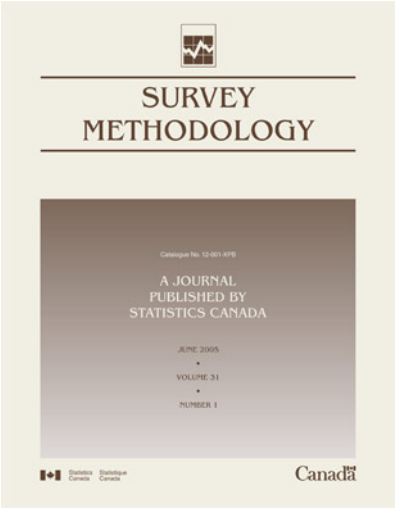
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. 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.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

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 cooperation and goodwill.

Vol. 32, No. 1, pp. 37-52
Statistics Canada, Catalogue No. 12-001

On Calibration Estimation for Quantiles

Torsten Harms and Pierre Duchesne [1]

Abstract

In this paper, we consider the estimation of quantiles using the calibration paradigm. The proposed methodology relies on an approach similar to the one leading to the original calibration estimators of Deville and Särndal (1992). An appealing property of the new methodology is that it is not necessary to know the values of the auxiliary variables for all units in the population. It suffices instead to know the corresponding quantiles for the auxiliary variables. When the quadratic metric is adopted, an analytic representation of the calibration weights is obtained. In this situation, the weights are similar to those leading to the generalized regression (GREG) estimator. Variance estimation and construction of confidence intervals are discussed. In a small simulation study, a calibration estimator is compared to other popular estimators for quantiles that also make use of auxiliary information.

Key Words: Calibration estimators; Quantiles; Ratio estimators; Difference estimators.

1. Introduction

In recent years, considerable attention has been given to the estimation of population distribution functions in the context of survey sampling. A particular target of this attention has been the median, which is often regarded as a more satisfactory location measure than the mean, especially when the variable of interest follows a skewed distribution. Traditional estimators of population means or totals can be usually substantially improved if relevant auxiliary information is made available. Consequently, the use of such auxiliary Information seems highly desirable in sample quantile estimators.

Using a model-based approach, Chambers and Dunstan (1986) considered quantile estimators based on an estimator of the distribution function which do incorporate auxiliary information. Rao, Kovar and Mantel (1990) have proposed design-based alternatives to the model-based approach. They used simulation experiments to compare two quantile estimators, based on ratio and difference estimators, to the simple design-based estimator which makes no use of the auxiliary information. It should be noted that neither of the two design-based proposals requires knowledge of the auxiliary information for each unit in the population; it rather suffices to know only the corresponding quantiles. While the model-based estimator proposed by Chambers and Dunstan (1986) can be more efficient than its design-based alternative if the model is correctly specified, Rao *et al.* (1990) have pointed out the advantage of the design-based estimators under model misspecification. Chambers, Dorfman and Hall (1992) have compared these two estimators theoretically with respect to their consistency, asymptotic bias and variance under a population model. Their main conclusion is that neither of the two methods is a sharp winner. Dorfman (1993) has reevaluated the simulation results obtained by Rao *et al.* (1990) and proposed a modified version of their methodology, using model-based arguments. Variance estimators in the model-based approach of Chambers and Dunstan (1986) and the design-based estimators of Rao *et al.* (1990) are discussed in Wu and Sitter (2001).

Other related works on quantile and median estimators include that of Kuk (1988) who proposes quantile estimators under pps (*proportional to size*) sampling and that of Kuk and Mak (1989) who use a method that is based on cross-classifying the individuals in the sample, according to the variable of interest and a single auxiliary variable. Meeden (1995) takes a different approach to construct a median estimator based on univariate auxiliary information, using the Bayesian concept of Polya sampling to impute all the target variable's unknown population values via a ratio-based approach. Rueda, Arcos and Martínez (2003) have recently built quantile estimators that extend ratio, difference and regression estimators in ways similar to those developed for the population mean.

In this paper, we follow the concept of calibration which was first introduced by Deville (1988) in order to derive a quantile estimator. The calibration approach has gained popularity in real applications, because the resulting estimators are easy to interpret and to motivate, relying, as they do, on sampling weights and natural calibration constraints. This approach was developed in the seminal work of Deville and Särndal (1992) as an alternative means of incorporating auxiliary information in the estimation of population totals. The so-called calibrated weights are found by minimizing a distance measure between the sampling weights and the new weights, which need to satisfy certain calibration constraints. For estimating totals the calibrated weights replace

1. Torsten Harms and Pierre Duchesne, Université de Montréal, Département de mathématiques et de statistique, CP 6128 Succursale Centre-Ville, Montréal, Québec, H3C 3J7, Canada. E-mail: duchesne@dms.umontreal.ca.

the original design weights used in Horvitz-Thompson type estimators. When the new weights are applied to the auxiliary variables available in the sample, they reproduce the known population totals of the auxiliary variables exactly; it is for this reason that the estimators in this class are called calibration estimators. See also Singh and Mohl (1996) who provide simple justifications of calibration estimators. They also present a very general and unifying treatment of calibration methods whose weights satisfy certain range restrictions and benchmark constraints.

Our fundamental aim is to propose calibration estimators for quantiles which are as easy to implement and interpret as the calibration estimators for totals developed by Deville and Särndal (1992). When compared to the quantile estimators available in the literature, the new calibration estimators should also be competitive with respect to their bias, variance, and coverage rates of the confidence intervals. Early calibration estimators for distribution functions and quantiles include those proposed by Kovačević (1997), who considered estimators of the distribution function calibrated on moments of the auxiliary variables. Harms (2003) has investigated a similar approach, with applications to the Finnish European Household Panel survey. Ren (2002) appears to have been the first to develop a unifying treatment of calibration estimators for distribution functions and quantiles. The calibration estimators for quantiles presented in this paper continue the work initiated by Ren (2002). We adhere to the original calibration paradigm for totals as closely as possible: when the parameter of interest is a total, it seems natural to calibrate on totals of the auxiliary variables. In the present context, since the parameter of interest corresponds to a quantile, the calibration constraints require that the weights are such that the sample quantile estimators of the auxiliary variables and their corresponding population quantiles are equal. In other words, the weighted quantile estimators for the auxiliary variables should yield exactly the population quantiles, which are assumed to be known. We present arguments which justify calibrating on quantiles, whenever the parameter of interest is itself a quantile. Interestingly, our methodology does not necessitate knowledge of the values of the auxiliary variables for all units in the population. Since the resulting estimators display a structural form very similar to the original calibration estimators for totals, it is expected that, under general conditions, the proposed estimators for quantiles will be asymptotically design-unbiased. Furthermore, these similarities allow us to derive variance estimators which admit a familiar form. Contrary to some of the other estimators, the proposed approach is also applicable to vectorial auxiliary variables (that is, when several auxiliary variables are available), while requiring only minimal auxiliary information. However, some restricttions may apply when the sample is highly unrepresentative of the sampled population or when the quantiles being estimated are very close to the population minimum or maximum. Note that highly unrepresentative samples can also cause problems for calibration estimators for totals commonly used; in such situations, the algorithm for computing calibration estimators may fail to converge for many distance measures of practical interest.

The organization of the paper is as follows: In section 2, some preliminaries are given, including a brief review of the calibration estimators for totals. The new calibration estimators for quantiles are developed in section 3.1. The standard distribution function can be interpreted as a Horvitz-Thompson estimator, providing a possible approach to the construction of a calibrated distribution function estimator. Quantile estimators are then naturally derived by inverting the distribution function estimator (see *e.g.*, Ren (2002)). As in calibration estimators for totals, design weights can be replaced by more general sampling weights, in order to take account the auxiliary information. However, for many situations of practical interest, it may happen that no solution exists for the calibration constraints when this kind of distribution function estimator is adopted, the reason being that this estimator corresponds to a step function. In order to avoid existence problems of solutions for the calibration constraints, a new distribution function estimator is introduced, based on the natural concept of interpolation. Under the common quadratic metric, an analytic representtation of the calibration weights is provided in section 3.2; variance estimators and confidence intervals are discussed in section 3.3. A practical aspect involves evaluating the methodology proposed with real populations and several sampling plans. Consequently, in section 4, we present a small simulation study where we compare our new approach, with respect to variance, bias and coverage rates of the confidence intervals, with that of Chambers and Dunstan (1986) as well as with some of the estimators proposed by Rao *et al.* (1990). Finally, concluding remarks are offered in section 5.

2. Some Preliminaries on Calibration Estimators

In this section, we present the fundamental concepts and notations useful for the sequel. We also give a brief review of calibration estimators for totals.

Let $U = \{1, ..., k, ..., N\}$ be a finite population of size N. Let $T_y = \sum_U y_k$ be the population total of the variable of interest y, (note that for a set A, $A \subseteq U$, $\sum_A$ will be used as shorthand for $\sum_{k \in A}$). A sample $s \subset U$ of size n is drawn according to a sampling plan. Let $\pi_k = \Pr(s \ni k)$ and $\pi_{kl} = \Pr(s \ni k, l)$ be the first and second order inclusion probabilities, respectively. We denote the design

weights $d_k = \pi_k^{-1}$ and $\hat{T}_{y,\mathrm{HT}} = \sum_s d_k y_k$ represents the Horvitz-Thompson (HT) estimator of T_y.

Let $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$ be a vector of auxiliary variables associated with unit k, $k \in U$. Calibration estimators naturally include auxiliary information in the estimation. Let $s = \{k_1, ..., k_n\}$, $s \subset U$. Starting with the vector of original weights $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$, new weights are found which, when applied to the auxiliary variables available in s, make it possible to retrieve the known population totals for the J auxiliary variables $\mathbf{T}_\mathbf{x} = \sum_U \mathbf{x}_k = (T_{x_1}, ..., T_{x_J})'$. The calibration estimator for totals are more precisely defined in Definition 1.

