Estimation of the distribution function with calibration methods

M. Rueda\textsuperscript{a}, S. Martínez\textsuperscript{b}, H. Martínez\textsuperscript{b}, A. Arcos\textsuperscript{a,*}

\textsuperscript{a}Department of Statistics and Operational Research, University of Granada, Spain
\textsuperscript{b}University of Almería, Spain

Received 18 June 2004; accepted 13 December 2005
Available online 23 February 2006

Abstract
The paper proposes a new calibration estimator for the distribution function of the study variable. This estimator is a distribution function unlike others estimators that use auxiliary information. Comparisons are made with existing estimators in two simulation studies.
© 2006 Elsevier B.V. All rights reserved.

MSC: 62D05
Keywords: Auxiliary information; Finite population; Calibration methods; Distribution function

1. Introduction
In sample surveys, supplementary population information is often used at the estimation stage to increase the precision of estimators of a population mean or total. The use of auxiliary information for estimating a finite population mean has been extensively studied, but relatively less effort has been devoted to the development of efficient methods for estimating the distribution function and the finite population quantiles.

In the presence of auxiliary information, there exist several procedures to obtain more efficient estimators for the population means and totals (in particular, customary ratio and regression estimators). Many of these techniques, when applied directly to the estimation of a distribution function, will produce unsatisfactory results. For instance, the ratio and regression type estimators for the distribution function suffer from several drawbacks, the obvious one being that they may not be a distribution function and can take values outside $[0,1]$.

Most of papers in this topic (Chambers and Dunstan, 1986; Rao et al., 1990; Dorfman and Hall, 1993) assumed a superpopulation model, and suggested model based estimators. All of these methods share the difficulty in the computation and the possible bias caused by the inadequacy of the assumed superpopulation model. These authors do not go deeply in the properties of the derived quantile estimators and do not study them in any simulations. Silva and Skinner (1995) carried out an exhaustive study of the properties of these estimators pointing out important problems.

The distribution function is a finite population mean of an indicator variable, but this estimation differs from the traditional estimation of a population mean in several fundamental aspects. The models specified for variables $y$ and $x$ in the estimation of a mean cannot be applied. Also, an estimator of $F(t)$ should have desirable properties (Silva

* Corresponding author.
E-mail address: arcos@ugr.es (A. Arcos).

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
and Skinner, 1995): the estimator should have the properties of a distribution function, uniqueness in the definition, availability of the variance-estimator, good conditional properties and should make efficient and flexible use of auxiliary information.

In this paper, we propose an estimator for the finite population distribution function using the calibration technique proposed by Deville and Särndal (1992). This estimator proposed in Section 2 is a genuine distribution function and possesses a number of attractive features which are studied in Sections 3, 4 and 5. The empirical comparison of the proposed method with the conventional estimators is performed in Section 6. Finally Section 7 contains some conclusions.

2. Calibration estimators of the distribution function

Consider a finite population \( U = \{1, \ldots, k, \ldots, N\} \), consisting of \( N \) different elements. Let \( s = \{1, \ldots, n\} \) be the set of \( n \) units included in a sample, selected according to a specified sampling design with inclusion probabilities \( \pi_k \) and \( \pi_k \) assumed to be strictly positive. Let \( y_k \) be the value of the study variable \( y \), for the \( k \)th population element, with which also is associated and auxiliary vector value \( x_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})' \). The values \( x_1, x_2, \ldots, x_N \) are known for the entire population but \( y_k \) is known only if the \( k \)th unit is selected in the sample \( s \).

The finite population distribution function of the study variable \( y \), is given by

\[
F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k),
\]

with

\[
\Delta(t - y_k) = \begin{cases} 
0 & \text{if } t < y_k, \\
1 & \text{if } t \geq y_k.
\end{cases}
\]

The distribution function \( F_y(t) \) can be estimated by the Horvitz–Thompson estimator, defined by

\[
\hat{F}_{yH}(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k),
\]

with \( d_k = 1/\pi_k \), the basic design weights.

The estimator \( \hat{F}_{yH}(t) \) is unbiased, but in general, is not a distribution function and does not use the auxiliary information provided by the vector \( x \). We shall modify the estimator \( \hat{F}_{yH}(t) \) to obtain new estimators of \( F_y(t) \), replacing the basic design weights \( d_k \) by new weights \( \omega_k \). This new set of weights is constructed with the calibration techniques (in Singh 2003, Chapter 5), there is a review of calibration estimation of population mean and variance).

We consider a new calibration estimator by first defining a pseudo-variable \( g_k = \vec{\beta} x_k \) for \( k = 1, 2, \ldots, N \), where

\[
\vec{\beta} = \left( \sum_{k \in s} d_k q_k x_k x_k' \right)^{-1} \cdot \sum_{k \in s} d_k q_k x_k y_k
\]

is a weighted estimator of the multiple regression coefficient \( \beta \) between \( y \) and \( x \). The \( q_k \) are known positive constants unrelated to \( d_k \).

We then define the calibration estimator

\[
\hat{F}_{yc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t - y_k),
\]

where the new weights \( \omega_k \) are modified from \( d_k = 1/\pi_k \) by minimizing the chi-square distance measure

\[
\Phi_s = \sum_{k \in s} \frac{(\omega_k - d_k)^2}{d_k q_k}
\]

(2)
subject to the calibration equations

$$\frac{1}{N} \sum_{k \in s} \omega_k A(t_j - g_k) = F_g(t_j), \quad j = 1, 2, \ldots, P. \quad (3)$$

The term $F_g(t_j)$ denotes the finite distribution function of the pseudo-variable $g$ evaluated at the point $t_j$, where $t_j$ for $j = 1, 2, \ldots, P$ are points that we choose arbitrarily and assume that $t_1 < t_2 < \cdots < t_P$.

Thus, with the new weights $\omega_k$, we would like to obtain a calibrated estimator that:

- is an asymptotically unbiased estimator;
- gives perfect estimates for the distribution function of the pseudo-variable $g$, when evaluated over the set of points $t_j$, $j = 1, 2, \ldots, P$.

