Variance Estimation for High Income Statistics

Wei Qian,
Statistics Canada, 150 Tunney's Pasture Dr., Ottawa, ON, K1A 0T6

Abstract

Statistics Canada produces high income statistics that provide information on demography, taxation, and income for high income populations. These statistics involve estimation of percentiles and quantities in percentile groups using a sample that covers around 1/5 of the Canadian population. For variance estimation, re-sampling methods may take unacceptable time due to the extremely large sample size; linearization cannot be applied directly because of the non-smoothness of the estimators. In this paper we propose a weighted estimating equations approach to derive linear variance estimators. Data from Prince Edward Island (P.E.I) is used to illustrate the results obtained.

Key Words: estimating equations, percentile group, variance estimation

1. Introduction

In early 2013, for the first time Statistics Canada published high income tables (1982-2010) through the Canadian Socio-Economic Information Management System (CANSIM), using data from the Longitudinal Administrative Databank (LAD). These tables provide statistics on demography, income and taxation in groups defined by income percentiles for various levels of geography (Canada, provinces, territories and census metropolitan area/census agglomeration). In this paper, we present the methods used to evaluate the quality of statistics in high income tables, more specifically variance estimation.

The paper is organized as follows. Section 2 briefly introduces the sampling design for the LAD. Section 3 summarizes the parameters of interest and their corresponding estimators in high income tables. In Section 4, linearization and re-sampling variance estimation methods are discussed, and linear variance estimators are derived using a unified estimating equations approach. In Section 5, the linear variance estimators are evaluated using the data of P.E.I. - the smallest province in Canada. The last section summarizes the findings.

2. Sampling Design for the LAD

High income tables are created mainly to provide information on taxfilers in different income groups.

The T1 family file (T1FF) is used as the frame. The T1FF is an annual cross-sectional file of all taxfilers and their families and created from personal income tax returns (T1) submitted to the Canada Revenue Agency (CRA) and other administrative files such as Child Tax Benefits. In 2010, the T1FF covered around 95% of the population.

The LAD is a 20% random sample of individuals from the T1FF. They are selected using Bernoulli sampling with equal selection probability of 1/5. In 2010, more than 5 million individuals were on the LAD. Once selected, the individual will be on the LAD file for any subsequent year if he or she is on the T1FF file for that year. Thus, the LAD is a longitudinal database and has been extensively used for longitudinal studies on income. More details about the LAD can be found in the *Longitudinal Administrative Data Dictionary* (Statistics Canada internal document, 2010). However, for high income tables the LAD was only used to produce annual cross-sectional estimates.

3. Parameter Estimation

In this section, we describe different types of statistics presented in high income tables, as well as their corresponding estimators. Table 1 summarizes the statistics in high income tables. The first item is the estimate of income percentile defining the percentile group. The others are the demographic, income and taxation characteristics. Items 15-20 are longitudinal characteristics, and their estimation depends on the sampling design of previous years. For longitudinal statistics, the associated variance estimation is much more complicated than that for cross-sectional statistics because of the dependence. In our work, the computation was simplified by treating longitudinal indicators as cross-sectional ones. For large sample sizes, the variance associated with percentile estimators is very small and the estimates are close to the actual values. Their CVs may also provide an idea of the quality of the longitudinal statistics.

Table 1: A summary of statistics in high income tables

	Statistics
1	Income threshold value
2	Number of tax filers
3	Percentage, married, males or females
4	Percentage married by sex
5	Median age
6	Median income
7	Average income
8	Share of income
9	Share of income, by sex
10	Median federal and provincial income taxes paid
11	Average federal and provincial income taxes paid
12	Share of federal and provincial income taxes paid
13	Percentage of income from wages and salaries
14	Percentage of income from wages and salaries, by sex
15	Percentage in the same quantile last year
16	Percentage in the same quantile 5 years ago
17	Percentage in top 5% last year
18	Percentage in top 5% five years ago
19	Percentage at least once in top 5% in preceding 5 years
20	Percentage always in top 5% in preceding 5 years

For methodological purposes, the parameters of interest can be summarized into six categories: the percentile of a distribution, and, within a percentile group, the mean, median, share, ratio and a function of them such as the product of share and ratio. The estimators for the different types of parameters in the high income tables are given below.

Consider a population of size N, such that $U = \{1, \dots, N\}$. Let X be the income variable defining the percentile group and Y be another variable (demographic, income or taxation) whose quantities are of interest. Let ξ_p denote the p$^{\text{th}}$ percentile of X and let γ_p denote the quantity of interest for Y in the top p$^{\text{th}}$ percentile group defined as $\{i \in U: x_i \geq \xi_p\}$. Then, (ξ_p, γ_p) are the parameters of interest. Let S be the sample drawn from U and $(\hat{\xi}_p, \hat{\gamma}_p)$ be estimators for (ξ_p, γ_p).

