Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators

by Eric Graf and Yves Tillé

June 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 61-79
Statistics Canada, Catalogue no. 12-001-X

Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators

Eric Graf and Yves Tillé[1]

Abstract

We have used the generalized linearization technique based on the concept of influence function, as Osier has done (Osier 2009), to estimate the variance of complex statistics such as Laeken indicators. Simulations conducted using the R language show that the use of Gaussian kernel estimation to estimate an income density function results in a strongly biased variance estimate. We are proposing two other density estimation methods that significantly reduce the observed bias. One of the methods has already been outlined by Deville (2000). The results published in this article will help to significantly improve the quality of information on the precision of certain Laeken indicators that are disseminated and compared internationally.

Keywords: influence function; EU-SILC survey; non-linear statistics; poverty and inequality indicators.

1 Introduction

Deville (2000) proposed that the precision of non-linear statistics in sampling designs be estimated using the generalized linearization method, which relies on the concept of influence function proposed by Hampel (1974) in the field of robust statistics. Osier (2009) applied these theories to estimate the variance of complex statistics such as the Laeken indicators (Eurostat 2005) in the European Statistics on Income and Living Conditions (EU-SILC) survey. Goga, Deville and Ruiz-Gazen (2009) extend the theory of Deville (2000) to two-sample surveys. Verma and Betti (2011) provide a comprehensive list of traditional poverty indicators and associated linearized variables, and they also compare the performance of the linearization technique with that of the jackknife repeated replication method. In this article, we will limit ourselves to poverty indicators published in the EU-SILC survey and focus on the way to estimate the income density function at various points in the distribution.

In Section 2, we review the required theoretical foundations, the expressions for the poverty and inequality indicators being studied, and the linearized variables of those indicators. Some linearized variables are dependent on the density function of the variable of interest, which is usually estimated using the Gaussian kernel estimation method. Two alternative methods are presented in Section 3. The R-language simulations are described and discussed in Section 4. We show that Gaussian kernel estimation can generate strong bias in the estimated variance of indicators when an estimate of the income density function is being used. We also show that the other two density estimation methods proposed in Section 3 reduce the observed bias, and this is discussed in the findings in the last part of this article.

2 Review of given poverty indicators and their linearized variables

Let U be a finite population consisting of N identifiable units $u_1, \ldots, u_k, \ldots, u_N$. To simplify the notation, let unit u_k be denoted simply by the index k. In practice the population U is a sampling frame with acceptable coverage of a given population for which we wish to make inferences. To each unit k is

1. Eric Graf and Yves Tillé, Institute of Statistics, Faculty of Economics and Business, University of Neuchâtel, Rue de la Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland. Email: eric.graf@unine.ch and yves.tille@unine.ch.

associated the value y_k for a given characteristic (in this case, income). Without loss of generality, to simplify the notation, assume that the values of y_k are distinct and sorted by order of magnitude, so that $y_k = y_{[k]}$. Data from sample surveys often contain duplicates, that is, a number of units with the same value y, as a result of rounding or range questions. In these cases and for this study, we can simply increase the values by a small (negligible), randomly selected, uniformly distributed amount so that the data may be sorted unambiguously.

Let S be a random sample of size n obtained using a sample design $p(s) = \mathrm{P}(S = s)$, for all $s \subset U$. In addition, let $\pi_k = \mathrm{P}(k \in s) > 0$ be the inclusion probability of unit k of U. As well, let $d_k = 1/\pi_k$ be the sampling weight, and let $w_k = w_k(s)$ be the estimation weight, which may be equal to d_k or may be more refined. For example, w_k may have been obtained after calibration (Deville and Särndal 1992) and therefore also reflect a non-response adjustment.

The estimators of poverty and inequality indicators are non-linear statistics that can't be expressed as regular functions of totals (that is, continuously differentiable up to the second order). In fact, they are rank statistics for the Gini coefficient and quantile statistics for the others. As Osier (2009) points out, their variance therefore can't be estimated using a Taylor linearization; the generalized linearization method is required instead (Deville 2000, Demnati and Rao 2004, and Osier 2009). An alternative for estimating variance would be to use bootstrap-type re-sampling techniques but, for the EU-SILC survey data, preference was given to the linearization technique, at least for a certain number of participating countries. Indeed, re-sampling methods often require more human and machine resources. As well, since Eurostat collaborates with some 30 countries that have different sampling designs and that may perform non-response adjustments and calibration to external sources, it seemed more appropriate to select an analytical solution for estimating variance. In addition, some countries might be using the existing SAS software POULPE (Ardilly and Osier 2007) to generate the required estimates. That was the case for initial tests using Swiss EU-SILC data. Here we use a procedure that, as Antal, Langel and Tillé (2011) point out, reconciles the approach introduced by Deville (2000) with that of Demnati and Rao (2004). Both approaches use the concept of *influence function* initially developed in the field of robust statistics (Hampel 1974). Antal *et al.* (2011) also state that the same linearized variables can be found by applying the method proposed by Graf (2011 and 2013) that constructs a linearized variable on the basis of a Taylor expansion with respect to sample inclusion indicators. Note also the work by Kovačević and Binder (1997) in which a linearization approach using estimating equations is developed.

Deville (2000) states that the influence of a unit k on a population parameter of interest θ is determined by an infinitesimal variation in the importance assigned to the unit. The parameter is expressed as a functional $\theta = T(M)$, where M is a measure allocating a mass of 1, $M(k) = M_k = 1$, only at points on the continuum corresponding to units $k \in U$. The specialization of the general measure M into a discrete measure turns the functional T, predefined on a continuum, into a discrete functional, in the same way as the total Y is defined as the sum of all y_k over the given finite population. The *influence function* of T, or the *linearized variable,* is defined as

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ for all } k \in U,$$

where δ_k is the Dirac measure for unit k $(\delta_k(i) = 1$ if $i = k$ and 0 otherwise$)$. In practice, we have known data only from a sample S and Deville (2000) defines a linearized variable $\hat{z}_k$, or *empirical* *influence function*, by (1) determining the limit above using differential calculus and (2) replacing the unknowns in the evaluation with the corresponding estimated quantities using the sample. Deville justifies this procedure by showing that

$$T(\hat{M}) - T(M) \approx \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k\right).$$