Now, we denote by

$$t' = (t_1, \ldots, t_P), \quad A(t - g_k)' = (A(t_1 - g_k), \ldots, A(t_P - g_k))$$

$$F_g(t) = (F_g(t_1), \ldots, F_g(t_P))', \quad \hat{F}_{GH}(t) = (\hat{F}_{GH}(t_1), \ldots, \hat{F}_{GH}(t_P))',$$

with $\hat{F}_{GH}(t_j)$ denotes the Horvitz–Thompson estimator for the distribution function of $g$ at the point $t_j$.

It is clear that the calibration equations given by (3), can be expressed by

$$\frac{1}{N} \sum_{k \in s} \omega_k A(t - g_k) = F_g(t). \quad (4)$$

By minimizing (2) subject to (3) we obtain the new weights

$$\omega_k = d_k + \frac{d_k q_k \lambda A(t - g_k)}{N}, \quad (5)$$

where $\lambda$ is the vector of Lagrange multipliers of dimension $P$ given by

$$\lambda = N^2 (F_g(t) - \hat{F}_{GH}(t))' \cdot T^{-1}$$

assuming that the inverse of the symmetric matrix $T$

$$T = \sum_{k \in s} d_k q_k A(t - g_k) A(t - g_k)'$$

exists. The resulting estimator can be written as

$$\hat{F}_{yc}(t) = \hat{F}_{yH}(t) + (F_g(t) - \hat{F}_{GH}(t))' \cdot \hat{D}, \quad (6)$$

where

$$\hat{D} = T^{-1} \cdot \sum_{k \in s} d_k q_k A(t - g_k) A(t - y_k)$$

and the new calibration weights are

$$\omega_k = d_k + d_k q_k N (F_g(t) - \hat{F}_{GH}(t))' \cdot T^{-1} \cdot A(t - g_k). \quad (7)$$

In the next section, we study the problem of the existence of $T^{-1}$. 
3. Existence and expression of $T^{-1}$

If $T$ is singular, the calibration process does not have solution. We, therefore, seek conditions that guarantee the existence of $T^{-1}$ and then, we can obtain a new expression of $T^{-1}$. With this new expression for $T^{-1}$, we can better study the properties of the estimator $\hat{F}_Y(t)$. We remember that the points $t_j$ are ordered, that is

$$t_1 < t_2 < \cdots < t_P$$

and we consider the $g$-values of sample units in ascending order

$$g(1) \leq g(2) \leq \cdots \leq g(n-1) \leq g(n).$$

Suppose that the value $t_i$ is bigger than the first $k_i$ sample values of the variable $g$, with $k_i > k_{i-1}$ for $i = 2, \ldots, P$; $k_1 > 0$ and $k_P \leq n$. Under these considerations

$$T = \sum_{k=1}^{k_1} d_{k}q_k A(t - g_k)A(t - g_k)'$$

$$= \begin{pmatrix}
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_1} d_{k}q_k & \cdots & \sum_{k=1}^{k_1} d_{k}q_k \\
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_2} d_{k}q_k & \cdots & \sum_{k=1}^{k_2} d_{k}q_k \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_2} d_{k}q_k & \cdots & \sum_{k=1}^{k_p} d_{k}q_k 
\end{pmatrix}.$$ 

If $T^{-1}$ exists then it is a $P \times P$ symmetric matrix of the form

$$T^{-1} = \begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1P} \\
a_{21} & a_{22} & \cdots & a_{2P} \\
\vdots & \vdots & \ddots & \vdots \\
a_{P1} & a_{P2} & \cdots & a_{PP}
\end{pmatrix}$$

such that

$$\begin{pmatrix}
a_{11} & a_{12} & \cdots & a_{1P} \\
a_{21} & a_{22} & \cdots & a_{2P} \\
\vdots & \vdots & \ddots & \vdots \\
a_{P1} & a_{P2} & \cdots & a_{PP}
\end{pmatrix} \begin{pmatrix}
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_1} d_{k}q_k & \cdots & \sum_{k=1}^{k_1} d_{k}q_k \\
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_2} d_{k}q_k & \cdots & \sum_{k=1}^{k_2} d_{k}q_k \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{k=1}^{k_1} d_{k}q_k & \sum_{k=1}^{k_2} d_{k}q_k & \cdots & \sum_{k=1}^{k_p} d_{k}q_k
\end{pmatrix} = I_{P \times P}.$$

We now prove that $T^{-1}$ exists by showing that the $a_{ij}$ are well-defined.

Multiplying the first row of $T^{-1}$ by $T$, we get the following equations:

$$a_{11} \cdot \sum_{k=1}^{k_1} d_{k}q_k + a_{12} \cdot \sum_{k=1}^{k_1} d_{k}q_k + \cdots + a_{1P} \cdot \sum_{k=1}^{k_1} d_{k}q_k = 1,$$

$$a_{11} \cdot \sum_{k=1}^{k_1} d_{k}q_k + a_{12} \cdot \sum_{k=1}^{k_2} d_{k}q_k + \cdots + a_{1P} \cdot \sum_{k=1}^{k_2} d_{k}q_k = 0,$$

$$\vdots$$

$$a_{11} \cdot \sum_{k=1}^{k_1} d_{k}q_k + a_{12} \cdot \sum_{k=1}^{k_p} d_{k}q_k + \cdots + a_{1P} \cdot \sum_{k=1}^{k_p} d_{k}q_k = 0,$$

$$a_{11} \cdot \sum_{k=1}^{k_1} d_{k}q_k + a_{12} \cdot \sum_{k=1}^{k_2} d_{k}q_k + \cdots + a_{1P} \cdot \sum_{k=1}^{k_p} d_{k}q_k = 0.$$
Now, taking the difference between each equation with the previous equation until the difference between the third equation with second equation, we obtain \( a_{ij} = 0 \) for \( j = 3, 4, \ldots, P \) because

\[ k_P > k_{P-1} > k_{P-2} > \cdots > k_3 > k_2. \]

Therefore, the first and second equations take the form

\[
\begin{align*}
11 \cdot \sum_{k=1}^{k_1} d(k)q(k) + 12 \cdot \sum_{k=1}^{k_1} d(k)q(k) &= 1, \\
11 \cdot \sum_{k=1}^{k_1} d(k)q(k) + 12 \cdot \sum_{k=1}^{k_2} d(k)q(k) &= 0
\end{align*}
\]

and solving these we obtain

\[
\begin{align*}
11 &= \frac{1}{\sum_{k=1}^{k_1} d(k)q(k)}, \\
12 &= -\frac{1}{\sum_{k=k_1+1}^{k_2} d(k)q(k)}
\end{align*}
\]

where

\[
\sum_{k=1}^{k_1} d(k)q(k) \neq 0 \quad \text{and} \quad \sum_{k=k_1+1}^{k_2} d(k)q(k) \neq 0
\]

because \( k_2 > k_1 > 0 \).