Since γ_p is a parameter of the population in a percentile group, the estimation of γ_p relies on the estimation of the percentile ξ_p. The weighted percentile estimator $\hat{\xi}_p$ is defined as

$$\hat{\xi}_p = \begin{cases} x_{(1)} & if\ \frac{1}{\hat{N}} w_1 > p \\ \frac{1}{2}\left(x_{(i)} + x_{(i+1)}\right) & if\ \frac{1}{\hat{N}} \sum_{j=1}^{i} w_j = p \\ x_{(i+1)} & if\ \frac{1}{\hat{N}} \sum_{j=1}^{i} w_j < p < \frac{1}{\hat{N}} \sum_{j=1}^{i+1} w_j \end{cases} \tag{1}$$

where $x_{(i)}$ is the ordered values of the income variable, w_i is the sampling weight (the inverse of the selection probability) associated with $x_{(i)}$ and $\hat{N} = \sum_{i \in S} w_i$. The weighted percentile estimator is consistent and its bias is negligible for large sample sizes. The estimate can be obtained from the SAS procedure PROC UNIVARIATE. Given $\hat{\xi}_p$, the percentile estimate, $\hat{\gamma}_p$ is defined below for the different types of parameters.

Case 1. γ_p is an average in the top p$^{\text{th}}$ percentile group

For both continuous income variables and categorical demography variables, many parameters in these tables can be expressed as an average. For example, the percentage of males in the population is the average of an indicator variable indicating male or not. Item 3, 7 and 11 in Table 1 can be expressed as an average. The estimator for an average is

$$\hat{\gamma}_p = \frac{\sum_{i \in S} w_i I\{x_i \geq \hat{\xi}_p\} y_i}{\sum_{i \in S} w_i I\{x_i \geq \hat{\xi}_p\}} \tag{2}$$

where $I\{x_i \geq \hat{\xi}_p\} = \begin{cases} 1 & if\ x_i \geq \hat{\xi}_p \\ 0 & \text{otherwise} \end{cases}$.

Case 2. γ_p is a median in the top p$^{\text{th}}$ percentile group

The top p^{th} percentile group is treated as a sub-population. Let S_p be the corresponding sub-sample such that $S_p = \{i: i \in S, x_i \geq \hat{\xi}_p\}$. Let $y_{(i)}$ be the ordered values of the variable Y for sampled units in S_p, and $w_{(i)}$ be the sampling weight associated with the unit whose y-value is $y_{(i)}$. Then, the estimator of the median of Y based on S_p is

$$\hat{\gamma}_p = \begin{cases} y_{(1)} & if\ \frac{1}{\hat{N}_p} w_1 > 0.5 \\ \frac{1}{2}\left(y_{(i)} + y_{(i+1)}\right) & if\ \frac{1}{\hat{N}_p} \sum_{j=1}^{i} w_j = 0.5 \\ y_{(i+1)} & if\ \frac{1}{\hat{N}_p} \sum_{j=1}^{i} w_j < 0.5 < \frac{1}{\hat{N}_p} \sum_{j=1}^{i+1} w_j \end{cases} \tag{3}$$

where $\hat{N}_p = \sum_{i \in S_p} w_i$.

Case 3. $\boldsymbol{\gamma_p}$ is a ratio in the top pth percentile group

Some statistics may be expressed as a ratio of the totals (average) of two variables. In Table 1, the estimator of the percentages of income from wages and salaries in the p$^{\text{th}}$ percentile group is defined as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\}\, y_i}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} x_i} \tag{4}$$

where Y is the wage and salaries and X is the income variable.

The estimator of the percentage of married by sex, can also be expressed as the ratio as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\}\, I\{\, i \text{ is married and male}\}}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\, i \text{ is male }\}},$$

where $I\{i \text{ is married and male}\} = \begin{cases} 1 & \text{if } i \text{ is married and male} \\ 0 & \text{otherwise} \end{cases}$ and $I\{i \text{ is male}\} = \begin{cases} 1 & \text{if } i \text{ is male} \\ 0 & \text{otherwise} \end{cases}$.

Longitudinal statistics are also treated as ratios. For example, the estimator of the percentage in top 5 percentiles at least once during the preceding five-year period (Item 19) is

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\}\, I\{\text{in top 5\% at least once during the preceding 5 year }\}}{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\} I\{\, i \text{ filed during the preceding 5 year }\}}.$$

The longitudinal indicator in the numerator depends on the 5th percentile estimates of the last five years. In the case of large sample sizes, those 5th percentile estimates are very close to the actual 5th percentile. Therefore, the longitudinal indicators are treated as fixed and the above estimator becomes a ratio of two indicator variables.