Definition 1 (Calibration estimator for totals). *Let* $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$ *be the design weights. The calibration estimator for totals takes the form* $\hat{T}_{y,\mathrm{cal}} = \sum_s w_{ks} y_k$, *where the weights* w_{ks}, $k \in s$ *are obtained as the following minimization problem with respect to the variable* $\mathbf{v} = (v_{k_1}, ..., v_{k_n})'$:

$$\mathbf{w} = \mathrm{argmin}_\mathbf{v}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \tag{1}$$

subject to the calibration constraints $\sum_s v_k \mathbf{x}_k = \mathbf{T}_\mathbf{x}$, *where* $\mathrm{D}(\cdot,\cdot)$ *denotes the distance measure and* $\mathbf{w} = (w_{k_1}, ..., w_{k_n})'$ *corresponds to the vector of the calibrated weights.*

For notational simplicity, we write $w_k \equiv w_{ks}$ in Definition 1 when no confusion is possible. It is common practice to let $x_{1k} \equiv 1$, $\forall k \in U$, and consequently $T_{x_1} = N$. This means that the calibrated weights satisfy the natural constraint $\sum_s w_k = N$. Many distance functions D are available in the literature (see, *e.g.*, Deville and Särndal (1992), Chen and Qin (1993), Thompson (1997)). Consider the quadratic distance function

$$\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s \frac{(v_k - d_k)^2}{d_k q_k}, \tag{2}$$

where q_k determines the importance of the unit $k \in s$ in the calibration problem. Heteroscedasticity problems can be handled using an appropriate choice of the q_k's. Solving the optimization problem (1) using the Lagrange multiplier technique (see Deville and Särndal (1992), among others), the weights $w_k = d_k(1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}_s)$ are obtained, where $\boldsymbol{\lambda}_s = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\mathbf{T}_\mathbf{x} - \hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}})$ and $\hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}}$ denotes the HT-estimator of $\mathbf{T}_\mathbf{x}$. This choice of distance function leads to the weights of the well-known generalized regression estimator (GREG) of Cassel, Särndal and Wretman (1976), which is studied in detail in Särndal, Swensson and Wretman (1992). Under minimal requirements for the disance measure D, Deville and Särndal (1992) have shown that all calibration estimators in this class are asymptotically equivalent to the GREG. For ease of interpretation and other cosmetic reasons, some users may want to have positive weights or restrict them to a specific interval (see also Singh and Mohl (1996)). In practical applications, these numerical features of the weights seem to be the main motivation for an alternative choice of D.

3. New Calibration Estimators

In this section we develop calibration estimators for quantiles, using ideas similar to those leading to the caliration estimators for population totals, as described in section 2. The new calibration estimators for quantiles are introduced in the next subsection, using interpolated distribution function estimators. Then, special attention is devoted to the quadratic distance function. The last subection presents variance estimation and the construction of confidence intervals.

3.1 Definition of the Calibration Estimators for Quantiles

Let $\mathbf{Q}_{\mathbf{x},\alpha} = (Q_{x_1,\alpha}, ..., Q_{x_J,\alpha})'$ denote the known vector of population quantiles for the vector of auxiliary variables $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$, $k \in U$. The Heavyside function $H(z)$ is given by:

$$H(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

The population distribution function of a scalar auxiliary variable x is defined in the usual way as $F_x(t) = N^{-1}\sum_U H(t - x_k)$, and the population quantile $Q_{x,\alpha}$ is obtained by letting $Q_{x,\alpha} = \inf\{t \mid F_x(t) \geq \alpha\}$.

The vector $\mathbf{Q}_{\mathbf{x},\alpha}$ contains quantiles of the auxiliary variables, obtained from information in past surveys or from available administrative sources. For example, for skewed distributions which are rather common in business and economic surveys, it seems more natural to keep in the record files the population medians rather than population means; in this case it seems natural to assume the knowledge of $\mathbf{Q}_{\mathbf{x},0.5}$. This suggests that, using the same approach as the one leading to calibration for totals described in section 2, the proposed estimator for the population quantile $Q_{y,\alpha}$ of the variable of interest y, noted $\hat{Q}_{y,\mathrm{cal},\alpha}$, could be obtained by inverting a certain estimator of the distribution function (that we discuss below), subject to calibration constraints such as $\hat{Q}_{x_j,\mathrm{cal},\alpha} = Q_{x_j,\alpha}$, $j = 1, ..., J$. Following the usual interpretation, if the calibrated weights allow us to retrieve the known population quantiles of the auxiliary variables then, under certain conditions, they should produce reasonable estimators for the quantile of the variable of interest y.

More precisely, the calibrated weights are obtained by solving the following optimization problem:

$$\mathbf{w} = \mathrm{argmin}_\mathbf{v}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \tag{3}$$

subject to the calibration constraints $\sum_s v_k = N$ and $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, \ldots, \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$.

The estimators $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha}$ and $\hat{Q}_{y,\text{cal},\alpha}$ rely on the vector of weights $\mathbf{w}$, stemming from the solution of the calibration problem (3). To calculate these estimators for quantiles, we need to construct $w-$weighted estimators of the distribution function for variables $\mathbf{x}$ and y. Based on the sampling weights $\mathbf{d}$, a natural estimator of the sampling distribution function is given by

$$\tilde{F}_y(t) = \sum_s d_k H(t - y_k) \Big/ \sum_s d_k, \tag{4}$$

which provides a consistent estimator of $F_y(t)$. Similarly, $F_{x_j}(t)$ can be consistently estimated by $\tilde{F}_{x_j}(t) = \sum_s d_k H(t - x_{jk}) / \sum_s d_k$, $j = 1, \ldots, J$. A $w-$weighted distribution function estimator of $F_{x_j}(t)$ is given by

$$\tilde{F}_{x_j,\text{cal}}(t) = \sum_s w_k H(t - x_{jk}) \Big/ \sum_s w_k. \tag{5}$$

A similar formula holds for $\tilde{F}_{y,\text{cal}}(t)$. These $w-$weighted estimators are considered in Ren (2002). However, if one estimates $Q_{x_j,\alpha}$ by $\hat{Q}_{x_j,\alpha} = \inf\{t \mid \tilde{F}_{x_j}(t) \geq \alpha\}$, or makes a similar estimation using a $w-$weighted version, then it is generally not possible to reach an exact solution of the calibration problem (3). Indeed, if the previous definition is used to estimate the quantiles by inverting the distribution function using the previous definitions, then the constraints in the optimization problem (3) will not, in general, be fulfilled unless the sample s contains precisely a unit k such that $x_{jk} = Q_{x_j,\alpha}$. When J is large, this problem can be more pronounced. Furthermore, even if the sample does contain such a value, it is sometimes not possible to obtain the weights needed to minimize the distance function, the reason being that under certain circumstances, the weights fulfilling the calibration constraints form an open set, whereas the optimal weights lie precisely on the border of this set. The following example illustrates this situation.

Example 1:
Consider a population U of size $N = 30$, such that the population median of x is $Q_{x,0.5} = 2$. A sample s of size $n = 3$ is drawn, and suppose that $x_k = k, \forall k \in s = \{1, 2, 3\}$. For simplicity, the distance measure $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s (v_k - d_k)^2$ is adopted; it is supposed that the sampling weights are $(d_1, d_2, d_3) = (15, 9, 6)$. Based on (5), the calibration constraint is $\tilde{Q}_{x,\text{cal},0.5} = \inf\{t \mid \tilde{F}_{x,\text{cal}}(t) \geq 0.5\} = 2$, which implies that $\sum_s w_k H(2 - x_k) \geq 15$ and $\sum_s w_k H(1 - x_k) < 15$. Equivalently, $w_1 + w_2 \geq 15$ and $w_1 < 15$. Thus we have to choose w_1 of the form $w_1 = 15 - \epsilon$, for $\epsilon > 0$. In this case, since $w_1 + w_2 + w_3 = 30$, we have that $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \epsilon^2 + (w_2 - 9)^2 + (w_2 - 9 - \epsilon)^2$, leading to the optimal solution $(w_1, w_2, w_3) = (15 - \epsilon, 9 + \epsilon/2, 6 + \epsilon/2)$. Consequently, for these weights $\mathrm{D}(\mathbf{v}, \mathbf{d}) = 3\epsilon^2/2$, which is obviously minimized when $\epsilon \to 0$. However, the limit reduces to $\mathbf{w} = (w_1, w_2, w_3) = (15, 9, 6)$ with $\mathrm{D}(\mathbf{w}, \mathbf{d}) = 0$, but based on these weights $\tilde{Q}_{x,\text{cal},0.5} = 1 \neq Q_{x,0.5} = 2$.

However, these difficulties can be naturally avoided by considering a smooth estimator of the distribution function. For simplicity, we consider here a distribution function estimator calculated using a linear interpolation (another possibility is discussed in section 5), which is precisely defined in Definition 2.

Definition 2 (Interpolated distribution function estimators). *Define*

$$\hat{F}_{y,\text{cal}}(t) = \frac{\sum_s w_k H_{y,s}(t, y_k)}{\sum_s w_k}, \tag{6}$$

$$\hat{F}_{x_j,\text{cal}}(t) = \frac{\sum_s w_k H_{x_j,s}(t, x_{jk})}{\sum_s w_k}, \tag{7}$$

where the Heavyside function H in (4) *and* (5) *is replaced by the slightly modified function*

$$H_{y,s}(t, y_k) = \begin{cases} 1, & y_k \leq L_{y,s}(t), \\ \beta_{y,s}(t) & y_k = U_{y,s}(t), \\ 0, & y_k > U_{y,s}(t), \end{cases} \tag{8}$$

where $L_{y,s}(t) = \max\{\{y_k, k \in s \mid y_k \leq t\} \cup \{-\infty\}\}$, $U_{y,s}(t) = \min\{\{y_k, k \in s \mid y_k > t\} \cup \{\infty\}\}$ *and* $\beta_{y,s}(t) = \{t - L_{y,s}(t)\}/\{U_{y,s}(t) - L_{y,s}(t)\}$. *The function* $H_{x_j,s}(t, x_k)$ *is defined similarly. The estimators* (6) *and* (7), *based on the functions* $H_{y,s}(t, y_k)$ *and* $H_{x_j,s}(t, x_k)$, *are called interpolated distribution function estimators of* $F_y(t)$ *and* $F_{x_j}(t)$, *respectively*.