Similarly, noting that \( T^{-1} \) is a symmetric matrix, we get for \( i = 2, 3, \ldots, P - 1 \) that

\[
\begin{align*}
i_{i,j} &= \frac{1}{\sum_{k=k_i+1}^{k_i+j-1} d(k)q(k)} + \frac{1}{\sum_{k=k_i+1}^{k_i+j} d(k)q(k)}, \\
i_{i,i+1} &= -\frac{1}{\sum_{k=k_i+1}^{k_i+j} d(k)q(k)}
\end{align*}
\]

with \( a_{ij} = 0 \) if \( j \neq i - 1, j \neq i \) and \( j \neq i + 1 \), where

\[
\sum_{k=k_i+1}^{k_i+j} d(k)q(k) \neq 0
\]

for all \( i = 2, 3, \ldots, P - 1 \) because \( k_{i+1} > k_i \).

Finally, multiplying the last row of \( T^{-1} \) by the last column of \( T \), we get

\[
a_{PP-1} \cdot \sum_{k=1}^{k_{P-1}} d(k)q(k) + a_{PP} \cdot \sum_{k=1}^{k_P} d(k)q(k) = 1
\]

and since \( a_{PP-1} = a_{P-1} \), we get

\[
a_{PP} = \frac{1}{\sum_{k=k_{P-1}+1}^{k_P} d(k)q(k)}.
\]

We have shown that if \( 0 < k_1 < k_2 < \cdots < k_{P-1} < k_P \leq n \) the \( a_{ij} \) are well-defined and therefore the matrix \( T^{-1} \) exists. It is of the form

\[
T^{-1} = \begin{pmatrix}
a_{11} & a_{12} & 0 & 0 & \cdots & 0 \\
a_{21} & a_{22} & a_{23} & 0 & \cdots & 0 \\
0 & a_{32} & a_{33} & \cdots & \cdots & 0 \\
0 & 0 & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \cdots & a_{PP-1} \quad a_{PP-1} \\
0 & 0 & 0 & 0 & \cdots & a_{PP-1} \quad a_{PP}
\end{pmatrix},
\]

where the \( a_{ij} \) are defined earlier.
Now that we have determined $T^{-1}$, we can substitute its entries $a_{ij}$ to obtain a new expression for the calibration weights (7). Thus, for $j = 1, \ldots, k_1$ we have

$$g(j) \leq t_1 < t_2 < \cdots < t_P.$$ 

Consequently $A(t - g(j))' = (1, \ldots, 1) = 1_P$ and then

$$\omega(j) = d(j) + d(j)q(j)N(F_g(t) - \widehat{F}_{GH}(t))' \cdot T^{-1} \cdot 1_P$$

$$= d(j) + d(j)q(j)N(F_g(t) - \widehat{F}_{GH}(t))' \cdot A,$$

with

$$A' = \left( \sum_{k=1}^{P} a_{1k}, \sum_{k=1}^{P} a_{2k}, \ldots, \sum_{k=1}^{P} a_{pk} \right).$$

From the structure of $T^{-1}$ given by (8) and considering the definition of $a_{ij}$ we derive

$$\sum_{k=1}^{P} a_{1k} = \frac{1}{\sum_{k=1}^{k_1} d(k)q(k)} \quad \text{and} \quad \sum_{k=1}^{P} a_{2k} = \cdots = \sum_{k=1}^{P} a_{pk} = 0.$$ 

The calibrated weights for the units $j = 1, \ldots, k_1$, are

$$\omega(j) = d(j) + \frac{d(j)q(j)N(F_g(t_1) - \widehat{F}_{GH}(t_1))}{\sum_{k=1}^{k_1} d(k)q(k)}, \quad j = 1, 2, \ldots, k_1.$$ 

In the same way, the weights for the sample units $j = k_{i-1} + 1, \ldots, k_i$ with $i = 2, \ldots, P$ are

$$\omega(j) = d(j) - \frac{d(j)q(j)N(F_g(t_{i-1}) - \widehat{F}_{GH}(t_{i-1}))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)q(k)} + \frac{d(j)q(j)N(F_g(t_i) - \widehat{F}_{GH}(t_i))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)q(k)}.$$ 

If $k_P < n$ we have for the units $j = k_P + 1, \ldots, n$ that

$$t_1 < t_2 < \cdots < t_P < g(j).$$

Consequently $A(t - g(j))' = 0_P$ for all $j = k_P + 1, \ldots, n$ and the weights for these units are $\omega(j) = d(j)$.

Finally, the calibration estimator $\hat{F}_{yc}(t)$ is given by the expression

$$\hat{F}_{yc}(t) = \hat{F}_{YH}(t) + \sum_{i=1}^{P} \left( F_g(t_i) - \widehat{F}_{GH}(t_{i-1}) \right) A_i,$$ 

where

$$A_1 = \frac{\sum_{k=1}^{k_1} d(k)q(k)A(t - y(k))}{\sum_{k=1}^{k_1} d(k)q(k)} - \frac{\sum_{k=k_1+1}^{k_2} d(k)q(k)A(t - y(k))}{\sum_{k=k_1+1}^{k_2} d(k)q(k)},$$

$$A_i = \frac{\sum_{k=k_{i-1}+1}^{k_i} d(k)q(k)A(t - y(k))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)q(k)} - \frac{\sum_{k=k_{i-1}+1}^{k_i+1} d(k)q(k)A(t - y(k))}{\sum_{k=k_{i-1}+1}^{k_i+1} d(k)q(k)}, \quad 2 \leq i \leq P - 1,$$

$$A_P = \frac{\sum_{k=k_{P-1}+1}^{k_p} d(k)q(k)A(t - y(k))}{\sum_{k=k_{P-1}+1}^{k_p} d(k)q(k)}.$$
4. Properties of \( \hat{F}_{yc}(t) \)