Case 4. $\boldsymbol{\gamma_p}$ is a share in the top pth percentile group

The share of a percentile group reflects the degree of income inequality in a population. It is defined as the ratio of total income (or tax) for the persons in the percentile group over that for all persons in the population. The estimator is given as

$$\hat{\gamma}_p = \frac{\sum_{i\in S} w_i I\{x_i \geq \hat{\xi}_p\}\, y_i}{\sum_{i\in S} w_i y_i}. \tag{5}$$

Case 5. $\boldsymbol{\gamma_p}$ is a product of share and ratio

The estimators of some parameters may not be as simple as the above cases, but they can be expressed as a function of them. For example, for men in a percentile group, the income share, γ_p, can be viewed as a product of share $\gamma_p^{(S)}$ and ratio $\gamma_p^{(R)}$:

$$\begin{aligned}\gamma_p &= \frac{income\ of\ men\ in\ a\ percentile\ group}{income\ of\ all} \\ &= \frac{income\ of\ all\ in\ a\ percentile\ group}{income\ of\ all} \times \frac{income\ of\ men\ in\ a\ percentile\ group}{income\ of\ all\ in\ percentile\ group} \\ &= \gamma_p^{(S)} \times \gamma_p^{(R)}.\end{aligned}$$

Accordingly, the estimator is defined as

$$\hat{\gamma}_p = \hat{\gamma}_p^{(S)} \times \hat{\gamma}_p^{(R)}, \tag{6}$$

where $\gamma_p^{(R)}$ and $\hat{\gamma}_p^{(R)}$ are given as **Case 3** and $\gamma_p^{(S)}$and $\hat{\gamma}_p^{(S)}$ are given in **Case 4**.

4. Variance Estimation

Two types of variance estimation methods are usually considered for household surveys: re-sampling and linearization. Bootstrap and jackknife are the two most popular re-sampling methods used for household surveys at Statistics Canada. The jackknife method is often used for surveys with multi-stage clustering design such as the Labour Force Survey (LFS). In this study, the jackknife is ruled out because it performs poorly for estimating the variance of non-smooth estimators such as sample percentiles. Bootstrap variance estimators are commonly used in household surveys, such as Survey of Labour and Income Dynamics (SLID). The advantage of the bootstrap method is that, 1) it works well for non-smooth estimators under simple sampling designs, and 2) it is easy to implement. It is not necessary to develop formulas for the different estimators. The bootstrap algorithm for Bernoulli sampling is very simple. The disadvantage of the bootstrap is the time and computational resources required. As stated previously, the LAD sample size now is more than 5 million records. Running the estimation process repeatedly on the LAD for all geography levels would take a tremendous amount of time. For example, for Ontario, more than 3 weeks was required to produce all tables. However, the bootstrap provides a tool to verify other variance estimators for some smaller domains; moreover, the bootstrap may be preferable for analytical purposes as the analysts can use the bootstrap samples to generate replicates of test statistics and then produce confidence interval estimates.

On the other hand, linearization methods have long been used in surveys and the theory is well developed. Standard variance estimation methods from textbooks can be used only for linear estimator, such as the Horvitz-Thompson (HT) estimator (see Särndal et al., 1991). For a smooth nonlinear estimator, Taylor linearization permits the nonlinear estimator to be approximated by a HT total estimator for a new variable - *linear variable*. Then, the variance of the nonlinear estimator may be approximated by the variance of an HT total estimator which, in turn, can be estimated by the standard methods. For example, suppose $\hat{\theta}$ is a non-linear “smooth” estimator and Z is the associated linear variable; then,

$$V(\hat{\theta}) \approx V(\textstyle\sum_{i \in S} w_i z_i), \tag{7}$$

where S is the sample and z_i is the value for the linearized variable attached to unit i. The problem with the linearization method is that a linear variable must be found for each estimator and the linearization method is not easily generalized. For example, if a quantity in a low percentile group is of interest, the formula for variance estimation developed for the top percentile group cannot be reused. However, the linearization method does not require replication therefore the computation is fast. In addition, it provides consistent variance estimates. The linearization variance estimation method is discussed below.

As stated previously, the sampling design for the LAD is very simple: Bernoulli sampling with the selection probability of 0.2. As a result, the variance formula given by (7) can be simplified as

$$V(\hat{\theta}) \approx 4 \textstyle\sum_{i \in U} z_i^2, \tag{8}$$

where U is the population. In the case where the number of individuals in the population is not available (the population counts in some small geographies may not be provided), the variance estimator is then given by

$$\hat{V}(\hat{\theta}) = 20 \textstyle\sum_{i \in S} \hat{z}_i^2, \tag{9}$$

where $\hat{z}_i$ is a proper estimator of z_i since z_i may involve some unknown finite population quantities.