The various quantities in (8) have easy interpretations: $L_{y,s}$ and $U_{y,s}$ represent the lower and upper neighbors of t in the sampled values $y_k, k \in s$, and $\beta_{y,s}(t)$ denotes the linear interpolation coefficient between these two quantities. In particular, for all $t \in \{y_k, k \in s\}$ we have $H_{y,s}(t, y_k) = H(t - y_k)$. Consequently, the relations $\hat{F}_{y,\text{cal}}(t) = \tilde{F}_{y,\text{cal}}(t)$ are satisfied for all $t \in \{y_k, k \in s\}$. For all the other values of t, $\hat{F}_{y,\text{cal}}(t)$ consists of a linear interpolation between these quantities. In the following example, Example 1 is revisited using the interpolated distribution function estimator (7).

Example 2:
In Example 1, using the interpolated version (7), the constraints are now $w_1 + w_2 + w_3 = 30$ and $(w_1 + w_2)/(w_1 + w_2 + w_3) = 0.5$. Consequently $w_3 = 15$, $w_1 + w_2 = 15$. Simple algebra shows that the optimal solution is $(w_1, w_2, w_3) = (10.5, 4.5, 15)$, which is now well-defined.

With the interpolated distribution function estimators, $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ and $\hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$ are now well defined $\alpha-$quantile estimators for all $\alpha \in (0, 1)$, as long as one can assure that the weights w_k are all strictly positive. Letting $\hat{Q}_{x_j,\text{cal},\alpha} = \hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$, we define the proposed calibration estimator

$\hat{Q}_{y,\text{cal},\alpha}$ for the quantile $Q_{y,\alpha}$, using the interpolated distribution function estimator given in Definition 2.

Definition 3 (Calibration estimator for quantiles). *Consider the optimization problem* (3), *subject to the calibration constraints* $\sum_s v_k = N$ *and* $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, \ldots, \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$. *Solving this optimization problem and denoting the resulting weights as* $\mathbf{w}$, *the proposed calibration estimator for quantiles of* $Q_{y,\alpha}$ *is defined by*

$$\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha), \tag{9}$$

where $\hat{F}_{y,\text{cal}}(t)$ *is given by* (6).

One of the appealing properties of the proposed estimator (9) is that it yields exact population quantiles when the relationship between y and a scalar auxiliary variable x is exactly linear. Assume that $y_k = a + bx_k$ holds perfectly for all units $k \in U$ and suppose that the units in the sample s are such that $x_k < Q_{x,\alpha} < x_l$ for some units x_k and $x_l, k, l \in s$. For the calibrated estimator (9), we have that $\hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = \alpha$. We need to distinguish the two cases, $b > 0$ and $b < 0$ (The case $b = 0$ is trivial since y_k is then identically equal to a constant). Firstly, consider the situation $b > 0$. Since the linear relation $y_k = a + bx_k$ is satisfied for all units k and since $b > 0$, the following relations hold: $L_{y,s}(a + bt) = a + bL_{x,s}(t); U_{y,s}(a + bt) = a + bU_{x,s}(t)$ and $\beta_{y,s}(a + bt) = \beta_{x,s}(t)$. These relations lead to $H_{y,s}(a + bt, y_k) = H_{x,s}(t, x_k)$. It follows that $\hat{F}_{y,\text{cal}}(a + bt) = \hat{F}_{x,\text{cal}}(t)$. Furthermore, $\hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = \alpha$ and using the relation $a + bQ_{x,\alpha} = Q_{y,\alpha}$, we deduce that $\hat{F}_{y,\text{cal}}(Q_{y,\alpha}) = \alpha$. Consequently, when an exact linear relationship holds and $b > 0, \hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha) = Q_{y,\alpha}$. Secondly, consider the case $b < 0$. We deduce in this case the following relations: $L_{y,s}(a + bt) = a + bU_{x,s}(t); U_{y,s}(a + bt) = a + bL_{x,s}(t); \beta_{y,s}(a + bt) = 1 - \beta_{x,s}(t)$ and $H_{y,s}(a + bt, y_k) = 1 - H_{x,s}(t, x_k)$. Since $b < 0$, the relationship between the quantiles of x and y is given by $a + bQ_{x,\alpha} = Q_{y,1-\alpha}$. Then, we deduce that $\hat{F}_{y,\text{cal}}(Q_{y,1-\alpha}) = \hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = 1 - \hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = 1 - \alpha$. Thus, in this situation, $Q_{y,1-\alpha}$ is estimated exactly by $\hat{Q}_{y,\text{cal},1-\alpha}$. This means that, when an exact relation holds, if $b > 0$ the proposed calibration estimator $\hat{Q}_{y,\text{cal},\alpha}$ yields perfect estimators with zero bias and variance of $Q_{y,\alpha}$. On the other hand, if $b < 0$ and calibrating on $Q_{x,\alpha}, Q_{y,1-\alpha}$ is estimated exactly by $\hat{Q}_{y,\text{cal},1-\alpha}$ (which makes sense because the perfect linear relationship between x and y is such that the slope parameter is negative).

Note that when $\hat{F}_{y,\text{cal}}$ and $\hat{F}_{x_j,\text{cal}}$ are invertible at points $Q_{y,\alpha}$ and $Q_{x_j,\alpha}$, the calibration constraints in (3) can be rewritten in terms of the distribution functions, that is the calibration constraints based on the quantiles are equivalent to $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha, j = 1, \ldots, J$. This means that the original calibration problem can be alternatively written in terms of distribution functions with the above constraints.

A natural question arises as to the existence of a solution to the optimization problem (3). Even when formulated with the interpolated distribution functions, it is not always possible to find a solution to (3). For example, if $Q_{x_j,\alpha}$ is smaller or larger than all values x_{jk} in the sample s, then $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha})$ will equal zero or one regardless of the choice of the weights $\mathbf{w}$. Thus in these cases it may happen that the calibration constraints cannot be fulfilled. However, when the sample's behavior differs widely from that of the target population, one should keep a very critical eye on any adjustment, and this situation can be considered somewhat extreme. In practice, this rarely occurs unless α is chosen very close to zero or one. Note that it may be impossible to obtain a solution when the sample size n is small. In these situations, the sample minimum or maximum could serve as a possible estimator or we could resort to the simple design-based estimator of the distribution function.

The second potential problem is that some weights w_k might be negative. In this case $\hat{F}_{y,\text{cal}}$ is no longer bijective. This is not a problem as long as $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ is still uniquely determined. This problem can be avoided by restricting all the weights to be strictly positive, using an appropriate metric $D(\cdot,\cdot)$. This approach has been adopted by Kovačević (1997) (for more details on distance functions yielding positive weights, see also Deville and Särndal (1992) and Singh and Mohl (1996)).

Remark 1:
The proposed distribution functions estimators (6) and (7) rely on a linear interpolation. In a unified way, the population distribution function, which is a step function as well, could also be defined using a linear interpolation. In practice, the two definitions differ only slightly in behavior, if the population N is sufficiently large. However, it should be noted that if the population size N is relatively small, it might be worth using an interpolation to define distribution functions.

Remark 2:
In the optimization problem (3), we calibrated on a particular quantile. This approach could be extended by allowing to calibrate on a finite set of quantiles, if such information is available. More precisely, suppose that for an auxiliary variable x, the α_m – quantiles $Q_{x,\alpha_m}, m = 1, \ldots, M$ are known, where $M < n - 1$. In this case, we could consider the calibration constraints $\hat{F}_{x,\text{cal}}(Q_{x,\alpha_m}) = \alpha_m, m = 1, \ldots, M$ and solve the optimization problem (3) with these additional calibration constraints. Naturally, this information yields a more complete description of the distribution of the auxiliary variables; so the efficiency of the calibration estimators is expected to be higher.

Remark 3:
The proposed calibration estimator (9) is obtained by calibrating on population quantiles. Another possibility has been considered by Ren (2002) who calibrated on population moments, up to order m, of the same distribution. More precisely, Ren (2002) has proposed calibration estimators for quantiles satisfying constraints of the form $\sum_s w_k x_k^m = \sum_U x_k^m$, $m = 0, 1, ..., M$. Calibration on different moments of the same distribution is closely related to calibrating on different quantiles of the same variable, and all these constraints provide a more complete description of the distribution of the auxiliary variable. For other generalizations of the calibration paradigm on moments, see also Ren and Deville (2000) and Harms (2003).

3.2 Analytical Solution of the Calibrated Weights when D is the Quadratic Metric

When the quadratic distance function (2) is adopted, an explicit solution of the optimization problem (3) can be derived. This situation is similar to the calibration estimators for totals, where the weights of the GREG estimator are explicitly obtained under the metric (2). A careful analysis of the estimation problem for quantiles reveals important similarities, the reason being that the estimators given by (7) are weighted sums of the variables $\{H_{x_j,s}(t, x_{jk}),\ k \in s\}$, $j = 1, ..., J$. This is stated in Proposition 1.

Proposition 1 (Calibrated weights for the quadratic metric). *Consider the quadratic distance function* (2)*. The vector of weights* $\mathbf{w}$ *which solves the optimization problem* (3) *satisfies the relation*:

$$w_k = d_k(1 + q_k \mathbf{a}_k' \boldsymbol{\lambda}_s), k \in s, \qquad (10)$$

where the vector $\boldsymbol{\lambda}_s = (\lambda_0, ..., \lambda_J)'$ *is determined via the* $J+1$ *constraints as*:

$$\boldsymbol{\lambda}_s = \left(\sum_s d_k\ q_k\ \mathbf{a}_k\ \mathbf{a}_k'\right)^{-1}\left(\mathbf{T}_\mathbf{a} - \sum_s d_k\ \mathbf{a}_k\right), \qquad (11)$$

with $\mathbf{T}_\mathbf{a} = (N, \alpha, ..., \alpha)'$ *and the components of* $\mathbf{a}_k = (1, a_{1k}, ..., a_{Jk})'$ *are given by*

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

with $j = 1, ..., J$.