The next question is if the estimator \( \hat{F}_{yc}(t) \) is a distribution function. For this, we have to verify if the following conditions are satisfied:

(i) \( \hat{F}_{yc}(t) \) is continuous on the right.
(ii) \( \hat{F}_{yc}(t) \) is monotone nondecreasing.
(iii) \( \lim_{t \to -\infty} \hat{F}_{yc}(t) = 0, \lim_{t \to +\infty} \hat{F}_{yc}(t) = 1. \)

It is easy to verify that condition (i) is satisfied. On the other hand, \( \hat{F}_{yc}(t) \) is not monotone nondecreasing, in general and with respect to the third condition, we have

\[
\lim_{t \to -\infty} \hat{F}_{yc}(t) = \lim_{t \to -\infty} \frac{1}{N} \sum_{k \in s} \omega_k A(t - y_k) = 0
\]

and

\[
\lim_{t \to +\infty} \hat{F}_{yc}(t) = \lim_{t \to +\infty} \frac{1}{N} \sum_{k \in s} \omega_k A(t - y_k) = \frac{1}{N} \sum_{k \in s} \omega_k + \frac{1}{N} \sum_{k = k_p + 1}^{n} \omega_{(k)} = \frac{1}{N} \sum_{k = 1}^{k_p} \omega_{(k)} + \frac{1}{N} \sum_{k = k_p + 1}^{n} d_{(k)}. \]

Now, by Eqs. (3), we have

\[
\frac{1}{N} \sum_{k = 1}^{k_p} \omega_{(k)} = F_g(t_P)
\]

consequently

\[
\lim_{t \to +\infty} \hat{F}_{yc}(t) = \frac{1}{N} \sum_{k = 1}^{k_p} \omega_{(k)} + \frac{1}{N} \sum_{k = k_p + 1}^{n} d_{(k)} = F_g(t_P) + \frac{1}{N} \sum_{k = k_p + 1}^{n} d_{(k)}. \]

This value is not equal to 1 in general.

Thus, the estimator \( \hat{F}_{yc}(t) \) is not a distribution function, in general, because conditions (ii) and (iii) are not satisfied. We are going to solve these problems and we will begin with condition (ii). Since \( \hat{F}_{yc}(t) \) is a calibration estimator, that is

\[
\hat{F}_{yc}(t) = \frac{1}{N} \sum_{k \in s} \omega_k A(t - y_k)
\]

it is easy to see that \( \hat{F}_{yc}(t) \) is monotone nondecreasing if \( \omega_k \) are positive for all sample units. For this reason we shall seek conditions that guarantee \( \omega_k \geq 0. \)

If we take \( q_k = c \) for all sample units, then for \( k = 1, \ldots, k_1 \)

\[
\omega_{(k)} = d_{(k)} + \frac{d_{(k)}c N(F_g(t_1) - \hat{F}_{GH}(t_1))}{\sum_{k=1}^{k_1} d_{(k)} c} = \frac{d_{(k)} N(F_g(t_1))}{\sum_{k=1}^{k_1} d_{(k)}} \geq 0.
\]
For $k = k_{i-1} + 1, \ldots, k_i$ with $i = 2, \ldots, P$

$$
\omega_{(k)} = d(k) - \frac{d(k)cN(F_{\hat{g}}(t_{i-1}) - \hat{F}_{GH}(t_{i-1}))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)c} + \frac{d(k)cN(F_{\hat{g}}(t_i) - \hat{F}_{GH}(t_i))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)c}
$$

$$
= \frac{Nd(k)(F_{\hat{g}}(t_i) - \hat{F}_{bH}(t_{i-1}))}{\sum_{k=k_{i-1}+1}^{k_i} d(k)c} \geq 0
$$

and finally, for $k = k_P + 1, \ldots, n$ the weights are $\omega_{(k)} = d(k) \geq 0$.

Thus, the estimator $\hat{F}_{yc}(t)$ is monotone nondecreasing if $q_k = c$ for all sample units.

We know that

$$
\lim_{t \to +\infty} \hat{F}_{yc}(t) = \frac{1}{N} \lim_{t \to +\infty} \sum_{k=1}^{n} \omega_k A(t - y_k) = \frac{1}{N} \sum_{k=1}^{n} \omega_k.
$$

In order to meet the condition $\lim_{t \to +\infty} \hat{F}_{yc}(t) = 1$, we look to the calibration equations (3). By taking $t_P$ sufficiently large, $F_{\hat{g}}(t_P) = 1$ and the calibration equations give $(1/N)\sum_{k=1}^{n} \omega_k = 1$ as required. This means that $k_P = n$. Therefore, the estimator $\hat{F}_{yc}(t)$ is a distribution function if we choose $q_k = c$ for all sample units and $t_P$ sufficiently large.

5. Asymptotic behaviour of $\hat{F}_{yc}(t)$

Next we shall establish the asymptotic behaviour of $\hat{F}_{yc}(t)$ by the following results:

When the vector $\beta$ is replaced by $\hat{\beta}$, the multiple regression coefficient between $y$ and $x$ the asymptotic behaviour of $\hat{F}_{yc}(t)$ is the same as the estimator

$$
\hat{F}_{yc}^B(t) = \hat{F}_{yH}(t) + (F_{b}(t) - \hat{F}_{bH}(t))' \hat{B},
$$

with

$$
\hat{B} = \left( \sum_{k \in s} d_k q_k A(t - b_k)A(t - b_k)' \right)^{-1} \left( \sum_{k \in s} d_k q_k A(t - b_k)A(t - y_k) \right),
$$

where $b_k = \beta' x_k$, $F_{b}(t) = (F_{b}(t_1), \ldots, F_{b}(t_P))'$, $F_{b}(t_j)$ for $j = 1, \ldots, P$ denotes the distribution function of the variable $b$ and $\hat{F}_{bH}(t) = (\hat{F}_{bH}(t_1), \ldots, \hat{F}_{bH}(t_P))'$ where $\hat{F}_{bH}(t_j)$ denotes the Horvitz–Thompson estimator for the distribution function of $b$ at the point $t_j$.