The key result is that, under asymptotic conditions described by Deville (2000), which are in theory satisfied when the sample is "sufficiently large", the variance of the estimated total of the variable $\hat{z}_k$ is an approximation of the variance of the (complex) statistic $\hat{\theta}$:

$$\text{var}\left[\sum_{k \in s} \hat{z}_k w_k\right] \approx \text{var}(\hat{\theta}).$$

The starting point of Deville's approach is therefore the population parameter and not the estimator that is proposed to be used for the evaluation using the sample. When the estimator used follows naturally from the population parameter expression (for example, the total Y approached by the Horvitz-Thompson estimator), the procedure is unambiguous. However, imprecision arises if we estimate the same total Y using the ratio estimator with an auxiliary variable *x*. In that case, Deville's approach, which does not specify the form of the total estimator to use, will yield a constant influence function equal to 1, instead of bringing the unknown ratio of interest into play.

An alternative that avoids these problems is the approach by Demnati-Rao, when used in Deville's framework, as done in Antal *et al.* (2011). They present the Demnati-Rao approach as resulting from Deville's framework when the measure M used is not the discrete measure defined for U described above, but rather the following measure defined for S, the sample:

$$\hat{M}(k) = w_k, k \in S$$

where w_k is a weight. By defining the measure for S, the starting point becomes the estimator and not the parameter; it is the parameter that is initially expressed as a functional, and not the population parameter to be estimated. That is, the functional corresponds to the estimator for which we are seeking a variance estimate using generalized linearization. We then obtain the linearized variable based on that functional as follows:

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{ for all } k \in S.$$

Antal *et al.* (2011) note that, to the extent that the functional in this limit is expressed as an explicit function of the variables that are the weights assigned by the measure $\hat{M}$ to the observations, this linearized variable is in fact a function of the partial derivatives with respect to the weights:

$$I[T(\hat{M})]_k = \frac{\partial T(\hat{M})}{\partial w_k}.$$

Antal *et al.* (2011) point out that the linearized variables that we will discuss below can be obtained using either approach. In fact, computing the limit using the Demnati-Rao approach does not necessarily result in the variance estimate suggested by Deville (2000). The practical approach used in this article might therefore be called the Deville-Demnati-Rao approach, in recognition of the theoretical framework provided by Deville (2000) and the practical algorithm for Deville's framework provided by Demnati and Rao (2004).

Using this method, the variance of $\hat{\theta}$ can be estimated for any sampling design, and a confidence interval can therefore be obtained by substituting the linearized variable in the variance formula for a total for the selected sampling design. If the sampling design is simple random sampling without replacement, the estimator of the variance of an inequality indicator $\hat{\theta}$ is defined as

$$\widehat{\text{var}}_{\text{lin}}[\hat{\theta}] = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{k\in S}(\hat{z}_k - \bar{z})^2, \tag{2.1}$$

where

$$\bar{z} = n^{-1}\sum_{k\in S}\hat{z}_k.$$

Below, we review the empirical definitions of the inequality indicators considered with respect to population income measurement, as well as the expressions for the linearized variables of the indicators as we have implemented them.

2.1 Gini coefficient

The Gini coefficient, G, ranges from 0 (complete equality, that is, all individuals earn the same amount) to 1 (complete inequality, that is, one individual has all the income and the other individuals have no income). The coefficient G is expressed on the basis of the cumulative income of a given proportion of the poorest individuals. If $\mathcal{Y}$ is a random variable representing income, $f(y)$ its density function and $F(y)$ its distribution function, then the *Lorenz curve* (Lorenz 1905) can be written as

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)\,dy}{\int_0^{\infty} yf(y)\,dy} = \frac{1}{\mathrm{E}(\mathcal{Y})}\int_0^{\alpha} F^{-1}(u)\,du.$$

The Gini coefficient represents twice the area between the Lorenz curve and the line of complete equality (the diagonal line $f_{eg}(x) = x$), as shown in Figure 2.1. Therefore, the Gini coefficient can be defined as

$$G = 2\int_0^1 [\alpha - L(\alpha)]\,d\alpha.$$

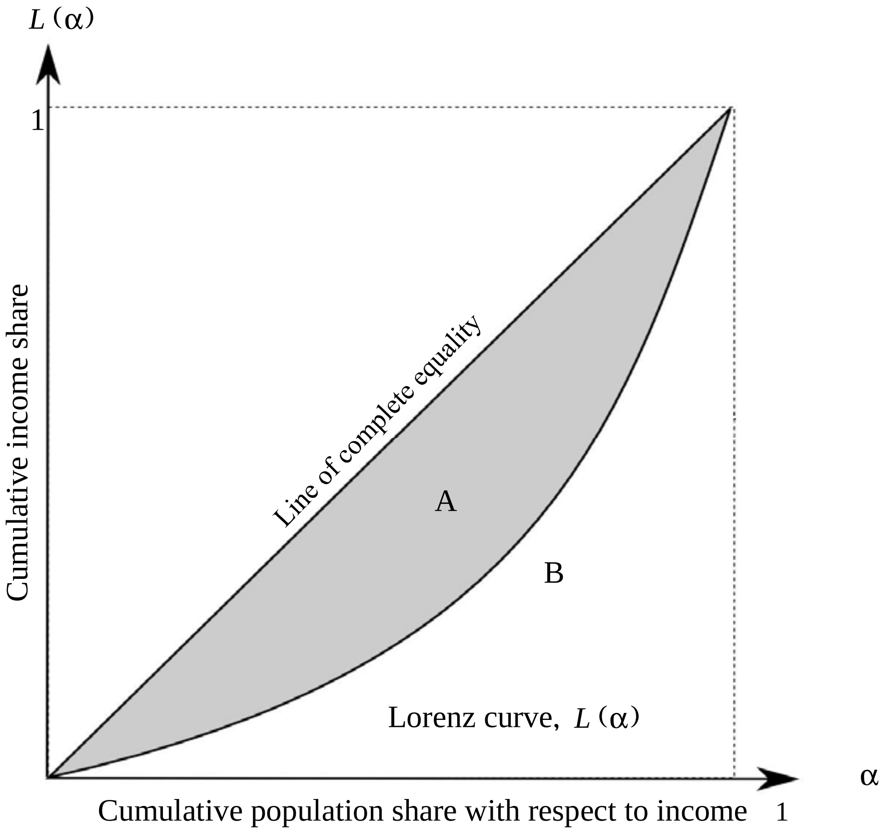

Figure 2.1 Gini coefficient, G, and Lorenz curve, $L(\alpha)$. $G = 2A$, $A + B = 1/2$