Proof. To prove Proposition 1, first note that, since the first constraint $\sum_s w_k = N$ must be satisfied, it follows that $\hat{F}_{x_j,\text{cal}}(t) = N^{-1}\sum_s w_k H_{x_j,s}(t, x_{jk})$. Proceeding as in Deville and Särndal (1992), we can show that the vector $\mathbf{a}_k = (1, a_{1k}, ..., a_{Jk})'$ satisfies

$$\mathbf{a}_k = \left(1, \frac{\partial \hat{F}_{x_1,\text{cal}}}{\partial w_k}, ..., \frac{\partial \hat{F}_{x_J,\text{cal}}}{\partial w_k}\right)'\Bigg|_{\sum_s w_k = N;\ \hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha,\ j=1,...,J}, \qquad (12)$$

that we now evaluate explicitly. Evaluating the derivatives, we have that $a_{jk} = N^{-1}H_{x_j,s}(t, x_{jk})$, $j = 1, ..., J$, evaluated at $t = Q_{x_j,\alpha}$. This leads to

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

$j = 1, ..., J$, as announced.

In (11), $\mathbf{T}_\mathbf{a}$ can be interpreted as the expected value of $\sum_s d_k\ \mathbf{a}_k$. The derived weights (10) in the distribution function estimator (6) rely on the variables $\mathbf{a}_k$, $k \in s$ defined by (12). Note that they correspond to a certain transformation of the auxiliary variable $\mathbf{x}_k$. The difference between the weights for totals and quantiles relies on this variable $\mathbf{a}_k$; when $\mathbf{a}_k$ is replaced by $\mathbf{x}_k$, we retrieve the original weights for totals. Consequently, it is useful to interpret this new variable. When estimating a total, the impact on the j^{th} calibration constraint is measured by x_{jk}, for each unit $k \in s$. In our framework, the impact of the unit k is now given by N^{-1} if $x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha})$; it corresponds to the factor $N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha})$ when $x_{jk} = U_{x_j,s}(Q_{x_j,\alpha})$ and it is null elsewhere. In section 5, we shall discuss other estimation problems, leading to different variables $\mathbf{a}_k$.

Noting the similarities between the estimation of totals and quantiles, variance estimation can also be considered. This issue is addressed in the next subsection.

3.3 Variance Estimation and Confidence Intervals

As described in the previous section, the estimator $\hat{Q}_{y,\text{cal},\alpha}$ displays several similarities to the usual GREG estimator for population totals. The transformed variables given by (12) provide the main difference between the calibration estimators for quantiles and totals. Interestingly, because of the structural similarity with the original calibration estimators, it is straightforward to derive a confidence interval for the proposed estimator $\hat{Q}_{y,\text{cal},\alpha}$. We consider the construction of confidence intervals following Woodruff's (1952) approach. The confidence interval is given in Result 1.

Result 1 (Woodruff confidence interval for the calibration estimator for quantiles). *The confidence interval based on Woodruff's* (1952) *approach, using the calibration estimator* (9) *for the quantile* $Q_{y,\alpha}$ *is given by*

$$[\hat{F}^{-1}_{y,\text{cal}}(\hat{c}_{1y}),\ \hat{F}^{-1}_{y,\text{cal}}(\hat{c}_{2y})], \tag{13}$$

where $\hat{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\}]^{1/2}$ *and* $\hat{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\}]^{1/2}$. *The resulting procedure yields an approximate confidence interval for* $Q_{y,\alpha}$ *at a specified* $1-\gamma$ *confidence level.*

Proof. Assuming that $\hat{F}_{y,\text{cal},\alpha}(Q_{y,\alpha})$ is approximately normally distributed, it follows that $\Pr(c_{1y} \le \hat{F}_{y,\text{cal},\alpha}(Q_{y,\alpha}) \le c_{2y})$ should approximately be equal to $1-\gamma$, if one chooses

$$c_{1y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\}]^{1/2}, \tag{14}$$

$$c_{2y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\}]^{1/2}, \tag{15}$$

where z_γ denotes the γ^{th} – quantile of the $N(0,1)$ standard normal distribution. Since $\hat{F}_{y,\text{cal},\alpha}(Q_{y,\alpha})$ represents essentially a sample mean, a possible variance estimator justified by the classical Taylor linearization is given by

$$\hat{V}\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\} = N^{-2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}(w_k\, e_k)\,(w_l\, e_l), \tag{16}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$; the weights $w_k, k \in s$, correspond to the calibrated weights (3) which reduce to (10) when D is the quadratic distance function (2); the residuals are given by $e_k = H_{y,s}(\hat{Q}_{y,\text{cal},\alpha},\ y_k) - \mathbf{a}_k'\,\hat{\mathbf{B}}_s$ where

$$\hat{\mathbf{B}}_s = \left(\sum_s w_k\ q_k\ \mathbf{a}_k\ \mathbf{a}_k'\right)^{-1}\sum_s w_k\ q_k\ \mathbf{a}_k\ H_{y,s}(Q_{y,\text{cal},\alpha},\ y_k)$$

represents the regression coefficient estimator. Since the constants c_{1y} and c_{2y} given by (14) and (15) rely on $V\{\hat{F}_{y,\text{cal}}(Q_{y,\alpha})\}$, we can estimate these quantities using the variance estimator (16).

In Result 1, note that Deville and Särndal (1992) advocated a w – weighted variance estimator similar to (16) for estimating the variance of the calibration estimators of the population totals. The performance of the proposed calibration estimator (9) and the confidence interval given by (13) are studied empirically in section 4.

4. Simulation Results

From a practical point of view, it is natural to inquire about the finite sample properties of the new calibration estimators and to compare them to popular estimators for quantiles available in the literature. In this section, simulation experiments are undertaken, to illustrate empirically the new estimators. In particular, their empirical bias and variance in real populations are investigated. The coverage properties of the confidence intervals represent another question of practical interest, which is also studied.

In partial answer to these questions, we carried out three small simulation studies. For several sampling plans and for real populations, the proposed calibration estimator for quantiles is compared to its popular competitors. In the next subsection 4.1, we describe in detail the populations investigated and we discuss the sampling plans chosen. In subsection 4.2, the estimators included in the empirical study are presented and, in subsection 4.3, the frequentist measures (empirical bias, variance and mean squared error, coverage rates of the confidence intervals) are described. Our empirical results are analyzed in subsection 4.4.

4.1 Description of the Real Populations and the Sampling Plans

The real populations are displayed in Figures 1 to 6. The first population, noted MU284, is taken from Särndal *et al.* (1992, Appendix B). This population consists of $N = 284$ municipalities in Sweden. We retain as variable of interest the population in 1985 (variable P85), and we assume that the auxiliary information available is the population in 1975 (variable P75). Both variables are measured in thousands. In Figure 1, the variable P85 is expressed as a function of P75; as expected, the relationship between P85 and P75 is strongly linear. The variable P85 follows a highly skewed distribution, as shown in Figure 2. In this population, 500 samples were drawn according to simple random sampling without replacement (SRS). In addition, the same study was carried out under a sampling plan with unequal probabilities, the Poisson (PO) sampling scheme. The properties of the PO sampling plan are described in Särndal *et al.* (1992). Due to the wide range of values for y, it was not possible to construct sample selection probabilities π_k of the form $\pi_k \propto y_k$, since this would mean that some π_k had to be greater than one. For the purpose of our illustration, we determined selection probabilities using the relation $\pi_k \propto 0.2y_k + 0.05$ (we recognize that these π_k's are idealized, since y_k is not available in practice). Under the SRS sampling plan (PO sampling plan), we considered the sample sizes (expected sample sizes) $n = 25$ and $n = 50$.

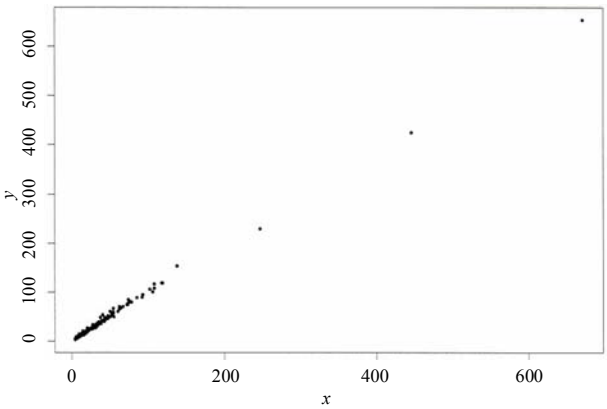


Figure 1. The Population MU284, where y = P85 and x = P75.

For the second study, we chose the MU284 population, but now made the variable of interest y = RMT85, which

represents the revenues from 1985 municipal taxation (in millions of kronor). Here the auxiliary variable chosen is x = REV84, which denotes real estate values according to 1984 assessments for each municipality (in millions of kronor). As can be seen in Figure 3, the relationship between x and y is somewhat spread out for larger values of x. The histogram of the variable RMT85 reveals that it follows a skewed distribution (Figure 4). For this study, 500 samples were drawn according to the SRS scheme of size $n = 25$ and $n = 50$.

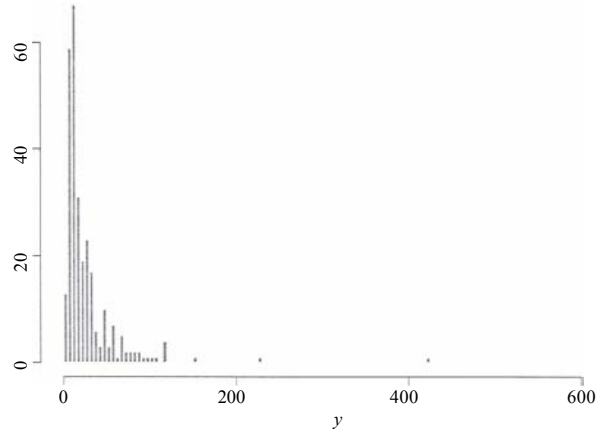


Figure 2. Histogram of the Variable P85 in the MU284 Population.