Binder (1983) introduced a unified estimating equations approach for estimating finite population parameters. The estimating equations approach assumes that the finite population is a sample from a superpopulation model and the sample is a subsample of the finite population. Any finite population parameter θ can be viewed as a solution of "census" estimating equations:

$$U(\theta) = \sum_{i \in U} u(\theta, y_i) = 0.$$

The estimator $\hat{\theta}$ can be found by solving the corresponding weighted estimating equations:

$$\hat{U}(\theta) = \sum_{i \in S} w_i u(\theta, y_i) = 0,$$

where $\hat{U}(\theta)$ is the HT total estimator of $U(\theta)$. Under regularity conditions, $\hat{\theta}$ is a consistent estimator of θ. For the case where θ is a parameter vector, u is a vector of the same dimension as θ.

Suppose θ_0 is the true value of θ. Taylor linearization around θ_0 leads to

$$\frac{1}{N}\left(\hat{U}(\hat{\theta}) - \hat{U}(\theta_0)\right) \approx \frac{1}{N}\left[\frac{\partial \hat{U}(\theta)}{\partial \theta}\right]_{\theta=\theta_0} \left(\hat{\theta} - \theta_0\right) \approx \left[\frac{\partial E[u(\theta;Y)]}{\partial \theta}\right]_{\theta=\theta_0} \left(\hat{\theta} - \theta_0\right).$$

where the expectation is under the superpopulation model. For non-smooth statistics, linearization is used after the following approximation,

$$\frac{1}{N}\left(\hat{U}(\hat{\theta}) - \hat{U}(\theta_0)\right) \approx E[u(\theta;Y)]|_{\hat{\theta}} - E[u(\theta;Y)]|_{\theta_0}.$$

The conditions for the approximation are discussed in Randles (1982), Shao and Rao (1994). Suppose that θ is a parameter vector. Then,

$$\hat{\theta} - \theta_0 \approx -\frac{1}{N}\left[\frac{\partial E[u(\theta;Y)]}{\partial \theta}\right]_{\theta=\theta_0}^{-1} \hat{U}(\theta_0).$$

Wang and Opsomer (2011) propose another method for non-smooth estimator that produces the similar result, and here we will not discuss it further. Then, the variance of $\hat{\theta}$ is

$$V(\hat{\theta}) \approx V\left(\textstyle\sum_{i \in S} w_i u_i^*\right) = \textstyle\sum_{k,l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right) u_k^* u_l^{*T} \qquad (10)$$

where $u_i^* = -\frac{1}{N}\left[\frac{\partial E[u(\theta;Y)]}{\partial \theta}\right]_{\theta=\theta_0}^{-1} u_i(\theta_0, y_i)$. Since u_i^* may involve unknown quantities, they can be replaced by the proper estimate $\hat{u}_i^*$. As a result, the variance estimator becomes

$$\hat{V}(\hat{\theta}) = \textstyle\sum_{k,l \in S} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}}\right) \hat{u}_k^* \hat{u}_l^{*T}. \qquad (11)$$

Thus, it remains to find u, u_i^* and $\hat{u}_i^*$.

High income statistics involve the estimation of percentiles (non-smooth statistics) and quantities in the top pth percentile group. The application of estimating equations approach to non-smooth statistics is also discussed in Binder and Kovacevic (1995) and Osier (2009).

Let $u_i = (u_{1i}, u_{2i})^T$ be estimating functions for (ξ_p, γ_p), where ξ_p is the pth percentile of the income variable X and γ_p is the quantity of interest for variable Y in the top pth

percentile group defined by ξ_p. In this study, X and Y are different variables, while both reference papers only discussed the case where X and Y are the same. This difference leads to the estimation of their conditional distributions for which a nonparametric method in Borkowf et al. (1996) is used.

Assume that X is a nonnegative continuous variable[1]. For ξ_p, the p$^{\text{th}}$ percentile of the variable X, and its estimator $\hat{\xi}_p$, the estimating equation and linear variable are

$$u_{1i} = I\{x_i \le \xi_p\} - p,$$
$$u_{1i}^{*} = -\frac{1}{f(\xi_p)}\left[I\{x_i \le \xi_p\} - p\right], \qquad and$$
$$\hat{u}_{1i}^{*} = -\frac{1}{\hat{f}(\hat{\xi}_p)}\left[I\{x_i \le \hat{\xi}_p\} - p\right],$$

where $\hat{u}_{1i}^{*}$ needs the estimation of $f(\xi_p)$ – the probability density function of X at ξ_p .