If a population U is finite, then the values of y_k will not be random and the Gini coefficient can be calculated as

$$G = \frac{2\sum_{k\in U} ky_k}{N\sum_{k\in U} y_k} - \frac{N+1}{N},$$

where the values of y_k are sorted by rank. For a sample, the Gini coefficient can be estimated as

$$\begin{aligned}\hat{G} &= \frac{2}{\hat{N}\hat{Y}}\sum_{k\in S} w_k \hat{N}_k y_k - \left(1 + \frac{1}{\hat{N}\hat{Y}}\sum_{k\in S} w_k^2 y_k\right)\\ &= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell \left|y_k - y_\ell\right|}{2\hat{N}\hat{Y}},\end{aligned}$$

where $\hat{N}_k = \sum_{\ell\in S} w_\ell \mathbf{1}_{[y_\ell \le y_k]}$ is the sum of the weights w_k, $\hat{Y} = \sum_{k\in S} w_k y_k$ is the estimated total income of the population, and $\hat{N} = \sum_{k\in S} w_k$ is the estimated size of the population. The expression can be simplified as follows if all the weights are equal to N/n :

$$\hat{G} = \frac{2\sum_{k\in S} ky_k}{n\sum_{k\in S} y_k} - \frac{n+1}{n}.$$

Note that the definition may vary by a factor of $n/(n-1)$ depending on the author (Osier 2009 and Eurostat 2004b); however, this subtlety becomes negligible if the sample is large enough.

Langel and Tillé (2012) combine the various approaches to obtain the same estimated linearized variable of the Gini coefficient for the sample:

$$\hat{z}_k^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}}\left[2\hat{N}_k(y_k - \hat{\bar{Y}}_k) + \hat{Y} - \hat{N}y_k - \hat{G}(\hat{Y} + y_k\hat{N})\right],$$

where $\hat{\bar{Y}}_k = \sum_{\ell=1}^{k} w_\ell y_\ell / \hat{N}_k$, and the values of y_ℓ are sorted and distinct.

2.2 Quintile Share Ratio (QSR or S_{80}/S_{20})

A good overview of this indicator is provided by Langel and Tillé (2012). Let q_{80} and q_{20} be the 80th and 20th percentiles of the distribution function $\mathrm{F}(y)$. The QSR is the ratio of the total income of the 20% of the population with the highest income to the total income of the 20% of the population with the lowest income. In the continuous case, the QSR can be defined as

$$\text{QSR} = \frac{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} > q_{80})}{\mathrm{E}(\mathcal{Y} \mid \mathcal{Y} < q_{20})} = \frac{1 - L(0.8)}{L(0.2)},$$

where $\mathcal{Y}$ is a random variable representing income. For finite populations, the QSR can be expressed and estimated for a sample on the basis of partial sums,

$$\widehat{\text{QSR}} = \frac{\hat{Y} - \hat{Y}_{0.8}}{\hat{Y}_{0.2}},$$

where, given the results obtained by Langel and Tillé (2011), we will use the following definition of the partial sum, which differs slightly from the official definition of Eurostat (2004a):

$$\hat{Y}_\alpha = \sum_{k \in S} w_k y_k H\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right), \tag{2.2}$$

with

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x < 1 \\ 1 & \text{if } x \ge 1. \end{cases}$$

To obtain the linearized variable of the QSR, we must first calculate the linearized variable of the partial sum (2.2), which is

$$I(Y_\alpha)_k = y_k H(\alpha N - k + 1) + \left[\alpha - \mathbf{1}_{[y_k < Q_\alpha]}\right] Q_\alpha,$$

where $Q_\alpha = y_i$, with $\hat{N}_{i-1} < \alpha\hat{N} \le \hat{N}_i$, corresponds to the first definition of the quantile of a finite population in the article by Hyndman and Fan (1996). Osier (2009) obtains a linearized variable that is dependent on the density of the variable $\mathcal{Y}$. However, Langel and Tillé (2011) have shown that the problem of estimating this density for the QSR can be avoided through a simplification, so that it is not necessary to make a kernel approximation of income density as proposed by Osier (2009).

The influence function of the QSR is dependent on the influence functions of the partial sums:

$$I\,(\mathrm{QSR})_k \;=\; z_k^{\mathrm{QSR}} \;=\; \frac{y_k - I\,(Y_{0.8})}{Y_{0.2}} - \frac{(Y - Y_{0.8})\,I\,(Y_{0.2})}{Y_{0.2}^2}.$$

By making the necessary substitutions, we can see that the estimated linearized variable for a sample is

$$\hat{z}_k^{\mathrm{QSR}} \;=\; \frac{y_k - \left\{y_k H\left(\frac{0.8\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.8}\left[0.8 - \mathbf{1}_{\left[y_k < \hat{Q}_{0.8}\right]}\right]\right\}}{\hat{Y}_{0.2}} - \frac{\left(\hat{Y} - \hat{Y}_{0.8}\right)\left\{y_k H\left(\frac{0.2\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.2}\left[0.2 - \mathbf{1}_{\left[y_k < \hat{Q}_{0.2}\right]}\right]\right\}}{\hat{Y}_{0.2}^2}. \tag{2.3}$$

2.3 Linearized variable of a quantile

Before we discuss poverty indicators, we should give a few details on the linearized variable of an α-order quantile, which can be expressed as

$$\hat{z}_k^{Q_\alpha} \;=\; -\frac{1}{f\,(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[\mathbf{1}_{\left[y_k \le \hat{Q}_\alpha\right]} - \alpha\right],$$

where the weighted quantile can be defined in a manner similar to the partial sum (2.2), and $f\,(\cdot)$ is an income density function that will be discussed in details in Section 3. Note that Eurostat (2004a) recommends the second definition by Hyndman and Fan (1996). We could dispute the Eurostat definition and use another quantile definition, for example $Q_\alpha = y_{k-1} + (y_k - y_{k-1})[\alpha N - (k-1)]$, where $\alpha N < k \le \alpha N + 1$, which is the fourth definition according to Hyndman and Fan (1996). We then estimate the quantile for a sample as follows:

$$\hat{Q}_\alpha \;=\; y_{k-1} + (y_k - y_{k-1})\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right).$$

The linearized variable of a quantile is dependent on the value of the income density function in that quantile. However, the actual income density is unknown and therefore must also be estimated using the sample. Deville (2000) and Osier (2009) suggest the use of Gaussian kernel estimation. We will discuss the problem of estimating f in more details in Section 3.