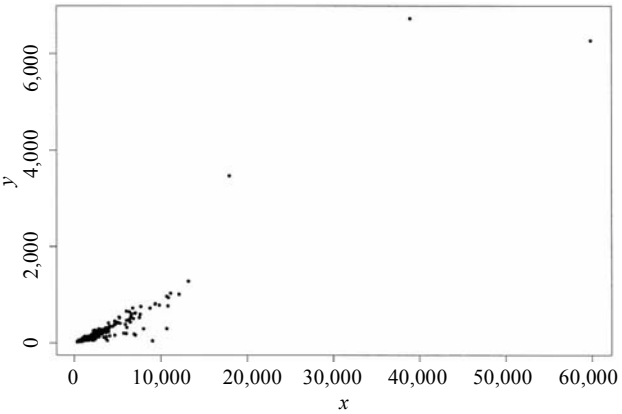


Figure 3. The Population MU284, where y = RMT85 and x = REV84.

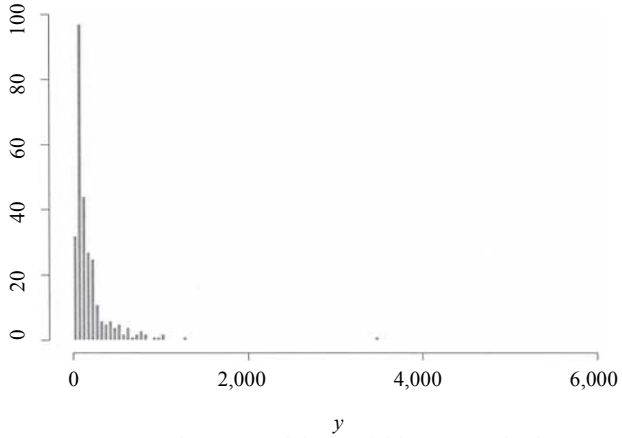


Figure 4. Histogram of the Variable RMT85 in the MU284 Population.

The third population is based on a random subsample of the *Survey of Labor and Income Dynamics*, noted SLID982. The survey was conducted at Statistics Canada in 1998. For simplicity's sake, only entries with no missing values were selected. The size of the subsample is $N = 2,000$ and for our purpose this is assumed to be a population (the original sample size of this survey is approximately 60,000). Taxable income (in thousands of dollars) is the target variable and the auxiliary variable is the duration in months of the current employment. From Figure 5, the linear relationship between taxable income and length of employment is less pronounced. However, the two variables do not appear to be independent. In Figure 6, the variable of interest exhibits a strong coefficient of skewness. We have drawn 500 samples from the SLID982 population, according to SRS and PO sampling plans. The sample sizes (expected sample size) $n = 100$ and $n = 200$ were considered. For PO sampling, the first order probabilities, $\pi_k, k \in U$, were defined according to two rules. Under the first rule, the π_k's were created such that π_k is approximately proportional to the variable of interest, that is taxable income (for the purpose of our study we assume that it is possible to create such π_k's). Since some y_k are negative in this population, we chose $p_{1k} = y_k - \min\{y_k,\ k \in U\} + 1$ and we defined $\pi_k = E(n_s)p_{1k} / \sum_U p_{1k}$, where $E(n_s)$ stands for the expected sample size, in our case $E(n_s) = 100$ and 200. Under the second rule, the π_k's were proportional to the entries in Table 1. This means that for each $k \in U$, there exists a factor p_{2k}, which is determined by the age-sex group of individual k. Then $\pi_k = E(n_s)p_{2k} / \sum_U p_{2k}$, where the factors p_{2k} are given in Table 1. The factors p_{2k} in Table 1 are based on a hypothetical sampling plan, in which we assume that these factors provide suitable size measures for the units in the various age-sex classes (see *e.g.*, Särndal *et al.* (1992, page 87)); for these units, more males than females are likely to be selected and, for both sexes, adults in the 27 to 37 and 38 to 46 age range are more likely to be included in the sample.

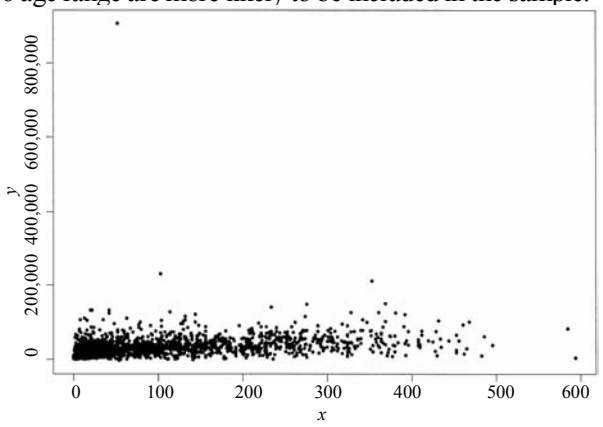


Figure 5. The SLID982 Population, where the Dependent Variable is the Taxable Income and Independent Variable is the Duration of Current Employment (in Months).

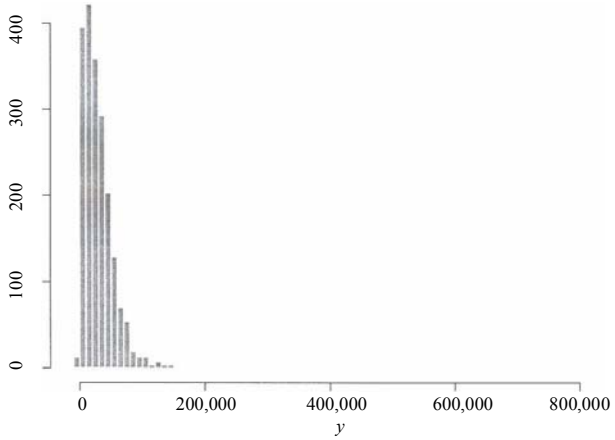


Figure 6. Histogram of the Taxable Income in the SLID982 Population.

Table 1
Factor p_{2k} by Age and Sex of Individual k, in the SLID982 Population

		Age			
		16–25	27–37	38–46	47–69
Sex	Male	3	6	5	4
	Female	1	2	3	2

In these three studies, we estimate the quartiles, that is the population parameters $Q_{y,\alpha}$ with $\alpha = 0.25$, 0.5 and 0.75. Since the variables of interest display highly skewed distributions, it might be particulary interesting to study the quantile corresponding to $\alpha = 0.75$, in addition to the median and the first quartile. The next section describes the estimators included in the study.

4.2 Estimators Included in the Empirical Study

Since one of our goals is to propose estimators with reasonable properties with respect to bias, variance and coverage rates of the confidence intervals, we compare the new estimator defined by (9) based on the metric (2) to some of the popular quantile estimators proposed in the literature.

First, we include the simple design-based estimator based on the inversion of the estimator $\hat{F}_y(t) = \sum_s d_k\, H_{y,s}(t, y_k) / \sum_s d_k$:

$$\hat{Q}_{y,\mathrm{HT},\alpha} = \hat{F}_y^{-1}(\alpha). \tag{17}$$

The estimator (17) does not make use of auxiliary information. A possible variance estimator is

$$\hat{V}\{\hat{F}_y(Q_{y,\alpha})\} = \hat{N}^{-2} \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_k) - \alpha)}{\pi_k} \right\} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_l) - \alpha}{\pi_l} \right\},$$

where $\hat{N} = \sum_s d_k$, and confidence intervals can be calculated using

$$[\hat{F}_y^{-1}(\tilde{c}_{1y}),\ \hat{F}_y^{-1}(\tilde{c}_{2y})],$$

where

$$\tilde{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}, \tag{18}$$

$$\tilde{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}. \tag{19}$$

For more details, see Särndal *et al.* (1992, page 202).

We also include in our empirical study the model-based estimator of Chambers and Dunstan (1986), which is motivated by a linear superpopulation model $y_k = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_k + \epsilon_k$, $k \in U$, where ϵ_k forms an identically and independently distributed sequence of random variables with mean zero and finite variance. Their estimator is defined as

$$\hat{Q}_{y,\mathrm{CD},\alpha} = \inf\{t \mid \hat{F}_{y,\mathrm{CD}}(t) \geq \alpha\}, \tag{20}$$

where $\hat{F}_{y,\mathrm{CD}}(t) = N^{-1}\{\sum_s H(t - y_k) + \sum_{U/s} \hat{G}(t - \hat{y}_k)\}$ represents a model-based estimator of the distribution function,

$$\hat{G}(u) = n^{-1} \sum_s H(u - \hat{\epsilon}_k) \tag{21}$$

denotes the empirical distribution function of the residuals $\hat{\epsilon}_k = y_k - \hat{y}_k$, $k \in s$, and $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_k$, $k \in U/s$ correspond to the least-squares predictions. Since the estimator (20) basically imputes the unknown y_k for $k \in U/s$, note that it necessitates a complete knowledge of $\mathbf{x}_k$ for $k \in U$.