To demonstrate this we assume the finite population embeds in a sequence of populations where $n$ and $N$ increase such that $(n/N) \to f$ when $n \to \infty$, and we will use a result due to Randles (1982) previously used by Rao et al. (1990) and Chambers and Dunstan (1986).

Randles study the asymptotic of some common family of statistics were it not for the fact that some vital parameters in the formulation of the statistics are unknown. He shows that if the statistic $T_n(\hat{\lambda})$ is a function of data and also uses the estimator $\hat{\lambda}$, which also is a function of data, consistently estimating the vector parameter $\lambda$, then $T_n(\hat{\lambda})$ and $T_n(\lambda)$ have the same limiting distribution provided $\partial \mu(\gamma)/\partial \gamma|_{\gamma=0} = (0, 0, \ldots, 0)$ where $\mu(\gamma) = \lim_{n \to +\infty} E_{\beta}[T_n(\gamma)]$ (\gamma is a mathematical symbol whose one particular value may be $\hat{\beta}$, the consistent estimator of $\beta$) and the expectation is taken when the true parameter is $\lambda$ (see Rao et al., 1990, p. 311).

Denote the calibration estimator by $\hat{F}_{yc}(t) = T_n(\hat{\beta}_1, \ldots, \hat{\beta}_J)$ (where $J$ is the number of auxiliary variables) and $\hat{F}_{yc}^B(t) = T_n(\beta_1, \ldots, \beta_J)$. $\hat{F}_{yc}(t)$ is a consistent estimator of vector $(\beta_1, \ldots, \beta_J)$. We find the limiting mean

$$
\mu(\gamma) = \lim_{n \to +\infty} E_{\beta}[T_n(\gamma)] = \hat{F}_y(t),
$$

where $\hat{F}_y(t)$ is the limiting value of $F_y(t)$ as $N \to \infty$. 
Therefore,

\[
\frac{\partial \mu(y)}{\partial \gamma} \bigg|_{\gamma=\beta} = \left( \frac{\partial \mu(y)}{\partial \gamma_1} \bigg|_{\gamma=\beta}, \frac{\partial \mu(y)}{\partial \gamma_2} \bigg|_{\gamma=\beta}, \ldots, \frac{\partial \mu(y)}{\partial \gamma_J} \bigg|_{\gamma=\beta} \right) = (0, 0, \ldots, 0).
\]

Using the above argument, we conclude that the distribution of \(T_n(\beta) (= \hat{F}_{yc}(t))\) is the same as that of \(T_n(\beta) (= \hat{F}^B_y(t))\). So the estimator \(\hat{F}_{yc}(t)\) is asymptotically design unbiased and has the same asymptotic variance as \(\hat{F}^B_y(t)\).

The asymptotic behaviour of the estimator \(\hat{F}^B_y(t)\) is the same that the asymptotic behaviour of the estimator

\[
\hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))'B,
\]

with

\[
B = \left( \sum_{k \in U} q_k A(t - b_k)A(t - b_k)' \right)^{-1} \left( \sum_{k \in U} q_k A(t - b_k)A(t - y_k) \right).
\]

Consequently \(\hat{F}_{yc}(t)\) is asymptotically normal and asymptotically design unbiased and its asymptotic variance is given by

\[
AV(\hat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} A_{kl}(d_kE_k)(d_lE_l),
\]

where \(A_{kl} = \pi_{kl} - \pi_k \pi_l\) and \(E_k = A(t - y_k) - A(t - b_k)'B\).

To see this, write \(\hat{F}^B_y(t)\) as follows:

\[
\hat{F}^B_y(t) = \hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))'B
\]

\[
= \hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))' (\hat{B} - B + B)
\]

\[
= \hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))'B + (F_b(t) - \hat{F}_{bH}(t))'(\hat{B} - B)
\]

Now \(\hat{F}_{bH}(t)\) is asymptotically design unbiased for \(F_b(t)\) and \(\hat{B}\) is asymptotically design unbiased for \(B\), so the product \((F_b(t) - \hat{F}_{bH}(t))'(\hat{B} - B)\) will be of lower order than the order of \(\hat{F}_{bH}(t)\). Consequently, the term \((F_b(t) - \hat{F}_{bH}(t))'(\hat{B} - B)\) is of lower order than \((F_b(t) - \hat{F}_{bH}(t))'B\). Then \(\hat{F}^B_y(t)\) is asymptotically unbiased and since the estimators \(\hat{F}_{YH}(t)\) and \(\hat{F}_{bH}(t)\) are asymptotically normal, the estimator \(\hat{F}^B_y(t)\) is asymptotically normal. The asymptotic variance of \(\hat{F}^B_y(t)\) is the same as the variance of the linearized statistic \(\hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))'B\) which is given by

\[
V\left(\hat{F}_{YH}(t) + (F_b(t) - \hat{F}_{bH}(t))'B\right) = V\left(\hat{F}_{YH}(t) + F_b(t)'B - \hat{F}_{bH}(t)'B\right)
\]

\[
= V\left(\hat{F}_{YH}(t) - \hat{F}_{bH}(t)'B\right)
\]

because \(F_b(t)'B\) is a constant.

Now

\[
\hat{F}_{YH}(t) - \hat{F}_{bH}(t)'B = \frac{1}{N} \sum_{k \in s} d_kA(t - y_k) - \frac{1}{N} \sum_{k \in s} d_kA(t - b_k)'B
\]

\[
= \frac{1}{N} \sum_{k \in s} d_k[A(t - y_k) - A(t - b_k)'B] = \frac{1}{N} \sum_{k \in s} d_kE_k,
\]

with \(E_k = A(t - y_k) - A(t - b_k)'B\).
Thus, the asymptotic variance of $\hat{F}_{yc}(t)$ is
\[
AV(\hat{F}_{yc}(t)) = V\left(\frac{1}{N}\sum_{k \in s} d_k E_k\right) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} A_{kl}(d_k E_k)(d_l E_l).
\]

So the asymptotic variance of $\hat{F}_{yc}(t)$ is given by (11).