Two possible methods can be used for the estimation of the density function for complex survey data. Francisco and Fuller (1991) use the density estimator

$$\hat{f}(x) = \frac{2z_{\alpha/2}\delta}{h_1 + h_2}$$

where

$$\delta^2 = mse\left\{\sum_{i\in S} w_i[I\{x_i \le x\} - p]\right\},$$

$z_{\alpha/2}$ is the $100\left(1 - \frac{\alpha}{2}\right)$-th percentile from the standard normal distribution, and h_1 and h_2 are found by solving

$$\inf_{h_1}\left\{\frac{1}{\hat{N}}\sum_{i\in S} w_i[I\{x_i \le x - h_1\} - p] \le -z_{\frac{\alpha}{2}}\delta\right\},$$
$$and \quad \inf_{h_2}\left\{\frac{1}{\hat{N}}\sum_{i\in S} w_i[I\{x_i \le x + h_2\} - p] \ge z_{\alpha/2}\delta\right\}.$$

Lohr and Buskirk (1999) propose a weighted kernel density estimator such that

$$\hat{f}(x) = \frac{1}{\hat{N}}\sum_{i\in S} w_i \phi_h(x - x_i),$$

where h is the bandwidth and

$$\phi_h(t) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{t^2}{2h^2}\right),$$

is the standard normal density rescaled by the bandwidth. The bandwidth is obtained by

$$h = 0.79\hat{Q}n^{-\frac{1}{5}},$$

where $\hat{Q}$ is the sample interquartile range (IQR). Note that the kernel density estimation is very sensitive to the choice of bandwidth, especially at the tail of the distribution.

In this paper, the method proposed by Francisco and Fuller (1991) is used. As suggested by Rao and Wu (1987), α is set to 0.05. Using the data from selected small domains, the variance estimates are shown to be very similar to the bootstrap variance estimates.

[1] Some individuals may have negative income values. Since we only consider estimating the parameters in the top percentile groups, setting these negative values to zero has little impact.

For a different variable Y and its corresponding quantity γ_p and estimator $\hat{\gamma}_p$, estimating functions and associated linearized variables are presented below for the different cases. An approach similar to that given in Binder and Kovacevic (1995) has been used to derive the linear variables.

Case 1. γ_p is an average in the top p^th^ percentile group

The mean estimator $\hat{\gamma}_p$ is used not only for continuous income variables but also for categorical demography variables. For example, to estimate the percentage of males, we only need to create a variable indicating male or not, the percentage of males is the average of the indicator variable. The estimating function for γ_p is

$$u_{2i} = I\{x_i \geq \xi_p\}(y_i - \gamma_p),$$

and the associated linearized variable is

$$u^*_{2i} = \frac{1}{N(1-p)}\left\{\left(\gamma_p - E\left[Y\middle|\xi_p\right]\right)\left(I_{\{x\leq\xi_p\}} - p\right) + I_{\{x\geq\xi_p\}}\left(y_i - \gamma_p\right)\right\}.$$

By replacing all unknown quantities above by proper estimators, the above formula becomes

$$\hat{u}^*_{2i} = \frac{1}{\hat{N}(1-p)}\left\{\left(\hat{\gamma}_p - \hat{E}\left[Y\middle|\hat{\xi}_p\right]\right)\left(I_{\{x\leq\hat{\xi}_p\}} - p\right) + I_{\{x\geq\hat{\xi}_p\}}\left(y_i - \hat{\gamma}_p\right)\right\}.$$

A nonparametric method is used to estimate $\left[Y\middle|\xi_p\right]$, the conditional expected value of Y given X at ξ_p. The nonparametric estimator (Nadaraya-Watson kernel estimator with the normal kernel and the same bandwidth h for $\hat{\xi}_h$) is given by

$$\hat{E}[Y|x] = \frac{\sum_{i\in S} w_i y_i \phi_h(x - x_i)}{\sum_{i\in S} w_i \phi_h(x - x_i)}.$$

Note that if Y and X are the same variable, then $\hat{E}[Y|x] = x$.

Case 2. γ_p is a median in the top p^th^ percentile group

Assume that Y is a continuous nonnegative variable. Denote $f_X(x)$ and $F_X(x)$ as the marginal density and cumulative distribution function (CDF) of X and $f_Y(y)$ and $F_Y(x)$ as the marginal density and CDF of Y. Denote $F_{X|Y}(x|y)$ as the conditional CDF of X given $Y = y$ and $F_{Y|X}(y|x)$ the conditional CDF of Y given $X = x$.

The estimating function for γ_p , the median of Y in the top p^th^ percentile group is

$$\mu_{2i} = I\{x_i \geq \xi_p\}[I\{y_i \leq \gamma_p\} - 0.5],$$

and the associated linearized variable is

$$u^*_{2i} = \frac{1}{[1 - F_{X|Y}(\xi_p|\gamma_p)]f_Y(\gamma_p)}\{[0.5 - F_{Y|X}(\gamma_p|\xi_p)](I\{x_i \leq \xi_p\} - p) + I\{x_i \geq \xi_p\}[I\{y_i \leq \gamma_p\} - 0.5]\}.$$

After replacing all the population quantities by their estimates, the formula becomes

$$\hat{u}^*_{2i} = \frac{1}{[1 - \widehat{F_{X|Y}}(\hat{\xi}_p|\hat{\gamma}_p)]\hat{f}_Y(\hat{\gamma}_p)}\{[\widehat{F_{Y|X}}(\hat{\gamma}_p|\hat{\xi}_p) - 0.5](I\{x_i \leq \hat{\xi}_p\} - p) + I\{x_i \geq \hat{\xi}_p\}[I\{y_i \leq \hat{\gamma}_p\} - 0.5]\},$$

where $\widehat{F_{X|Y}}$, $\hat{f}_Y$, and $\widehat{F_{Y|X}}$ are the estimators of $F_{X|Y}$, f_Y and $F_{Y|X}$ respectively.