The construction of a confidence interval for $\hat{Q}_{y,\mathrm{CD},\alpha}$ relies on estimating the variance $V\{\hat{F}_{y,\mathrm{CD}}(t)\}$. However, this variance estimation problem creates difficulties, since any analytical variance formula depends on the assumed model. Furthermore, such analytical expressions involve kernel density estimators, which are numerically intensive and depend on a kernel function and a bandwidth. For all these reasons, we decide to implement the delete-one jackknife variance estimators studied in Wu and Sitter (2001), who have shown the consistency of the proposed variance estimators. In the context of survey sampling, various resampling methods, including the jackknife, are introduced in Kovar, Rao and Wu (1988). The jackknife technique involves deleting a unit and re-calculating the estimator. Let $s_i = s/\{i\}$ be the sample without unit i. Consider $\hat{\beta}_{0i}$ and $\hat{\boldsymbol{\beta}}_i$, the regression estimators of β_0 and $\boldsymbol{\beta}$ calculated on s_i. Under a simple regression model, define

$$F_i^* = (n-1)^{-1} \sum_{k \in s_i} \left[N^{-1} \sum_{l \in U/s} H\{\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{\beta}_i(x_l - x_k) - y_k\} \right].$$

A consistent variance estimator of $V\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\mathrm{CD},\alpha})\}$ is given by

$$\hat{V}_{y,\mathrm{CD}}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\} = \frac{n-1}{n}\sum_{i\in s}\ (F_i^* - \bar{F}^*)^2 + \frac{f(1-f)}{N-n}\sum_{k\in U/s} \hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\,\{1-\hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\},$$

where $f = n/N$ is the sampling fraction, $\bar{F}^* = n^{-1}\sum_s F_i^*$, and $\hat{G}$ is given by (21). Based on $\hat{V}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\}$, it is now possible to calculate the confidence intervals for $Q_{y,\alpha}$ using the inversion approach.

Since our method necessitates only the knowledge of the vector of quantiles $\mathbf{Q}_{\mathbf{x},\alpha}$, we include in our study the ratio and difference estimators for the quantiles studied in Rao *et al.* (1990):

$$\hat{Q}_{y,\mathrm{ra},\alpha} = Q_{x,\alpha}(\hat{Q}_{y,\mathrm{HT},\alpha}/\hat{Q}_{x,\mathrm{HT},\alpha}), \tag{22}$$

$$\hat{Q}_{y,\mathrm{diff},\alpha} = \hat{Q}_{y,\mathrm{HT},\alpha} + \hat{R}(Q_{x,\alpha} - \hat{Q}_{x,\mathrm{HT},\alpha}), \tag{23}$$

where $\hat{Q}_{y,\mathrm{HT},\alpha}$ is given by (17) and $\hat{Q}_{x,\mathrm{HT},\alpha}$ is calculated similarly; the ratio estimator given by $\hat{R} = \sum_s d_k\, y_k / \sum_s d_k\, x_k$ provides a consistent estimator of $R = \sum_U y_k / \sum_U x_k$. Note that the estimators (22) and (23) are elaborated based on a scalar auxiliary variable, that is $J = 1$. Valid variance estimators of (22) and (23) are given by:

$$\hat{V}(\hat{Q}_{y,\mathrm{ra},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \left(\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\right)^2 \hat{V}(\hat{Q}_{x,\mathrm{HT},\alpha}) - 2\,\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\,\hat{Q}_{x,\mathrm{HT},\alpha}),$$

$$\hat{V}(\hat{Q}_{y,\mathrm{diff},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \hat{R}^2\,\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) - 2\hat{R}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\,\hat{Q}_{x,\mathrm{HT},\alpha}).$$

These variance estimators rely on the variance of $\hat{Q}_{y,\mathrm{HT},\alpha}$, and the covariance between $\hat{Q}_{y,\mathrm{HT},\alpha}$ and $\hat{Q}_{x,\mathrm{HT},\alpha}$ which are estimated using Woodruff's (1952) approach:

$$\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) = \frac{W_y^2}{4z_{1-\gamma/2}^2},$$

$$\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha},\,\hat{Q}_{x,\mathrm{HT},\alpha}) = \frac{W_y W_x \hat{C}\{\hat{F}_x(Q_{x,\alpha}),\,\hat{F}_y(Q_{y,\alpha})\}}{4z_{1-\gamma/2}^2[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}},$$

where $W_y = \hat{F}_y^{-1}(\tilde{c}_{2y}) - \hat{F}_y^{-1}(\tilde{c}_{1y})$ and $W_x = \hat{F}_x^{-1}(\tilde{c}_{2x}) - \hat{F}_x^{-1}(\tilde{c}_{1x})$ denote the Woodruff intervals associated with y and x, with $\tilde{c}_{1y}$ and $\tilde{c}_{2y}$ defined by (18) and (19), $\tilde{c}_{1x} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$, $\tilde{c}_{2x} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$ and

$$\hat{C}\{\hat{F}_y(Q_{y,\alpha}),\ \hat{F}_x(Q_{x,\alpha})\} = \hat{N}^{-2}\sum\sum_s\ \frac{\Delta_{kl}}{\pi_{kl}}\left\{\frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha},\ y_k) - \alpha}{\pi_k}\right\}\left\{\frac{H_{x,s}(\hat{Q}_{x,\mathrm{HT},\alpha},\ x_l) - \alpha}{\pi_l}\right\}.$$

Summarizing, we expect $\hat{Q}_{y,\mathrm{CD},\alpha}$ to perform well when the linear model describes the population adequately. This motivates the comparison of the new methodology with a model-based estimator. Furthermore, it seems of interest to evaluate $\hat{Q}_{y,\mathrm{cal},\alpha}$ and the leading design-based proposals, such as $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{ra},\alpha}$. The estimators $\hat{Q}_{y,\mathrm{cal},\alpha}$, $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{ra},\alpha}$ use $Q_{x,\alpha}$ only to improve the estimations and they take into account the sampling plan; these estimators are natural competitors. Note that the different estimators included in our study are elaborated under different assumptions on the dimension of the vector of the auxiliary variable $\mathbf{x}$, and on the availability of $\mathbf{x}_k$. Table 2 provides a comparison of the different estimators described in this section.

Table 2
Comparison of the Proposed Calibration Estimators and of Some Leading Estimators for Quantiles Proposed in the Literature, with Respect to the Dimension J of $\mathbf{x}$ and the information requirement on $\mathbf{x}$

Estimator	Dimension of $\mathbf{x}$	Information requirements on $\mathbf{x}$
$\hat{Q}_{y,\mathrm{HT},\alpha}$	n.a.	none
$\hat{Q}_{y,\mathrm{CD},\alpha}$	$J \geq 1$	$\mathbf{x}_k, k \in U/s$
$\hat{Q}_{y,\mathrm{ra},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{diff},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{cal},\alpha}$	$J \geq 1$	$Q_{\mathbf{x},\alpha}$

4.3 Frequentist Measures

Our goal it to evaluate the estimators with respect to bias and variance. Other important considerations are the mean squared error (MSE) and the coverage rates of the confidence intervals.

Let $\hat{Q}_{y,\alpha}$ be an estimator of the population quantile $Q_{y,\alpha}$. Assume $\hat{Q}_{y,\alpha}^{(v)}$ is the estimator of the quantile calculated using the sample v, $v = 1, \ldots, K$. The Monte Carlo mean E_{MC}, the Monte Carlo bias B_{MC}, and the Monte Carlo variance V_{MC} are given by the usual formulas, that is

$$E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\hat{Q}_{y,\alpha}^{(v)},$$

$$B_{\mathrm{MC}} = E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) - Q_{y,\alpha},$$

$$V_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\{\hat{Q}_{y,\alpha}^{(v)} - E_{\mathrm{MC}}(\hat{Q}_{y,\alpha})\}^2.$$

Our main criterion for determining efficiency is the Monte Carlo MSE, defined by $\mathrm{MSE}_{\mathrm{MC}} = K^{-1}\sum_{v=1}^{K}(\hat{Q}_{y,\alpha}^{(v)} - Q_{y,\alpha})^2$. The confidence intervals are calculated at the 95% confidence level, according to the procedures described in the previous sections. For an estimator $\hat{Q}_{y,\alpha}^{(v)}$ and its variance estimator $\hat{V}^{(v)}$, $v = 1, \ldots, K$, the coverage rates at the 95% confidence level are calculated as

$$\mathrm{CR}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K} I\left(\left\{Q_{y,\alpha} \in \left[\hat{Q}_{y,\alpha}^{(v)} - 1.96\sqrt{\hat{V}^{(v)}},\ \hat{Q}_{y,\alpha}^{(v)} + 1.96\sqrt{\hat{V}^{(v)}}\right]\right\}\right),$$

where $I(A)$ is the indicator function of the set A. The coverage rates are given below the column CR. We recall that we adopt $K = 500$ for all studies.

4.4 Discussion of the Empirical Results

The results are presented in Tables 3 to 8. We first discuss the results from Tables 3 to 4, when sampling the MU284 population with SRS and PO sampling plans. As can be seen, all the estimators display a similar behavior in both studies. The model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ appears to be the most efficient among those analyzed when examining $\alpha = 0.75$ and is in general very efficient. This was expected, since the relationship between $x = \mathrm{P75}$ and $y = \mathrm{P85}$ is strongly linear and the model-based estimator assumes a simple regression model. However, for $\alpha = 0.25$ the differences in efficiency are less pronounced with respect to the other estimators based on auxiliary information. Among the estimators using only $Q_{x,\alpha}$ as information on the auxiliary variable, a rather similar performance is obtained. When the sample size is small, coverage rates usually deviate from the 95% nominal level. This is particulary true for the coverage rates of $\hat{Q}_{y,\mathrm{cal},\alpha}$, which are somewhat underestimated. However, some improvement is observed at $n = 50$, illustrating the consistency of the procedures studied. On the other hand, those of $\hat{Q}_{y,\mathrm{ra},\alpha}$ and $\hat{Q}_{y,\mathrm{diff},\alpha}$ are always one. This suggests that the variances are overestimated for these estimators. Due to an important component of bias in the MSE, the coverage rates of the model-based estimator sometimes deteriorate as the sample size increases. The best coverage rates are obtained by using the simple HT estimator, $\hat{Q}_{y,\mathrm{HT},\alpha}$, which is however less efficient than the other estimators.