An estimator for this variance (following Deville and Särndal, 1992), is
\[
\hat{V}(\hat{F}_{yc}(t)) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{A_{kl}}{\pi_{kl}} (\omega_k e_k)(\omega_l e_l),
\]
where $\omega_k$ are the calibrated weights (7) and $e_k = A(t - y_k) - A(t - g_k)'D$ with
\[
\hat{D} = \left(\sum_{k \in s} d_k q_k A(t - g_k)A(t - g_k)\right)^{-1} \left(\sum_{k \in s} d_k q_k A(t - g_k)A(t - y_k)\right).
\]

6. Simulation studies

In this section we compare the precision of the estimator $\hat{F}_{yc}(t)$ with the following estimators: $\hat{F}_{CD}$ (Chambers and Dunstan, 1986), $\hat{F}_{RK}(t)$ (Rao et al., 1990), $\hat{F}_{ps}(t)$, (Silva and Skinner, 1995) the difference estimator $\hat{F}_{d}(t)$ and the ratio estimator $\hat{F}_{R}(t)$. To do it, we consider two simulation studies, the first simulation involves a population with a good linear relation between $x$ and $y$ while the other does not.

The first study was carried out with a natural population called FAM1500. The population consists of 1500 families living in an Andalusian province. The variable of interest, $y$, denotes the cost of food, and two auxiliary variables $x_1$ and $x_2$ denote family income and other costs, respectively. These data have been taken from Fernández and Mayor (1994). Both auxiliary variables were used in the calibration process to obtain the variable $y$ at the points $Q_y(0.25)$, $Q_y(0.5)$ and $Q_y(0.75)$. Thus, it seems reasonable to use the corresponding values of the variable $y$ in the calibration equations.

In the construction of the estimator $\hat{F}_{yc}(t)$, we have chosen the following points $t_i$:
\[
t_1 = Q_y(0.25), \quad t_2 = Q_y(0.5), \quad t_3 = Q_y(0.75), \quad t_4 = \max_{k \in U} g_k.
\]

The reasons for these choices are

- In the Section 4, we saw that to guarantee $\lim_{t \to +\infty} \hat{F}_{yc}(t) = 1$, it is sufficient to take a very large value for $t_4$ and for this reason we choose $t_4 = \max_{k \in U} g_k$.
- The points $t_1$, $t_2$, and $t_3$ are selected because the simulation is based on the estimation of the distribution function $F_y(t)$ at the points $Q_y(0.25)$, $Q_y(0.5)$ and $Q_y(0.75)$. Thus, it seems reasonable to use the corresponding values of the variable $g$ in the calibration equations.

Finally, the constants $q_k$ were taken to be 1, since with this choice, we guarantee that $\hat{F}_{yc}(t)$ is monotone nondecreasing, as we have shown in Section 4.

We selected 1000 samples for three different sample sizes, $n = 75$, $n = 100$ and $n = 125$ under simple random sampling without replacement (SRSWOR) and under probability proportional to the $x_2$ variable, with Midzuno sampling. For each estimator, we calculated estimates of the distribution function of the study variable $y$ at the points $Q_y(0.25)$, $Q_y(0.5)$ and $Q_y(0.75)$. Thus, for each estimator we have 1000 estimates of $F_y(t)$ for $t = Q_y(0.25)$, $Q_y(0.5)$ and $Q_y(0.75)$. With these estimates we calculated the relative bias and the relative root mean square errors of the estimators. For any estimator $\hat{F}_y(t)$, say, we define the relative bias as
\[
\frac{1}{\hat{F}_y(t)} \left[ \frac{1}{1000} \sum_{r=1}^{1000} (\hat{F}_y^{(r)}(t) - \hat{F}_y(t)) \right]
\]
Table 1
Relative bias (RB) and relative root mean square error (RMSE) of estimators for the FAM1500 population under simple random sampling

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB $Q_y(0.25)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.5)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.75)$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{F}_{CD}$</td>
<td>-0.014638</td>
<td>0.229165</td>
<td>-0.004426</td>
<td>0.132019</td>
<td>-0.001994</td>
<td>0.075569</td>
</tr>
<tr>
<td>$\hat{F}_d$</td>
<td>0.016156</td>
<td>0.370702</td>
<td>0.005101</td>
<td>0.185967</td>
<td>0.003501</td>
<td>0.117057</td>
</tr>
<tr>
<td>$\hat{F}_R$</td>
<td>0.031181</td>
<td>0.312104</td>
<td>0.015107</td>
<td>0.190767</td>
<td>0.003162</td>
<td>0.140633</td>
</tr>
<tr>
<td>$\hat{F}_{ps}$</td>
<td>-0.008356</td>
<td>0.247420</td>
<td>-0.002157</td>
<td>0.141352</td>
<td>0.001993</td>
<td>0.075654</td>
</tr>
<tr>
<td>$\hat{F}_{RKM}$</td>
<td>-0.008585</td>
<td>0.247420</td>
<td>-0.002062</td>
<td>0.141352</td>
<td>0.001993</td>
<td>0.075654</td>
</tr>
<tr>
<td>$\hat{F}_{yc}$</td>
<td>-0.005932</td>
<td>0.228999</td>
<td>0.000766</td>
<td>0.129433</td>
<td>0.003848</td>
<td>0.075654</td>
</tr>
<tr>
<td>$n = 100$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{F}_{CD}$</td>
<td>-0.007418</td>
<td>0.188484</td>
<td>-0.000757</td>
<td>0.110151</td>
<td>-0.004358</td>
<td>0.063357</td>
</tr>
<tr>
<td>$\hat{F}_d$</td>
<td>0.000965</td>
<td>0.286768</td>
<td>-0.000223</td>
<td>0.153958</td>
<td>-0.000697</td>
<td>0.095592</td>
</tr>
<tr>
<td>$\hat{F}_R$</td>
<td>0.013981</td>
<td>0.237792</td>
<td>0.007419</td>
<td>0.158029</td>
<td>-0.002592</td>
<td>0.111917</td>
</tr>
<tr>
<td>$\hat{F}_{ps}$</td>
<td>-0.000219</td>
<td>0.197765</td>
<td>-0.000776</td>
<td>0.115670</td>
<td>0.001166</td>
<td>0.066195</td>
</tr>
<tr>
<td>$\hat{F}_{RKM}$</td>
<td>-0.002198</td>
<td>0.195597</td>
<td>-0.000665</td>
<td>0.114646</td>
<td>0.001236</td>
<td>0.065187</td>
</tr>
<tr>
<td>$\hat{F}_{yc}$</td>
<td>0.001112</td>
<td>0.180277</td>
<td>0.000509</td>
<td>0.103749</td>
<td>-0.00130</td>
<td>0.060203</td>
</tr>
<tr>
<td>$n = 125$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{F}_{CD}$</td>
<td>-0.002036</td>
<td>0.163754</td>
<td>-0.000554</td>
<td>0.097284</td>
<td>-0.001503</td>
<td>0.055954</td>
</tr>
<tr>
<td>$\hat{F}_d$</td>
<td>0.006197</td>
<td>0.260715</td>
<td>-0.000591</td>
<td>0.136630</td>
<td>0.001134</td>
<td>0.084079</td>
</tr>
<tr>
<td>$\hat{F}_R$</td>
<td>0.018999</td>
<td>0.211960</td>
<td>0.006875</td>
<td>0.142094</td>
<td>0.002082</td>
<td>0.098520</td>
</tr>
<tr>
<td>$\hat{F}_{ps}$</td>
<td>0.005376</td>
<td>0.172035</td>
<td>-0.000081</td>
<td>0.099769</td>
<td>0.002909</td>
<td>0.058328</td>
</tr>
<tr>
<td>$\hat{F}_{RKM}$</td>
<td>0.006573</td>
<td>0.171634</td>
<td>-0.000095</td>
<td>0.098873</td>
<td>0.002659</td>
<td>0.057863</td>
</tr>
<tr>
<td>$\hat{F}_{yc}$</td>
<td>0.005349</td>
<td>0.158406</td>
<td>-0.001436</td>
<td>0.089597</td>
<td>0.002075</td>
<td>0.054663</td>
</tr>
</tbody>
</table>