For the estimation of the conditional distribution $F_{X|Y}$, one can follow Borkowf et al. (1997),

$$\begin{aligned} F_{X|Y}(\xi_p|\gamma_p) &= P(X \le \xi_p | Y = \gamma_p) \\ &= P\big(F_X(X) \le p \big| F_Y(Y) = F_Y(\gamma_p)\big), \end{aligned}$$

which leads to

$$\widehat{F_{X|Y}}(\xi_p|\gamma_p) = \frac{\sum_{i\in S} w_i I\{\widehat{F_X}(x_i) \le p, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\sum_{i\in S} w_i I\{|\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}},$$

where $\hat{\delta}^2 = \frac{0.8}{n}\widehat{F_Y}(\hat{\gamma}_p)(1 - \widehat{F_Y}(\hat{\gamma}_p))$ and $z_{\alpha/2}$ is the $100\left(1 - \frac{\alpha}{2}\right)$-th percentile from the standard normal distribution. Similarly, the conditional CDF $F_{Y|X}(\gamma_p|\xi_p)$ is given by

$$\widehat{F_{Y|X}}(\hat{\gamma}_p|\hat{\xi}_p) = \frac{\sum w_i I\{\widehat{F_Y}(y_i) \le \widehat{F_Y}(\hat{\gamma}_p), |\widehat{F_X}(x_i) - p| \le z_{\alpha/2}\hat{\delta}^*\}}{\sum w_i I\{|\widehat{F_X}(x_i) - p| \le z_{\alpha/2}\hat{\delta}^*\}},$$

where $\hat{\delta}^{*2} = 0.8p(1-p)/n$.

Using the approach used by Francisco and Fuller (1991), the marginal density estimator for $h(\gamma_p)$ is given by

$$\widehat{F_Y}(\hat{\gamma}_p) \approx \frac{\sum_{i\in S} w_i I\{|\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\hat{N}(Y_{max} - Y_{min})},$$

where $Y_{max} = \max\{y_i: i \in S, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}$ and $Y_{min} = \min\{y_i: i \in S, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}$. Hence,

$$[1 - \widehat{F_{X|Y}}(\hat{\xi}_p|\hat{\gamma}_p)]\hat{f}_Y(\hat{\gamma}_p) = \frac{\sum_{i\in S} w_i I\{\widehat{F_X}(x_i) > p, |\widehat{F_Y}(y_i) - \widehat{F_Y}(\hat{\gamma}_p)| \le z_{\alpha/2}\hat{\delta}\}}{\hat{N}(Y_{max} - Y_{min})}.$$

Case 3. $\boldsymbol{\gamma_p}$ is a ratio in the top p$^{\text{th}}$ percentile group

Within a percentile group, the proportion of total income from wages and salary is a ratio. Let Y be the variable in the numerator and Z be the variable in the denominator in this ratio, the estimating function and linear variable for the ratio are

$$\mu_{2i} = I\{x_i \ge \xi_p\}(y_i - \gamma_p z_i),$$

$$\mu_{2i}^* = \frac{\left\{(E[Y|\xi_p] - \gamma_p\xi_p)\left(I_{\{x_i \le \xi_p\}} - p\right) + I_{\{x_i \ge \xi_p\}}(y_i - \gamma_p z_i)\right\}}{E[ZI(X \ge \xi_p)]} \ and$$

$$\hat{\mu}_{2i}^* = \frac{\left\{(\hat{E}[Y|\hat{\xi}_p] - \hat{\gamma}_p\hat{\xi}_p)\left(I_{\{x_i \le \hat{\xi}_p\}} - p\right) + I_{\{x_i \ge \hat{\xi}_p\}}(y_i - \hat{\gamma}_p z_i)\right\}}{\hat{E}[ZI(X \ge \hat{\xi}_p)]}.$$

where $\hat{E}[ZI(X \ge \hat{\xi}_p)] = \frac{\sum_i w_i z_i I(x_i \ge \hat{\xi}_p)}{\sum_i w_i}$.