In addition to the problem of estimating the income density function, Croux (1998) shows that the empirical influence function of the median income is not a consistent estimator of the corresponding theoretical influence function. For a positive variable (such as income), the empirical influence function of the median income (the case discussed in Croux's article) converges toward an exponential distribution, the expectation of which is the influence function. It is not robust to large proportions of extreme values. It can be said to lack robustness in that the value of the estimator for a sample can differ greatly from the actual value for the population as a result of outliers (that is, values that are relatively very large) in the sample (see Hampel (1974) for a basic idea of robustness for infinite populations, and Beaumont, Haziza and Ruiz-Gazen (2013) for recent thoughts on this topic for finite population sampling).

2.4 Median income and at-risk-of-poverty threshold

Let $\hat{m} = \hat{Q}_{0.5}$ be the estimated median income of the sample. The At Risk of Poverty Threshold (ARPT) is defined as 60% of the median income:

$$\begin{aligned} \mathrm{ARPT} &= 0.6F^{-1}(0.5) \\ \widehat{\mathrm{ARPT}} &= 0.6\hat{Q}_{0.5} = 0.6\hat{m}. \end{aligned}$$

This is an absolute measure that is scale-dependent. The linearized variable of the ARPT is proportional to that of the median income:

$$\hat{z}_k^{\mathrm{ARPT}} = I(\mathrm{ARPT})_k = 0.6I(\mathrm{MED})_k = -\frac{0.6}{f(\hat{m})}\frac{1}{\hat{N}}\left[\mathbf{1}_{[y_k \le \hat{m}]} - 0.5\right].$$

2.5 At Risk of Poverty Rate

The At Risk of Poverty Rate (ARPR), where $\mathrm{ARPR} \in [0,1]$, is the share of the population with an income below the ARPT: $\mathrm{ARPR} = F(\mathrm{ARPT})$. The ARPR is scale-independent, like the Gini coefficient, QSR and relative median poverty gap (see Section 2.7). The official Eurostat definition (Eurostat 2004a) of the estimated ARPR for a sample is

$$\widehat{\mathrm{ARPR}} = \frac{\sum_{y_k < \widehat{\mathrm{ARPT}}} w_k}{\hat{N}}.$$

The linearized variable of the ARPR is defined by Osier (2009) as

$$\begin{aligned} \hat{z}_k^{\mathrm{ARPR}} &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\mathrm{ARPT}}]} - \widehat{\mathrm{ARPR}}\right) - \frac{f\left(\widehat{\mathrm{ARPT}}\right)}{f(\hat{m})}\frac{0.6}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}]} - 0.5\right) \\ &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \widehat{\mathrm{ARPT}}]} - \widehat{\mathrm{ARPR}}\right) + f\left(\widehat{\mathrm{ARPT}}\right)\hat{z}_k^{\mathrm{ARPT}}. \end{aligned}$$

Here, the income density function must be estimated at two points, namely the median income and the ARPT.

2.6 Median income of individuals below the ARPT

The median income of individuals below the ARPT is $m_p = F^{-1}(1/2\,F(\mathrm{ARPT}))$. It is estimated in the same way as any other quantile, the exact definition of which may vary. The linearized variable of m_p (Osier 2009) is dependent on that of the ARPR:

$$\hat{z}_k^{m_p} = \frac{1}{f(\hat{m}_p)}\frac{\hat{z}_k^{\mathrm{ARPR}}}{2} - \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k \le \hat{m}_p]} - F(\hat{m}_p)\right).$$

The estimated income density therefore appears three times, namely in the median income and ARPT for $\hat{z}_k^{\mathrm{ARPR}}$, and in the median income of individuals below the ARPT, m_p.

2.7 Relative Median Poverty Gap

The relative median poverty gap (RMPG) is the relative difference between the ARPT and the median income of individuals below the ARPT. $\mathrm{RMPG} = 0$ if the income of all “poor” individuals is equal to the ARPT, and $\mathrm{RMPG} = 1$ if the income of all these individuals is zero. The RMPG is a measure of the extent to which the “poor” individuals are poor:

$$\mathrm{RMPG} = \frac{\mathrm{ARPT} - m_p}{\mathrm{ARPT}}.$$

The estimated RMPG for a sample has already been described. The influence of each observation on the RMPG is defined by Osier (2009):

$$\hat{z}_k^{\mathrm{RMPG}} = \frac{\hat{m}_p \hat{z}_k^{\mathrm{ARPT}} - \widehat{\mathrm{ARPT}}\,\hat{z}_k^{m_p}}{\widehat{\mathrm{ARPT}}^2}.$$

The estimated income distribution density appears four times: once in the calculation of $\hat{z}_k^{\mathrm{ARPT}}$ and three times in the calculation of $\hat{z}_k^{m_p}$.