Table 5 shows the result for the second population, which is the MU284 population but with $y = \mathrm{RMT85}$ and $x = \mathrm{REV84}$. Figure 3 seems to show a heteroscedasticity phenomenon in this population. In view of this, since the ratio estimator is justified when the underlying population displays such behavior, it is not surprising that the ratio estimator $\hat{Q}_{y,\mathrm{ra},\alpha}$ performs well in this particular situation; if outperforms $\hat{Q}_{y,\mathrm{diff},\alpha}$ in several cases. For a small sample size, the ratio estimator generally behaves better than $\hat{Q}_{y,\mathrm{cal},\alpha}$. However, for $n = 50$, the calibration estimator appears to perform as well or slightly better than the ratio estimator. In this experiment, the bias and variance of the model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ increase the MSE substantially. Furthermore, in some cases, confidence intervals for this estimator could not be obtained, since the Woodruff method is not appropriate in cases with extremely large variance (the Woodruff interval becomes too large and the linearity of the distribution function within this interval can thus no longer be assumed). We suspect that a model taking into account heteroscedasticity might improve the performance of the model-based estimator. This highlights the fact that to obtain high efficiency with model-based estimators, the model must be correctly specified.

The results in Table 6 to 8 concern the SLID982 population, under SRS and PO sampling plans with two rules for the π_k's. All the estimators in Table 6 perform reasonably well in estimating the first quartile and the median, except for the ratio estimator $\hat{Q}_{y,\mathrm{ra},\alpha}$ which is the least efficient. Since the relationship between the dependent and independent variables is not precisely a linear model, this may partially explain the poor performance of the ratio estimator in this case. The relationship between x and y is not proportional and so the difference estimator $\hat{Q}_{y,\mathrm{diff},\alpha}$ appears preferable to $\hat{Q}_{y,\mathrm{ra},\alpha}$. However, for $\alpha = 0.75$, these estimators show the highest MSE, being both the least efficient. Interestingly, in this part of the experiment $\hat{Q}_{y,\mathrm{cal},\alpha}$ dominates the design-based estimators in terms of MSE. However, for small α, $\hat{Q}_{y,\mathrm{diff},\alpha}$ and $\hat{Q}_{y,\mathrm{cal},\alpha}$ perform similarly. It should be noted that for a larger sample size, $\hat{Q}_{y,\mathrm{cal},\alpha}$ and $\hat{Q}_{y,\mathrm{CD},\alpha}$ give the best efficiencies for the median and the third quartile. In fact, the model-based estimator $\hat{Q}_{y,\mathrm{CD},\alpha}$ slightly outperforms $\hat{Q}_{y,\mathrm{cal},\alpha}$, but it should be noted that it uses more auxiliary information than $\hat{Q}_{y,\mathrm{cal},\alpha}$.

Tables 7 and 8 present results under PO sampling plans. In general, design-based estimators perform much like those under SRS sampling plan. This is not the case for the model-based estimator; it is less efficient, likely because it does not incorporate the information about the sampling plan. More precisely, Table 7 presents simulation results under PO sampling, using the first rule for the π_k's, $k \in U$. Coverage rates of the model-based estimator are particularly disappointing in this experiment; the components of bias were too important in the MSE. The design-based estimators provide much closer empirical coverage rates, to the nominal 95% confidence level. For moderate and large α, $\hat{Q}_{y,\mathrm{cal},\alpha}$ is the most efficient estimator. In fact, the calibration estimator $\hat{Q}_{y,\mathrm{cal},\alpha}$ performs well in this

experiment. Finally, Table 8 presents results obtained under PO sampling with the second rule for the π_k's. In this case, $\hat{Q}_{y,\mathrm{ra},\alpha}$. is the least efficient estimator for the first quartile and the median, and $\hat{Q}_{y,\mathrm{diff},\alpha}$ is the least efficient for $\alpha = 0.75$. In general, $\hat{Q}_{y,\mathrm{cal},\alpha}$ dominates the other estimators in this situation, offering the highest efficiency.

Table 3
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = P85, x = P75, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	$\mathrm{MSE}_{\mathrm{MC}}$	CR	B_{MC}	V_{MC}	$\mathrm{MSE}_{\mathrm{MC}}$	CR
0.25	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.0343	0.5075	0.5077	0.886	−0.0499	0.2437	0.2457	0.828
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	−0.0266	2.3196	2.3157	0.952	0.0035	1.1087	1.1065	0.936
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.1444	0.3869	0.4070	1.000	−0.0774	0.1684	0.1741	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.1486	0.3901	0.4114	1.000	−0.0734	0.1723	0.1774	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	0.4855	0.2791	0.5143	0.906	0.5485	0.1981	0.4985	0.824
0.5	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.2762	1.6499	1.7229	0.918	−0.2835	0.9585	1.0370	0.944
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	0.2605	12.5161	12.5589	0.922	−0.0064	5.8466	5.8349	0.916
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.2586	0.8828	0.9479	1.000	−0.4296	0.6701	0.8533	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.2775	0.9898	1.0648	1.000	−0.4331	0.7492	0.9352	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	0.9431	0.4054	1.2940	0.866	0.9884	0.2410	1.2175	0.714
0.75	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.6229	3.3241	3.7055	0.614	−0.3661	1.8107	1.9411	0.710
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	−0.1414	53.1951	53.1088	0.948	−0.3692	18.8586	18.9572	0.964
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.7925	3.0021	3.6242	1.000	−1.0004	1.4594	2.4573	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.8230	3.4379	4.1083	1.000	−1.0396	1.5267	2.6044	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	0.4343	0.5108	0.6984	0.954	0.4485	0.2618	0.4624	0.974

Table 4
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = P85, x = P75, Under PO Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	$\mathrm{MSE}_{\mathrm{MC}}$	CR	B_{MC}	V_{MC}	$\mathrm{MSE}_{\mathrm{MC}}$	CR
0.25	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.0441	0.4886	0.4896	0.888	−0.0169	0.2601	0.2599	0.828
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	−0.1698	2.2825	2.3068	0.936	−0.0384	1.1828	1.1819	0.928
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.1509	0.3857	0.4076	1.000	−0.0913	0.2100	0.2179	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.1634	0.3821	0.4080	1.000	−0.0877	0.2149	0.2221	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	0.6709	0.3310	0.7805	0.896	0.8792	0.1339	0.9066	0.554
0.5	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.3610	1.4881	1.6155	0.920	−0.3236	0.8833	0.9863	0.936
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	−0.0612	11.3969	11.3778	0.926	−0.2712	5.2672	5.3302	0.906
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.3735	1.0009	1.1385	1.000	−0.4130	0.5486	0.7181	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.3962	1.1271	1.2818	1.000	−0.4217	0.5962	0.7729	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	1.1740	0.4947	1.8719	0.820	1.3297	0.2146	1.9822	0.474
0.75	$\hat{Q}_{y,\mathrm{cal},\alpha}$	−0.6420	2.6605	3.0674	0.608	−0.4476	1.6212	1.8183	0.708
	$\hat{Q}_{y,\mathrm{HT},\alpha}$	−0.6200	51.2934	51.5752	0.956	−0.6632	17.3625	17.7677	0.966
	$\hat{Q}_{y,\mathrm{ra},\alpha}$	−0.8686	2.8841	3.6329	1.000	−0.9683	1.6494	2.5837	1.000
	$\hat{Q}_{y,\mathrm{diff},\alpha}$	−0.9025	2.9826	3.7911	1.000	−1.0177	1.6340	2.6665	1.000
	$\hat{Q}_{y,\mathrm{CD},\alpha}$	0.4620	0.4501	0.6627	0.982	0.5388	0.2329	0.5228	0.980

Table 5
Monte Carlo Simulation Results for Sampling from the MU284 Population, y = RMT85, x = REV84, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 25$				$n = 50$			
α	Estimator	B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	1.0161	51.5421	52.4714	0.892	0.6499	24.0662	24.4404	0.954
	$\hat{Q}_{y,\text{HT},\alpha}$	0.3733	110.2572	110.1760	0.960	0.3383	47.2921	47.3120	0.962
	$\hat{Q}_{y,\text{ra},\alpha}$	3.0025	65.4135	74.2979	0.998	2.3856	30.7284	36.3580	0.992
	$\hat{Q}_{y,\text{diff},\alpha}$	2.5952	107.7891	114.3084	0.994	2.4083	55.6977	61.3862	0.986
	$\hat{Q}_{y,\text{CD},\alpha}$	−16.5165	1661.0257	1930.4983	0.990	−17.3217	820.7447	1119.1443	0.960
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−1.6219	215.0326	217.2330	0.870	−0.3419	118.2125	118.0930	0.922
	$\hat{Q}_{y,\text{HT},\alpha}$	0.0075	763.6236	762.0964	0.910	−0.3977	331.2357	330.7314	0.914
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7712	212.8298	212.9988	0.996	−0.2810	136.4382	136.2443	0.996
	$\hat{Q}_{y,\text{diff},\alpha}$	0.3415	283.6718	283.2210	0.998	−1.0104	201.3707	201.9889	0.998
	$\hat{Q}_{y,\text{CD},\alpha}$	17.6124	190.0045	499.8199	n.a.	13.5037	100.2106	282.3611	0.566
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−5.3477	1023.6924	1050.2431	0.826	−4.7339	443.0660	464.5896	0.926
	$\hat{Q}_{y,\text{HT},\alpha}$	−4.6352	3526.8202	3541.2514	0.938	−5.8890	1242.4858	1274.6812	0.940
	$\hat{Q}_{y,\text{ra},\alpha}$	−1.4390	980.5573	980.6669	0.994	−2.0070	555.5135	558.4305	1.000
	$\hat{Q}_{y,\text{diff},\alpha}$	−5.3988	1464.7867	1491.0041	0.996	−3.9008	744.1604	757.8881	1.000
	$\hat{Q}_{y,\text{CD},\alpha}$	49.3038	2753.8212	5179.1826	n.a.	49.4089	1488.9734	3927.2324	0.596