Case 4. $\boldsymbol{\gamma_p}$ is a share for the top p$^{\text{th}}$ percentile group

The estimating function and linear variable for the share are

$$\mu_{2i} = [I\{x_i \ge \xi_p\} - \gamma_p]y_i$$

$$\mu_{2i}^* = \frac{1}{\mu_y}\{E[Y|\xi_p](I\{x_i \le \xi_p\} - p) + (I\{x_i \ge \xi_p\} - \gamma_p)y_i\},$$

and

$$\hat{\mu}_{2i}^* = \frac{1}{\hat{\mu}_y}\{\hat{E}[Y|\hat{\xi}_p](I\{x_i \le \hat{\xi}_p\} - p) + (I\{x_i \ge \hat{\xi}_p\} - \hat{\gamma}_p)y_i\},$$

where $\hat{\mu}_y = 1/\hat{N}\sum_{i\in S} w_i y_i$.

Case 5. γ_p is a product of share and ratio

When γ_p is a product of a ratio (as defined in **Case 3)** and a share (as defined in **Case 4)** such that $\gamma_p = \gamma_p^{(S)}\gamma_p^{(R)}$, the estimating function and linear variable are

$$\mu_{2i} = \gamma_p^{(S)}\mu_{2i}^{(S)} + \gamma_p^{(R)}\mu_{2i}^{(R)},$$
$$\mu_{2i}^{*} = \gamma_p^{(S)}\mu_{2i}^{(S)*} + \gamma_p^{(R)}\mu_{2i}^{(R)*} \quad and$$
$$\hat{\mu}_{2i}^{*} = \hat{\gamma}_p^{(S)}\hat{\mu}_{2i}^{(S)*} + \hat{\gamma}_p^{(R)}\hat{\mu}_{2i}^{(R)*},$$

where $\mu_{2i}^{(R)}$, $\mu_{2i}^{(R)*}$, and $\hat{\mu}_{2i}^{(R)*}$ have been previously given for a ratio in **Case 3** and $\mu_{2i}^{(S)}$, $\mu_{2i}^{(S)*}$, and $\hat{\mu}_{2i}^{(S)*}$ for a share in **Case 4.**

A special case: γ_p is the count in the top p^th percentile group

The count in the top p^th percentile group where $\gamma_p = N(1-p)$ and $\hat{\gamma}_p = \hat{N}(1-p)$ is a special case of parameter of interest. The variance of $\hat{\gamma}_p$ is

$$V(\hat{\gamma}_p) = 4N(1-p)^2.$$

Hence, the corresponding CV estimate is given by

$$\widehat{CV}(\hat{\gamma}_p) = \sqrt{\hat{V}(\hat{\gamma}_p)}/\hat{\gamma}_p = 2/\sqrt{\hat{N}}.$$

5. Performance Evaluation of Variance Estimation

In this section, the variance estimators are evaluated. The linear variance estimates are compared to both the approximate true variance calculated from the T1FF and the bootstrap variance estimates. This evaluation is only done for Prince Edward Island (P.E.I.) as it yields the largest variances at the provincial level.

Suppose θ is the parameter of interest and $\hat{\theta}$ is a consistent estimator of θ. The relative bias (RB) is defined as

$$RB(\hat{\theta}) = (\hat{\theta} - \theta)/\theta \times 100\%.$$

The approximate CV (ACV) of $\hat{\theta}$ is defined as

$$ACV(\hat{\theta}) = \sqrt{AV(\hat{\theta})}/\theta \times 100\%,$$

where $AV(\hat{\theta})$ is computed by the formula given in (10). The true parameter values are computed from the T1FF. The CV estimator on the sample is defined as

$$\widehat{CV}(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})}/\hat{\theta} \times 100\%.$$

The bootstrap CV estimate is based on 1,000 bootstrap replicates (bootstrap weights) and defined as

$$\widehat{CV}^b(\hat{\theta}) = \sqrt{\hat{V}^b(\hat{\theta})}/\hat{\theta} \times 100\%.$$

The replicates were generated, using the pseudo-population approach (see Beaumont and Patak, 2012).

Variance estimates for selected high income statistics are produced. Since the parameter estimators have large sample properties, it is expected that the CVs for other provinces should be smaller than that for P.E.I.

Table 2 lists the approximate CV and CV estimates generated using the linearization method and bootstrap method for the top 1% income group in P.E.I.. Small differences are observed between the approximate CV and two estimated CVs. CV estimates from two methods are very similar with largest difference being 0.8% for the product of share and ratio. Studies on other income group show that the quality of estimation for large group usually is better because of the larger sample size.