3 Estimating the income density function

In a design-based approach with a finite population, inference is made in relation to the sampling design $\mathrm{P}(S)$ used to select a sample S from a finite population U of size N. In this approach, only the sample inclusion indicators are random; all other quantities are fixed. The population income distribution function is then a step function: $F_y(x) = \sum_{k \in U} \mathbf{1}_{y_k \le x} / N$, and its derivative, the density function, does not exist due to discontinuities. If a model-based approach with a super-population model to justify the income density function term is not desired, then the distribution function must be artificially smoothed to make it differentiable. Therefore, our use of “density function” is not quite correct. For purposes of smoothing, Deville (2000) and Osier (2009) suggest using Gaussian kernel estimation to estimate the income density function:

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-u^2/2}, \qquad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k \in S} w_k K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \tag{3.1}$$

where h is the bandwidth that Osier estimates using $\hat{h} = \hat{\sigma}\hat{N}^{-0.2}$ and $\hat{\sigma}$ is the estimated standard deviation of the empirical income distribution:

$$\hat{\sigma} = \sqrt{\frac{\sum_{k \in S} w_k y_k^2}{\hat{N}} - \left(\frac{\sum_{k \in S} w_k y_k}{\hat{N}}\right)^2} = \sqrt{\frac{\sum_{k \in S} w_k y_k^2}{\hat{N}} - \bar{y}_w^2}.$$

Note that this estimate of σ is not robust, since it is very sensitive to the extreme values of y. Income data often have a distribution tail extending to the right with values that may be extremely high; these are "representative outliers" as defined by Chambers (1986) and Hulliger (1999). As the simulations in Section 4 will show, this can generate a strong bias in the variance estimates. Verma and Betti (2011) also use kernel estimation, recalling that Silverman (1986) states that the choice of kernel is not critical to ensure that $\hat{f}(y)$ converges toward $f(y)$, but the choice of bandwidth is. They use a value recommended by Silverman for distributions with a positive skewness coefficient, $h = 0.79(\hat{Q}_{75} - \hat{Q}_{25})\hat{N}^{-0.2}$. In their findings, they point out that the linearization method may be problematic because of irregularities in the empirical density function. They also state that these problems are all the more cause for concern because survey data often contain groups of observations with the same value (due to rounding or range questions), which can make estimating the density more complicated. The rest of this article describes the solutions we are proposing to reduce bias in variance estimates.

3.1 Using the logarithm

One solution that produces very good results, as shown below, is simply to use the logarithm to estimate the density of x. If $v = \log(x + a)$, where x is the income and a is a positive real number equals to, say, $(|\min_k(y_k)| + 1)$ where there may be negative or zero incomes (ignoring that a would be estimated), then

$$F_v(v) = \mathrm{P}(\mathcal{V} \le v) = \mathrm{P}(\log(\mathcal{Y} + a) \le v) = \mathrm{P}(\mathcal{Y} \le e^v - a) = F_y(e^v - a),$$

where $\mathcal{V}$ and $\mathcal{Y}$ would be random variables. Therefore,

$$f_v(v) = \frac{dF_v(v)}{dv} = \frac{dF_y(e^v - a)}{dv} = f_y(e^v - a)e^v.$$

That is, $f_v(v) = f_y(x)(x + a)$, which gives us the following estimator for the density of x:

$$\hat{f}_2(x) = \frac{\hat{f}_v(v)}{x + a} = \frac{\hat{f}_y(\log(x + a))}{x + a}. \tag{3.2}$$

The estimated density at x for $\mathcal{Y}$ can therefore be determined by estimating the density of the logarithm of the variable divided by the value of the variable at a given point. This property is valid for finite populations. Using the logarithm has the advantage of reducing the leveraging effect of large income values in the kernel density approximation calculation. Simulations show that this simple method significantly reduces bias.

3.2 Nearest neighbour with minimum bandwidth

Deville (2000) outlines another density estimation method that is a "nearest neighbour" method (see Silverman 1986) using the kernel

$$K_D(u) = \begin{cases} \dfrac{1}{b - a} & \text{if } a \le u < b \\ 0 & \text{otherwise,} \end{cases},$$

where $u = y_k$ and the choice of a and b, with $x \in [a, b]$, is to be determined and could depend on x. The distance $(b - a)$ represents the bandwidth h. The density estimate would therefore be

$$\begin{aligned} \hat{f}_D(x, a, b) &= \frac{1}{\hat{N}} \sum_{k \in S} K_D(y_k) \\ &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b - a} \mathbf{1}_{y_k \in [a, b[} \\ &= \frac{\hat{F}_y(b) - \hat{F}_y(a)}{b - a}, x \in [a, b[\end{aligned} \tag{3.3}$$

where $\hat{F}_y(x) = \sum_{k \in S} w_k \mathbf{1}_{y_k \le x} \big/ \hat{N}$.

Note that the density estimate (3.3) is not a continuous function and would not be suitable for estimating density values at the end tails of the distribution. Since our work relies little on distribution tails, we shall consider this approach as an option.

Our second proposal for estimating the density of x is based on the idea above. It is a nearest neighbour method, but also imposes a minimum bandwidth. Specifically, our method requires the use of at least p observations nearest to point x with minimum bandwidth $h(p) \ge h_{\text{opt}}$ where

$$h_{\text{opt}} = \frac{0.9 \min(\hat{\sigma}, \hat{Q}_{75} - \hat{Q}_{25})}{1.34 \sqrt[5]{\hat{N}}}$$

is the rule of thumb (Silverman 1986) for determining the bandwidth. This is also the default bandwidth value in the R function `density`. This solution is more robust than (3.1) and avoids the problems that arise when a number of values y_k are very close to each other, which is often the case because the individuals interviewed tend to round their income.

As the values $y_k, k = 1,...,n,$ are assumed to be ordered by rank, the width $h(p)$ of the window around x is initially determined by the p nearest observations, where $p \ll n.$ In the simulations discussed in the next section, after various trials, p was initially set at 30. The density of x is imputed to be the estimated density at the nearest observed point y_j that is less than or equal to $x,$ that is, $j = \max(k \mid y_k \leq x), k = 1,...,n.$ The bandwidth at x in fact depends on the p_j nearest observations around $y_j,$ with $p_j \geq p,$ which will be denoted $h(p_j)$ in the rest of this article. The density is therefore estimated only at observed points, with no smoothing or interpolation between the values $\hat{f}(y_j).$ The algorithm for estimating $\hat{f}(y_j)$ is as follows (see also Figure 3.1):

