as a sum of two variance terms:

$$V\hat{a}r(\hat{\mu}_y^D) \approx V\hat{a}r \left(\frac{1}{F_x(\lambda_x)} \sum_{i \in s_1} w_i^* (y_i - \mu_y^D) I_i(\lambda_x) \right) + C^2 V\hat{a}r \left(\sum_{i \in s_2} w_i^* u_2(x, \lambda_x) \right)$$

where $C = C_{x,y}(\mu_y^D, \lambda_x)$. Obviously the variance of the domain mean is increased only by the portion of the variance of $\hat{\lambda}_x$.

4.1 Estimation of $C_{x,y}(\mu_y^D, \lambda_x)$

In order to estimate the factor $C = C_{x,y}(\mu_y^D, \lambda_x)$ we need to estimate the density function at λ_x (or at $\alpha\lambda_x$) and the conditional mean $\mu_{y|x=\lambda_x} = E(y|x=\lambda_x)$.

For estimation of the density function f_x at $x=\xi$, Binder and Kovacevic (1995) suggested the following estimate

$$\hat{f}_x(\xi) = \frac{\hat{F}_x(\xi + h/2) - \hat{F}_x(\xi - h/2)}{h} \quad (19)$$

for some suitably small h .

Similarly, the conditional mean can be estimated as

$$\hat{\mu}_{y|x=\xi} = \frac{\sum_{i \in s} w_i^* y_i I(\xi - h/2 \leq x_i \leq \xi + h/2)}{\hat{F}_x(\xi + h/2) - \hat{F}_x(\xi - h/2)} \quad (20)$$

for a suitably small h .

5. NUMERICAL EXAMPLE

Here we present an example based on the data from the Canadian Survey of Labour and Income Dynamics (SLID) for 1993. The domain of interest is 'the low income families'. The domain defining parameter λ_x is the half of the median family income (after tax). The parameter of interest, μ_y^D , is the average family income (after tax). This is an example where the variable of interest y and the domain defining variable x are the same. For the variance estimation we used the Taylor linearization method as implemented in SUDAAN.

The half of the median family income is estimated from the same sample as $\hat{\lambda}_x = \$19,479$ (with s.e. = \$185). The average income (after tax) for a low income family is estimated as $\hat{\mu}_y^D = \$13,210$. If the variability of $\hat{\lambda}_x$ is ignored (Case 1), the standard error of $\hat{\mu}_y^D$ is obtained as \$131. When accounting properly for variability of $\hat{\lambda}_x$

(Case 2), the standard error is \$142. Assuming that $\hat{\lambda}_x$ is estimated from an independent sample (Case 3) increases the standard error of $\hat{\mu}_y^D$ to \$148. For estimation of the density function and the conditional mean we tried several values of h (see formulae (19) and (20)). The presented results are based on $h=2 \times 185$. Considering the variance estimation in Case 2 as an appropriate one we conclude that when ignoring the variability of the estimated $\hat{\lambda}_x$ by treating it as a fixed known parameter (Case 1) we underestimate the variance by 8.4%. If we treat $\hat{\lambda}_x$ as if it is estimated from an independent sample when it is estimated from the same sample as $\hat{\mu}_y^D$, thus ignoring the covariance between $\hat{\lambda}_x$ and $\hat{\mu}_y^D$, we overestimate the variance of $\hat{\mu}_y^D$ for 4.1%.

APPENDIX 1

Proofs of some approximations used in the paper:

$$F_x(\hat{\lambda}_x) \approx F_x(\lambda_x) + o(\hat{\lambda}_x - \lambda_x). \quad (A1)$$

The Taylor linearization gives

$$F_x(\hat{\lambda}_x) - F_x(\lambda_x) \approx (\hat{\lambda}_x - \lambda_x) f_x(\lambda_x).$$

Approximation (A1) holds based on the assumed consistency of $\hat{\lambda}_x$, i.e. $\hat{\lambda}_x - \lambda_x \rightarrow 0$ as $n/N \rightarrow 1$ and $N \rightarrow \infty$.

Approximation (16) we express equivalently as

$$\int_{-\infty}^{\hat{\lambda}_x} \int_{-\infty}^{\lambda_x} (y - \mu_y^D) dF_{x,y}(x,y) \approx (\hat{\lambda}_x - \lambda_x) f_x(\lambda_x) E(y - \mu_y^D | x = \lambda_x). \quad (A2)$$

To prove it we start from the left side of (A2)

$$\begin{aligned} \int_{-\infty}^{\hat{\lambda}_x} \int_{-\infty}^{\lambda_x} (y - \mu_y^D) dF_{x,y}(x,y) &\approx \int_{-\infty}^{\hat{\lambda}_x} (y - \mu_y^D) \left[\int_{\lambda_x}^{\hat{\lambda}_x} f_{x,y}(x,y) dx \right] dy \\ &\approx \int_{-\infty}^{\hat{\lambda}_x} (y - \mu_y^D) \left[(\hat{\lambda}_x - \lambda_x) f_{x,y}(\lambda_x, y) \right] dy \\ &= (\hat{\lambda}_x - \lambda_x) f_x(\lambda_x) \int_{-\infty}^{\hat{\lambda}_x} (y - \mu_y^D) \frac{f_{x,y}(\lambda_x, y)}{f_x(\lambda_x)} dy \\ &= (\hat{\lambda}_x - \lambda_x) f_x(\lambda_x) E(y - \mu_y^D | x = \lambda_x). \end{aligned}$$

REFERENCES

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical*

Association, Vol. 72, No. 358, 355- 367.

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 Kovacevic, M.S. (1995). Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach. *Survey Methodology*, Vol. 21, No.2, 137-145.

Caron, N. and Chambaz, C. (1998). Mobility Matrices and Computation of Associated Precision. *Proceedings of Statistics Canada Symposium 98 "Longitudinal Analysis for Complex Surveys"*, 85-91.

Casady, R.J., Dorfman, A.H. and Wang, S. (1998). Confidence Intervals for Domain Parameters When the Sample Size is Random. *Survey Methodology*, Vol 24, No.1, 1998, pp.57-68.

Kovacevic, M.S. (2000). Variance Estimation for Gross Flows When the States Are Estimated from the Same Sample: Example of Quantile Groups. *Unpublished manuscript*.

Kovacevic, M.S. and Binder, D.A. (1997). Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach. *Journal of Official Statistics*, Vol. 13, No.1, 41-58.

Pierce, D.A. (1982). The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics. *The Annals of Statistics*. Vol. 10, No. 2, 475-478.

Randles, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *The Annals of Statistics*, Vol. 10, No. 2, 462-474.

Shao, J. and Rao, J.N.K. (1993). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhya*, Vol 55, B, 393-414.

Sarndal, C-E., Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag New York, Inc.