and the relative root mean square error as

$$\frac{1}{F_y(t)} \left[ \frac{1}{1000} \sum_{r=1}^{1000} \left( \hat{F}_y^{(r)}(t) - F_y(t) \right)^2 \right]^{1/2},$$

where $\hat{F}_y^{(r)}(t)$ is the $r$th estimate at the point $t$ calculated with the estimator $\hat{F}_y(t)$.

In general, the conclusions are similar for both types of sampling (Tables 1 and 2):

- The Chambers Dunstan estimator and the calibration estimator $\hat{F}_{yc}(t)$ gives the lowest RMSE, but the RB of the Chambers Dunstan estimator is usually larger than the RB of the calibration estimator.
- The ratio and difference estimators have the largest RMSE in all cases.
- The performance of the calibration estimator is very good. In general, the RB and RMSE are smaller than those of the others estimators.

We have also included a second simulation with a population that not involves a good linear relation between $x$ and $y$ to see the effect of the departure from the assumption of a linear relation between the study variable and the auxiliary vector on these estimators. Our procedure for producing $\hat{F}_{yc}(t)$ can be applied to any population regardless of the underlying relation between $y$ and $x$. The population is a small population of 80 factories (Murthy, 1967) which has also been studied by Kuk and Mak (1989, 1994). The variable of interest is output and the auxiliary variable is the number of workers. An examination of the scatter plot (Fig. 1) indicates that the linearity assumption is no longer valid.

We selected 1000 samples for sample sizes $n = 25$, $n = 30$ and $n = 35$ under simple random sampling without replacement (SRSWOR) and under probability proportional to the variable fixed capital, with Midzuno sampling. Tables 3 and 4 show results for this population.
Table 2
Relative bias (RB) and relative root mean square error (RMSE) of estimators for the FAM1500 population under Midzuno sampling

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB (Q_y(0.25))</th>
<th>RMSE</th>
<th>RB (Q_y(0.5))</th>
<th>RMSE</th>
<th>RB (Q_y(0.75))</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{F}_{CD})</td>
<td>(-0.038899)</td>
<td>0.179382</td>
<td>(-0.082503)</td>
<td>0.129033</td>
<td>(-0.054316)</td>
<td>0.081414</td>
</tr>
<tr>
<td>(\hat{F}_d)</td>
<td>(-0.038505)</td>
<td>0.293505</td>
<td>(-0.086254)</td>
<td>0.167089</td>
<td>(-0.054957)</td>
<td>0.106154</td>
</tr>
<tr>
<td>(\hat{F}_R)</td>
<td>(-0.018991)</td>
<td>0.235478</td>
<td>(-0.079070)</td>
<td>0.165719</td>
<td>(-0.058354)</td>
<td>0.120820</td>
</tr>
<tr>
<td>(\hat{F}_{ps})</td>
<td>(-0.032524)</td>
<td>0.188579</td>
<td>(-0.080719)</td>
<td>0.130605</td>
<td>(-0.051899)</td>
<td>0.081051</td>
</tr>
<tr>
<td>(\hat{F}_{yc})</td>
<td>(-0.032921)</td>
<td>0.178404</td>
<td>(-0.081050)</td>
<td>0.124473</td>
<td>(-0.051380)</td>
<td>0.078090</td>
</tr>
</tbody>
</table>

\(n = 75\)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB (Q_y(0.25))</th>
<th>RMSE</th>
<th>RB (Q_y(0.5))</th>
<th>RMSE</th>
<th>RB (Q_y(0.75))</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{F}_{CD})</td>
<td>(-0.008455)</td>
<td>0.197444</td>
<td>(-0.000969)</td>
<td>0.134165</td>
<td>(-0.004669)</td>
<td>0.064772</td>
</tr>
<tr>
<td>(\hat{F}_d)</td>
<td>0.001068</td>
<td>0.308159</td>
<td>(-0.000368)</td>
<td>0.196488</td>
<td>(-0.000706)</td>
<td>0.097587</td>
</tr>
<tr>
<td>(\hat{F}_R)</td>
<td>0.016986</td>
<td>0.247839</td>
<td>(-0.009846)</td>
<td>0.177927</td>
<td>(-0.002795)</td>
<td>0.151812</td>
</tr>
<tr>
<td>(\hat{F}_{ps})</td>
<td>(-0.003688)</td>
<td>0.207379</td>
<td>(-0.009917)</td>
<td>0.167786</td>
<td>(-0.001583)</td>
<td>0.078497</td>
</tr>
<tr>
<td>(\hat{F}_{yc})</td>
<td>0.001567</td>
<td>0.194275</td>
<td>(-0.000575)</td>
<td>0.126788</td>
<td>(-0.000156)</td>
<td>0.066843</td>
</tr>
</tbody>
</table>