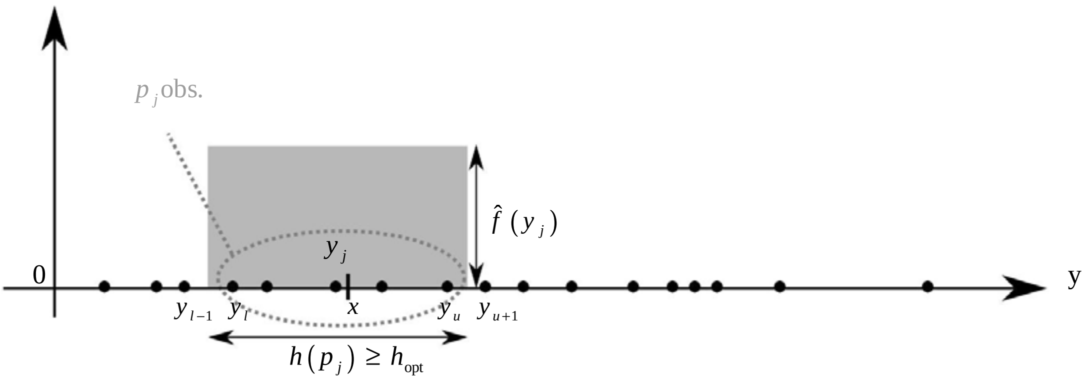

Figure 3.1 Window width $h(p_j)$

1. The initial width of the window around point $y_j,$ where $p_j = p,$ is defined as

$$h(p_j) = \frac{y_u + y_{u+1}}{2} - \frac{y_\ell + y_{\ell-1}}{2}; \; u = \begin{cases} j + p_j/2 - 1 & \text{if } p_j \text{ is even} \\ j + \lfloor p_j/2 \rfloor & \text{if } p_j \text{ is odd} \end{cases}$$
$$\ell = j - \lfloor p_j/2 \rfloor.$$

2. If the width of the resulting window $h(p_j)$ is less than h_{opt}, increment the two bounds:

   upper bound: $u \rightarrow u + 1,$ as long as $u < n,$

   lower bound: $l \rightarrow l - 1,$ as long as $l > 1,$

   which implies that $p_j \rightarrow p_j + 2,$ unless $u = n$ or $l = 1,$ in which case there is no longer the same number of points on each side of $y_j.$

3. Repeat step 2 until $h(p_j) \geq h_{\text{opt}}.$

4. The estimated density at x can then be written as

$$\hat{f}(x) = \hat{f}(y_j) = \begin{cases} \dfrac{p_j}{nh(p_j)} & \text{without weighting,} \\ \dfrac{\sum\limits_{p_j \text{ closest to } y_j} w_j^{\text{std}}}{nh(p_j)} & \text{with weighting,} \end{cases}$$

with standardized weights $w_k^{\text{std}} = w_k / \bar{w}$, $k = 1, \ldots, n$.

The number of observations p_j used in the calculation may vary, and it depends on the local curvature of the empirical distribution function. The condition $h(p_j) \geq h_{opt}$ guarantees a minimum window width in places where numerous observations would be concentrated over a small interval. The procedure is made even more solid by combining this approach with the preceding approach, that is, by estimating the density of the logarithm of the variable divided by its (non-logarithmic) value:

$$\hat{f}_3(x) = \frac{\hat{f}(\log(x + a))}{x + a}. \tag{3.4}$$

3.3 Robustness of the linearized variable

As stated above, for the median or for other quantiles, Croux (1998) points out that the empirical influence function or the linearized variable estimated using the sample is not as robust as it appears to be, even if the density function is known. We confirmed this for the EU-SILC data used in the model simulations with a Generalized Beta distribution of the second kind (GB2) by means of the R function `profml.gb2` (Graf and Nedyalkova 2011). For small samples $(n \leq 100)$, the potential bias of the linearized variable resulting from too many outliers may also bias the variance estimate calculated using the linearized variable. For larger samples $(n \geq 1{,}000)$, a maximum relative bias in the variance estimated using the empirical *versus* theoretical linearized variable may reach 5%. However, it is below the percentage in absolute terms three times out of four.

4 Results

Simulations were conducted on three sets of real data to compare and assess the different density function estimation methods, $\hat{f}_1(x)$, see (3.1), $\hat{f}_2(x)$, see (3.2) and $\hat{f}_3(x)$, see (3.4). These methods are required to estimate the variance of certain poverty and inequality indicators.

1. The first dataset contains equivalent household incomes from the EU-SILC survey conducted by the Swiss Federal Statistical Office in 2009. It includes 17,534 individuals with a non-zero income.
2. The second dataset also comes from the 2009 EU-SILC survey, but is limited to salaried individuals. It contains salaries from the register of the Central Compensation Office that has

been linked with the survey respondents. We therefore have no non-response issues, and there are 7,922 individuals with a non-zero income.

3. The third test file, named *Ilocos*, comes with the R package `ineq` (Zeileis 2012). It contains 632 observations, which are household incomes in Ilocos, one of the 16 regions of the Philippines. The data come from two surveys by the National Statistics Office of the Philippines, in 1997 and in 1998.

The three datasets have a positive skewness coefficient, which is typical of income distributions. Each data set is considered to be one population, and we initially selected 10,000 simple random samples without replacement of various sizes. The values of the various indicators were calculated for each sample, giving us a Monte Carlo estimate of their variance, $\text{var}_{\text{sim}}(\hat{\theta})$, for a poverty or inequality indicator θ. The variance estimator using linearization is denoted $\widehat{\text{var}}_{\text{lin}}(\hat{\theta})$ and is calculated using the linearization variable $\hat{z}^{\hat{\theta}}$ estimated for each sample:

$$\widehat{\text{var}}_{\text{lin}}(\hat{\theta}) = \frac{N(N-n)}{n}\text{var}\left(\hat{z}_S^{\hat{\theta}}\right),$$

where n is the size of the sample used for the simulations and

$$\text{var}\left(\hat{z}_S^{\hat{\theta}}\right) = \frac{1}{n-1}\sum_{k\in S}\left(\hat{z}_{S,k}^{\hat{\theta}} - \bar{z}_S^{\hat{\theta}}\right)$$

where $\bar{z}_S^{\hat{\theta}} = n^{-1}\sum_S \hat{z}_{S,k}^{\hat{\theta}}$, see (2.1).