Table 6
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under SRS Sampling Plan. The Number of Replications is Set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.1360	3.0390	3.0514	0.956	0.2331	1.6787	1.7297	0.934
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0596	3.6099	3.6062	0.946	0.0499	1.9277	1.9263	0.918
	$\hat{Q}_{y,\text{ra},\alpha}$	0.3067	6.8815	6.9618	0.970	0.0910	3.0743	3.0764	0.958
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.0504	2.9691	2.9657	0.980	0.0198	1.6139	1.6111	0.952
	$\hat{Q}_{y,\text{CD},\alpha}$	1.1042	2.1180	3.3329	0.922	1.1392	1.2937	2.5888	0.826
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.4034	6.3364	6.4865	0.966	−0.1402	2.9940	3.0076	0.940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4157	7.4589	7.6168	0.918	−0.1894	3.5865	3.6151	0.928
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7015	41.8314	42.2399	0.958	0.2238	18.7005	18.7131	0.952
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.4859	14.2083	14.4160	0.970	−0.2740	6.6184	6.6803	0.974
	$\hat{Q}_{y,\text{CD},\alpha}$	0.5702	3.5420	3.8601	0.952	0.6697	1.7559	2.2009	0.932
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.4164	12.4657	12.6142	0.952	−0.2384	5.9118	5.9568	0.950
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.5913	12.5456	12.8701	0.930	−0.3519	6.5496	6.6603	0.926
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7404	48.6836	49.1345	0.954	0.2967	18.5786	18.6294	0.966
	$\hat{Q}_{y,\text{diff},\alpha}$	0.3288	53.6456	53.6464	0.954	0.1841	21.7552	21.7456	0.966
	$\hat{Q}_{y,\text{CD},\alpha}$	0.5966	8.3416	8.6809	0.954	0.5413	4.3692	4.6535	0.936

Table 7
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the First Rule for the Construction of the π_k, $k \in U$. The number of replications is set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.1393	4.8403	4.8500	0.956	0.1603	2.8293	2.8493	0.922
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.0477	5.8276	5.8182	0.934	−0.0227	3.5939	3.5872	0.924
	$\hat{Q}_{y,\text{ra},\alpha}$	0.1648	9.5171	9.5252	0.980	0.1263	4.8687	4.8749	0.972
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1418	4.7045	4.7152	0.960	−0.0464	2.9213	2.9176	0.936
	$\hat{Q}_{y,\text{CD},\alpha}$	3.9150	3.5279	18.8477	0.584	3.9114	1.9163	17.2112	0.194
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1746	8.2437	8.2577	0.944	−0.2413	3.6477	3.6986	0.940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.2824	10.1117	10.1712	0.916	−0.3343	4.5023	4.6050	0.936
	$\hat{Q}_{y,\text{ra},\alpha}$	0.6558	50.4938	50.8228	0.944	0.4263	26.5883	26.7169	0.948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.5975	17.0315	17.3544	0.972	−0.3496	8.9060	9.0104	0.970
	$\hat{Q}_{y,\text{CD},\alpha}$	4.3173	4.4061	23.0363	0.484	4.0937	2.0711	18.8252	0.184
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.2229	12.1861	12.2114	0.942	−0.2113	6.5823	6.6138	0.952
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4150	14.2935	14.4371	0.934	−0.2786	7.6597	7.7220	0.934
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7861	47.3844	47.9077	0.980	−0.1344	19.5992	19.5781	0.958
	$\hat{Q}_{y,\text{diff},\alpha}$	0.4347	52.3845	52.4687	0.972	−0.3409	23.8277	23.8962	0.958
	$\hat{Q}_{y,\text{CD},\alpha}$	4.4114	7.7023	27.1478	0.654	4.3549	4.1566	23.1136	0.392

Table 8
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the Second Rule for the Construction of the π_k, $k \in U$. The Number of Replications is Set at $K = 500$

		$n = 100$				$n = 200$			
α	Estimator	BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,\text{cal},\alpha}$	0.2392	3.4402	3.4906	0.962	0.1674	1.5214	1.5464	0.952
	$\hat{Q}_{y,\text{HT},\alpha}$	0.0267	4.0027	3.9954	0.940	−0.0370	1.6995	1.6975	0.958
	$\hat{Q}_{y,\text{ra},\alpha}$	0.4402	7.4350	7.6139	0.970	0.1850	3.0687	3.0968	0.978
	$\hat{Q}_{y,\text{diff},\alpha}$	0.0528	3.2842	3.2804	0.972	−0.0127	1.4718	1.4690	0.964
	$\hat{Q}_{y,\text{CD},\alpha}$	2.1458	3.0460	7.6444	0.876	1.9785	1.3010	5.2130	0.690
0.5	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1410	6.5627	6.5695	0.942	−0.2850	2.9662	3.0415	0.954
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.2133	7.6604	7.6906	0.928	−0.2876	3.6017	3.6772	0.926
	$\hat{Q}_{y,\text{ra},\alpha}$	1.0245	43.2773	44.2402	0.930	−0.3075	17.7242	17.7833	0.948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0.1973	14.5261	14.5360	0.958	−0.6111	6.2988	6.6596	0.978
	$\hat{Q}_{y,\text{CD},\alpha}$	2.2140	4.5617	9.4543	0.834	1.8882	2.0393	5.6005	0.738
0.75	$\hat{Q}_{y,\text{cal},\alpha}$	−0.1985	12.6334	12.6476	0.952	−0.0022	5.6442	5.6329	0.966
	$\hat{Q}_{y,\text{HT},\alpha}$	−0.4012	13.5045	13.6384	0.922	−0.1078	6.2239	6.2231	0.934
	$\hat{Q}_{y,\text{ra},\alpha}$	0.7968	44.0650	44.6118	0.958	0.3727	19.1830	19.2836	0.960
	$\hat{Q}_{y,\text{diff},\alpha}$	0.4613	49.6620	49.7755	0.960	0.2340	22.1292	22.1397	0.966
	$\hat{Q}_{y,\text{CD},\alpha}$	2.6329	9.6723	16.5850	0.854	2.6729	4.1179	11.2541	0.738

5. Concluding Remarks

In this paper, we have developed quantile estimators based on the calibration paradigm. The estimators are particularly easy to implement and to interpret, since they focus on weights and calibration constraints. Furthermore, they require only the population quantiles of the auxiliary variables, which can be vectorial. When the quadratic metric is adopted, analytic expressions can be obtained for calibrated weights as well as variance estimators, which are similar to those for the calibration estimator for totals. From a practical point of view, an appealing consequence of the new methodology is that the proposed estimators are easy to calculate; it suffices to transform the auxiliary variables and then use existing software to compute the calibration estimators.

In a small simulation study, we compared the calibration estimator for quantiles, under the quadratic metric, to other leading quantile estimators available in the literature. The proposed estimator performed reasonably well in our empirical experiments; its performance was often preferable or at least similar to that of other estimators using the same amount of information. The model-based estimator incorporating much more information about the auxiliary variables appeared preferable under SRS sampling and a correctly specified model, but was outperformed by the new estimator when the first order inclusion probabilities were unequal. In general, the proposed estimator compared very well with the design-based alternatives of Rao *et al.* (1990).

While, in this paper, we have concentrated on the estimation of quantiles by calibrating on known population quantiles for the auxiliary variables, calibration estimators can be extended to other important estimation problems of interest in survey sampling. The formulation of these problems all lead to different transformed variables, that we have noted $\mathbf{a}_k$ in this paper. For example, it is possible to formulate a calibration problem for the well-known Gini coefficient and then show that the solution to this calibration problem will give weights analogous to those derived in this paper; however these weights can only be determined numerically. More work is needed in this direction, in order to extend calibration estimators to a more general framework, which would include totals, quantiles, and Gini coefficients as special cases. Another challenging research avenue concerns the choice of the distribution function estimator. In this paper, we have advocated a distribution function estimator calculated using a linear interpolation. Alternatively, we could consider kernel distribution function estimator (see *e.g.*, Altman and Léger (1995)). Kernel density estimation from complex surveys is elaborated in Bellhouse and Stafford (1999). This means that, in $\hat{F}_{y,\,\mathrm{cal}}(t)$, the function $H_{y,s}(t, y_k)$ could be replaced by a general kernel, which would, however, depend on an additional parameter, the bandwidth. Note that the linear interpolation in the present paper avoids the choice of a bandwidth, which is often a delicate matter. Developing a general framework for calibration problems of a certain functional, and kernel distribution function estimators, are left for future studies.

Acknowledgements

The authors thank two anonymous referees for their thoughtful comments and suggestions, which greatly enhanced the paper. Discussions and comments from Raymond Chambers, Christian Léger, Éric Rancourt, Ulrich Rendtel and participants in the 32nd meeting of the Statistical Society of Canada and of the 2004 Joint Statistical Meeting are gratefully acknowledged. The first author was supported by a scholarship from the German Academic Exchange Service (DAAD) and the second author by grants from the National Science and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les technologies du Québec (Canada).

References

Altman, N., and Léger, C. (1995), Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195-214.

Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.

Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of finite population distribution functions. *Biometrika*, 79, 577-582.

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80, 107-116.

Deville, J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquêtes par sondage. In *Essais en l'Honneur d'Edmont Malinvaud*, (Eds, A. Monfort, and J.J. Laffond), *Economica*, Paris, 915-929.

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

Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.

Harms, T. (2003), Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators, Chintex working paper #13, Federal Statistical Office, Germany.

Kovačević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, 139-144.

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 (Supp.), 25-45.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75, 97-103.

Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society*, Series B (Methodological), 51, 261-269.

Meeden, G. (1995). Median estimation using auxiliary information. *Survey Methodology*, 21, 71-77.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, *INSEE Méthodes*, Tome 1, 100, 263-289.

Ren, R., and Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments, II^{ème} Colloque Francophone sur les Sondages. Bruxelles.

Rueda, M.M., Arcos A. and Martínez, M.D. (2003). Difference estimators of quantiles in finite populations. *Test*, 12, 481-496.

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

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

Thompson, M. (1997). *Theory of Sample Surveys*. Chapman & Hall, New York.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 625-646.

Wu, C., and Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, 29, 289-308.