Table 2. Comparison of CV estimates for linearization and Bootstrap for estimates of the top 1% group for P.E.I (2009)

					CV		
Measure	**Variables**	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	**RB**	$\boldsymbol{ACV}(\hat{\boldsymbol{\theta}})$	$\widehat{\boldsymbol{CV}}(\hat{\boldsymbol{\theta}})$	$\widehat{\boldsymbol{CV}}^{\boldsymbol{b}}(\hat{\boldsymbol{\theta}})$
Threshold	Income	131,115	130,824	-0.2%	1.9%	1.8%	2.0%
Total	Counts	1,098	1,089	-0.8%	0.6%	0.6%	0.6%
Mean	% of Male	78.20%	80.3%	2.7%	3.2%	3.0%	3.0%
	% of Married	82.40%	83.5%	1.3%	2.8%	2.7%	2.9%
	Income	214,589	212,371	-1.0%	5.7%	4.6%	4.6%
	Tax	65,700	63,132	-3.90%	7.4%	6.1%	6.0%
Median	Income	166,331	166,192	-0.1%	3.0%	2.4%	2.5%
	Tax	48,666	48,923	0.5%	3.6%	3.7%	3.9%
	Age	53	53.5	0.9%	1.1%	1.9%	2.2%
Ratio	Wage/Income	59.70%	63.0%	5.50%	4.8%	4.8%	4.9%
Share	Income	6.6%	6.5%	-1.5%	5.2%	4.2%	4.3%
	Tax	11.40%	11.0%	-3.5%	6.5%	5.3%	5.3%
Share By Ratio	Income by males		5.3%			5.0%	5.8%

Table 3 gives the 95% confidence interval (CI) estimates from both methods. The linear CI estimates is calculated as $\left(\hat{\boldsymbol{\theta}} - 1.96\sqrt{\hat{V}(\hat{\boldsymbol{\theta}})}, \hat{\boldsymbol{\theta}} + 1.96\sqrt{\hat{V}(\hat{\boldsymbol{\theta}})}\right)$; the bootstrap CI is based on 1000 bootstrap replicates $\left(2\hat{\boldsymbol{\theta}} - \hat{\theta}^*_{(0.025)}, 2\hat{\boldsymbol{\theta}} - \hat{\theta}^*_{(0.975)}\right)$. Two methods produce very similar results. This implies that the asymptotic normality assumption of the estimators is satisfied. It should be noted that the linearization intervals may be slightly shorter than the bootstrap intervals, which is very common for those two methods.

Table 3. Comparison of 95% Confidence Interval (C.I.) estimates from linearization and Bootstrap for parameters in the top 1% group for P.E.I (2009)

		95% Confidence Interval			
		Linearization		Bootstrap	
Measure	**Variables**	**lower**	**upper**	**lower**	**upper**
Threshold	Income	126,209	135,439	125,486	135,660
Total	Counts	1,076	1,102	1,068	1,098
Mean	% of Male	75.6%	85.0%	75.6%	85.0%
	% of Married	79.1%	87.9%	79.0%	88.3%
	Income	193,224	231,518	193,297	231597
	Tax	55,584	70,680	55,301	70,335
Median	Income	158,700	173,684	158,136	172947
	Tax	45,567	52279	46,883	53,789
	Age	51.5	55.5	52.0	56.0
Ratio	Wage in Income	57.1%	68.9%	56.4%	68.6%
Share	Income	6.0%	7.0%	5.9%	7.0%
	Tax	9.9%	12.1%	9.8%	12.1%

6. Summary

In this paper, two methods for variance estimation have been considered for statistics in the high income tables: linearization and bootstrap. The estimates of CVs and CIs for these two methods are very close. However, in practice, the linearization method is employed as the computing time for the bootstrap method is extreme. Note that the linearization method requires the first derivative for each estimator.

Acknowledgements

I would like to thank Jean François Beaumont, Christian Nadeau, and Wisner Jocelyn for providing many useful comments on the paper. I also thank Cynthia Bocci and Yong You for reviewing the paper.

References

Beaumont, J.F. and Patak, Z. 2012. On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127-148.

Binder, D., and Kovacevic, M. 1995. Estimating some measures of income inequality from survey data. *Survey Methodology*, 21,137-145.

Binder, D. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Borkowf, C.B., Gail, M.H., Carroll, R., and Gill, R.D. 1997. Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of marginal distributions. Biometrics, 53, 1054 – 1069.

Francisco, C.A. and Fuller, W.A. 1991. *Quantile estimation with a complex survey design*. The Annals of Statistics, 19, 454 – 469.

Lohr, S. and Buskirk, T. 1999. Density estimation with complex survey data. SSC Annual Meeting, June, 1999, *Proceedings of the Survey Methods Section*.

Osier,G. 2009. Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Randles. R.H. 1982. On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

Särndal, C.E., Swensson, B. and Wretman, J. 1991. Model Assisted Survey Sampling. New York: Wiley.

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

Statistics Canada. 2010. Longitudinal Administrative Data Dictionary. Catalogue no. 12-585-X.

Wang, J.C. and Opsomer, J.D. 2011. On Asymptotic Normality and Variance Estimation for Nondifferentiable Survey Estimators. *Biometrika*, 98, 1, 91-106.