The quality of the variance estimator using linearization is assessed by comparing the expected Monte Carlo value of the variance estimated using linearization, denoted $E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right]$, with the "true" Monte Carlo variance $\text{var}_{\text{sim}}(\hat{\theta})$ in terms of relative bias:

$$\text{RB}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] = \frac{E_{\text{sim}}\left[\widehat{\text{var}}_{\text{lin}}(\hat{\theta})\right] - \text{var}_{\text{sim}}(\hat{\theta})}{\text{var}_{\text{sim}}(\hat{\theta})}. \tag{4.1}$$

For the second data set (EU-SILC 2009, income of salaried individuals) we also, in a second step, selected 10,000 random samples without replacement under a stratified sampling design, and then calibrated the sampling weights to agree with the eight known sociodemographic marginal totals for the population of 7,922 individuals. The five strata used correspond to the age groups of the salaried individuals (see Table 4.1).

The eight calibration cells were obtained by crossing the three following dichotomous variables (auxiliary calibration variables):

1. MARIÉ, which indicates whether or not the individual is married;
2. CHEF, which indicates whether or not the individual's job is a management position; and
3. HOMME, which indicates the individual's sex.

The totals for the population of 7,922 individuals for these calibration cells are shown in Table 4.2.

Table 4.1
Strata used in simulations with 2009 EU-SILC data and three sample sizes (income of salaried individuals, $N = 7,922$)

Stratum h	Description	N_h	%	n_h		
1	individuals under 25	1,187	15.0	75	112	150
2	26- to 35-year-olds	1,359	17.2	86	129	171
3	36- to 45-year-olds	2,137	27.0	135	202	270
4	46- to 55-year-olds	1,864	23.5	117	177	235
5	individuals over 55	1,375	17.4	87	130	174
	TOTAL	7,922	100.0	500	750	1,000

Table 4.2
Calibration margins in simulations with 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

Margin	MARIÉ	CHEF	HOMME	Population total	%
1	0	0	0	1,487	18.8
2	0	0	1	1,208	15.2
3	0	1	0	323	4.1
4	0	1	1	457	5.8
5	1	0	0	1,759	22.2
6	1	0	1	1,278	16.1
7	1	1	0	328	4.1
8	1	1	1	1,082	13.7
			TOTAL	7,922	100.0

For each stratified sample, a calibration (linear method) was performed to make the sums of the weights agree with the eight margins shown above. Point estimates of the indicators and their linearized variable were computed for each sample using the calibrated weights.

Variance was estimated using the method developed by Deville (2000), which consists of linearizing also with respect to the calibration by calculating the residuals $e^{\hat{\theta}}$ of the regression (weighted by the sampling weights) of the linearized variables of the indicators for the auxiliary calibration variables. The variance of the total of the residuals thus calculated, under a stratified random sampling plan without replacement is therefore an estimator of the variance of the estimated indicator; it is the quantity of interest:

$$\widehat{\text{var}}_{\text{lin}}(\hat{\theta}) = \sum_{h=1}^{H} \frac{N_h}{n_h}(N_h - n_h)\, s^2_{e_h^{\hat{\theta}}} \tag{4.2}$$

where

$$s^2_{e_h^{\hat{\theta}}} = \frac{1}{n_h - 1} \sum_{k \in S_h} \left(e_k^{\hat{\theta}} - \bar{e}^{\hat{\theta}}\right)^2$$

The quality of the variance estimator using linearization is assessed analogously to the procedure for simple random sampling, see (4.1).

Tables 4.3, 4.4 and 4.5 show the relative bias of the variance for the three data sets used and described above, using simple random sampling. Table 4.6 shows the relative bias of the variance using stratified random sampling with calibrated weights. The upper portions of the tables give the values for the Gini coefficient and QSR, which do not require estimating the income density function. The estimation of their variance works well. Note that there is a problem involving the underestimation of the variance of the Gini coefficient in the case of stratification with calibration (Table 4.6).

For the first data set, Table 4.3 does not reveal any major differences except that the estimation of income density using $\hat{f}_3(x)$ gives results that are more conservative. In fact, the relative bias remains of the same order of magnitude, but positive, while it is negative for the other two methods of estimating density. For the second data set, Table 4.4 shows that it is essential to use the logarithm or the nearest neighbour method with minimum bandwidth. The latter, all relative bias falls under 10% when the sample sizes are sufficiently large (see last column in the table). Simulations on the same data with a stratified sampling plan and calibration strengthen and confirm these results (see Table 4.6). For the third data set, Table 4.5 shows the same trends, although the results are less stable as a result of the small sample and population sizes. This is not surprising, since the minimum number of neighbours to consider is fixed at 30. In this case, for the Ilocos data set, simulations with a smaller value of p fixed at 10 makes no difference ultimately, because the condition $h(p_j) \geq h_{\text{opt}}$ automatically increases it above 30.

Furthermore, generally speaking, we can see that the greater the use of Gaussian kernel density estimation - $\hat{f}_1(x)$ - the greater the error. In fact, the relative bias of the variance for the median income of individuals below the ARPT and for the RMPG are almost systematically greater in absolute value that those for the other indicators. For the RMPG, the error may be offset (as in Table 4.3) if there are enough observations, since the density estimation appears in both the numerator and the denominator.

Table 4.3
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (equivalent household income, $N = 17,534$)

Indicator	Sample size (sampling rate) $n = 500\,(2.9\ \%)$			$n = 750\,(4.3\ \%)$			$n = 1,000\,(5.7\ \%)$		
GINI	-0.02			-0.02			-0.02		
QSR	0.01			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.08	-0.06	0.04	-0.09	-0.07	0.03	-0.09	-0.07	0.04
ARPR	-0.05	-0.01	-0.00	-0.09	-0.06	-0.05	-0.08	-0.05	-0.03
RMPG	-0.09	-0.07	*0.15*	*-0.10*	-0.07	*0.12*	-0.09	-0.06	*0.14*
MEDP	*-0.16*	*-0.12*	0.09	*-0.19*	*-0.13*	0.05	*-0.18*	*-0.11*	0.07
MED	-0.08	-0.06	0.05	-0.08	-0.06	0.04	-0.08	-0.06	0.04

Table 4.4
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

Indicator	Sample size (sampling rate) $n = 500\,(6.3\ \%)$			$n = 750\,(9.5\ \%)$			$n = 1,000\,(12.6\ \%)$		
GINI	-0.03			-0.03			-0.02		
QSR	-0.00			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0.07	0.05	*0.13*	0.06	0.04	*0.10*	0.06	0.03	0.08
ARPR	-0.05	-0.04	-0.02	-0.05	-0.04	-0.01	-0.06	-0.05	-0.02
RMPG	*0.61*	*0.12*	*0.15*	*0.60*	*0.11*	0.08	*0.59*	0.09	0.05
MEDP	*0.73*	*0.17*	*0.18*	*0.72*	*0.16*	*0.10*	*0.72*	*0.15*	0.07
MED	0.07	0.04	*0.13*	0.06	0.04	*0.10*	0.05	0.03	0.07

Table 4.5
Relative bias (4.1) of the variance obtained with 10,000 simple random samples without replacement from Ilocos data (household income, $N = 632$)

Indicator	Sample size (sampling rate) $n = 50\,(7.9\ \%)$			$n = 63\,(10.0\ \%)$		
GINI	*-0.16*			*-0.13*		
QSR	0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.05	-0.06	-0.01	-0.03	-0.03	-0.01
ARPR	*-0.31*	-0.01	*-0.12*	*-0.33*	-0.03	*-0.18*
RMPG	*1.55*	*0.83*	*0.26*	*1.54*	*0.16*	*0.39*
MEDP	*1.02*	*0.28*	*-0.26*	*1.05*	0.07	*-0.11*
MED	0.04	0.03	0.08	0.07	0.07	0.09

Table 4.6
Relative bias (4.1) of the variance obtained with 10,000 stratified random samples without replacement, with weights calibrated to eight sociodemographic margins, from the 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)

Indicator	Sample size (sampling rate) $n = 500\,(6.3\ \%)$			$n = 750\,(9.5\ \%)$			$n = 1,000\,(12.6\ \%)$		
GINI	*-0.21*			*-0.20*			*-0.20*		
QSR	-0.06			-0.06			-0.07		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.07	-0.09	-0.01	-0.08	*-0.10*	-0.04	-0.09	*-0.11*	-0.06
ARPR	*-0.10*	*-0.10*	-0.08	-0.07	-0.06	-0.05	-0.06	-0.06	-0.05
RMPG	*0.63*	*0.13*	*0.13*	*0.61*	*0.11*	0.08	*0.59*	*0.10*	0.04
MEDP	*0.71*	*0.16*	*0.15*	*0.68*	*0.13*	0.09	*0.66*	*0.12*	0.04
MED	-0.07	-0.09	-0.01	-0.08	*-0.10*	-0.04	-0.08	*-0.11*	-0.06

In short, we see that the variance can be overestimated $\left(\mathrm{RB}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] > 0\right)$ or underestimated $\left(\mathrm{RB}\left[\widehat{\mathrm{var}}_{\mathrm{lin}}(\hat{\theta})\right] < 0\right)$ depending on the indicator and the data set. The use of the logarithm $\left(\hat{f}_2(x)\right)$ provides significant improvement. The nearest neighbour method $\left(\hat{f}_3(x)\right)$ eliminates all problems if there is enough data (as in Tables 4.3, 4.4 and 4.6). Slight problems arise with this method when the samples are small (as in Table 4.5). Illogical variations and bias that persist in the tables may also be the result of a lack of robustness in the linearized variables for certain samples, as stated in Section 3.3.

5 Conclusions

In a number of countries, national sample surveys publish extrapolated values for the Laeken indicators (Eurostat 2005), since they are key indicators that make it possible to direct decision makers with regard to political and social matters. It is therefore critical that we be able to quantify the precision of these measures, which raises the issue of the appropriateness of the precision estimates available. This article shows that a substantial improvement may be made in the precision estimates for poverty and inequality indicators by using a (local) estimate of the income density or given monetary variable.

The simulations conducted show that the Gaussian kernel density estimation method currently implemented in most cases is not recommended without at least using the logarithm as proposed in

Section 3.1; otherwise, there may be significant bias in the estimated variance. The nearest neighbour method (Section 3.2), which also imposes a minimum bandwidth, may yield even better results, especially if there are agglomerations of observations with certain values in the given data. However, this method requires setting a minimum number p of neighbours on the basis of the data used. If few observations are available, the use of the logarithm is preferable instead. In all cases, we hope that this work will help raise awareness of the importance of being meticulous during the implementation of calculations for the linearized variable of any indicator involving quantiles.

6 Acknowledgments

This work was carried out under a collaboration agreement between the Institute of Statistics at the University of Neuchâtel and the Swiss Federal Statistical Office (SFSO). We would also like to specifically thank the Income, Consumption and Living Conditions unit of the SFSO for providing us with data from the Swiss portion of the EU-SILC survey. Thanks also to Matti Langel and Anne Massiani for their support during our investigations.

References

Antal, E., Langel, M. and Tilllé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. *Proceedings 58th World Statistical Congress*, Dublin.

Ardilly, P., and Osier, G. (2007). Cross-sectional variance estimation for the French “Labour Force Survey”. *Survey Research Methods*, 1, 75-83.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters*, 37, 331-340.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 1, 17-27.

Deville, J.-C. (2000). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 219-230.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc; the gender pay gap. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04.

Eurostat (2004b). Theoretical study of the gini index. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04.

Eurostat (2005). The continuity of indicators during the transition between ECHP and EU-SILC. Working and study documents of the Office for Official Publications of the European Communities, Luxembourg.

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, 96, 691-709.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis, Theory and Applications*, (Ed., V. Pawlosky-Glahn and A. Buccianti), Oxford: Wiley, chapter 9, 114-127.

Graf, M. (2013). A simplified approach to linerarization variance for surveys. *University of Neuchâtel*, working document.

Graf, M., and Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 54-63.

Hyndman, R.J., and Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician*, 50, 361-365.

Kovačević, M.S., and Binder, D.A. (1997). Variance estimation for measures of income inequality and polarization - The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Langel, M., and Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Langel, M., and Tillé, Y. (2012). Variance estimation of the Gini index: Revisiting a result several times published. In *Press in Journal of the Royal Statistical Society - Series A*.

Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *American Statistical Association*, 9, 209-219.

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

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Londres.

Verma, V., and Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics*, 38, 1549-1576.

Zeileis, A. (2012). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-10.