\(n = 100\)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB (Q_y(0.25))</th>
<th>RMSE</th>
<th>RB (Q_y(0.5))</th>
<th>RMSE</th>
<th>RB (Q_y(0.75))</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{F}_{CD})</td>
<td>(-0.002156)</td>
<td>0.166899</td>
<td>0.000587</td>
<td>0.104402</td>
<td>(-0.001769)</td>
<td>0.055663</td>
</tr>
<tr>
<td>(\hat{F}_d)</td>
<td>0.006575</td>
<td>0.274758</td>
<td>(-0.000624)</td>
<td>0.167745</td>
<td>0.001466</td>
<td>0.097655</td>
</tr>
<tr>
<td>(\hat{F}_R)</td>
<td>0.020665</td>
<td>0.224754</td>
<td>0.008438</td>
<td>0.174843</td>
<td>(-0.002247)</td>
<td>0.104729</td>
</tr>
<tr>
<td>(\hat{F}_{ps})</td>
<td>0.006417</td>
<td>0.194296</td>
<td>(-0.000999)</td>
<td>0.115843</td>
<td>0.003245</td>
<td>0.064746</td>
</tr>
<tr>
<td>(\hat{F}_{yc})</td>
<td>0.007089</td>
<td>0.193666</td>
<td>(-0.000147)</td>
<td>0.124623</td>
<td>0.003893</td>
<td>0.069425</td>
</tr>
<tr>
<td>(\hat{F}_{RKM})</td>
<td>0.005783</td>
<td>0.167263</td>
<td>(-0.001677)</td>
<td>0.094667</td>
<td>0.002239</td>
<td>0.058249</td>
</tr>
</tbody>
</table>

From our simulation study we observed the following:

- The \(\hat{F}_{CD}\) estimator has a serious problem of bias, especially for the lower quantiles of the distribution, something already noted by Chambers and Dunstan (1986), Kuk and Mak (1994) and Silva and Skinner (1995). This is expected because the relation between \(y\) (output) and \(x\) (number of workers) is not linear over this population.
- We found no evidence of any significant bias for the other estimators considered.
- In terms of efficiency the best overall performance is achieved by our proposed calibration estimator. It seems from Table 3 that the calibration estimator does remarkably better than the other estimators.
### Table 3
Relative bias (RB) and relative root mean square error (RMSE) of estimators for the Murthy population under simple random sampling

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB $Q_y(0.25)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.5)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.75)$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 25$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.167226</td>
<td>0.260399</td>
<td>0.159317</td>
<td>0.168082</td>
<td>0.049461</td>
<td>0.053308</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>0.014480</td>
<td>0.322890</td>
<td>0.000425</td>
<td>0.120465</td>
<td>-0.000530</td>
<td>0.047615</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>0.006647</td>
<td>0.273522</td>
<td>-0.001902</td>
<td>0.116168</td>
<td>-0.001299</td>
<td>0.050640</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>0.008002</td>
<td>0.280144</td>
<td>-0.001057</td>
<td>0.109349</td>
<td>0.001029</td>
<td>0.067566</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.003911</td>
<td>0.220026</td>
<td>0.007204</td>
<td>0.089002</td>
<td>0.002950</td>
<td>0.036948</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>-0.000396</td>
<td>0.112105</td>
<td>-0.000238</td>
<td>0.056157</td>
<td>-0.000001</td>
<td>0.000001</td>
</tr>
<tr>
<td>$n = 30$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.15001</td>
<td>0.220659</td>
<td>0.148933</td>
<td>0.156003</td>
<td>0.045931</td>
<td>0.049375</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>0.01401</td>
<td>0.279089</td>
<td>0.006550</td>
<td>0.100580</td>
<td>-0.000677</td>
<td>0.041265</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>0.00728</td>
<td>0.221477</td>
<td>0.004106</td>
<td>0.097708</td>
<td>-0.001453</td>
<td>0.043217</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>0.00830</td>
<td>0.225827</td>
<td>0.007471</td>
<td>0.091010</td>
<td>0.000343</td>
<td>0.061385</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.00517</td>
<td>0.175653</td>
<td>0.010511</td>
<td>0.076623</td>
<td>0.001411</td>
<td>0.033367</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>0.000096</td>
<td>0.096459</td>
<td>0.000588</td>
<td>0.049719</td>
<td>-0.000001</td>
<td>0.000001</td>
</tr>
<tr>
<td>$n = 35$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.142120</td>
<td>0.202902</td>
<td>0.131710</td>
<td>0.138492</td>
<td>0.041756</td>
<td>0.044773</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>0.004600</td>
<td>0.241171</td>
<td>0.000877</td>
<td>0.091489</td>
<td>0.000432</td>
<td>0.035610</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>-0.000590</td>
<td>0.197894</td>
<td>-0.000738</td>
<td>0.087631</td>
<td>-0.000041</td>
<td>0.037408</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>-0.002080</td>
<td>0.200459</td>
<td>-0.000603</td>
<td>0.081556</td>
<td>-0.001433</td>
<td>0.050769</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.006930</td>
<td>0.157648</td>
<td>0.004265</td>
<td>0.067272</td>
<td>0.001404</td>
<td>0.029070</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>-0.001270</td>
<td>0.084535</td>
<td>0.000712</td>
<td>0.041950</td>
<td>-0.000001</td>
<td>0.000001</td>
</tr>
</tbody>
</table>

### Table 4
Relative bias (RB) and relative root mean square error (RMSE) of estimators for the Murthy population under Midzuno sampling

<table>
<thead>
<tr>
<th>Estimator</th>
<th>RB $Q_y(0.25)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.5)$</th>
<th>RMSE</th>
<th>RB $Q_y(0.75)$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 25$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.176420</td>
<td>0.266382</td>
<td>0.161910</td>
<td>0.170437</td>
<td>0.050743</td>
<td>0.054681</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>0.009263</td>
<td>0.325295</td>
<td>0.004111</td>
<td>0.117849</td>
<td>-0.001460</td>
<td>0.046568</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>0.000946</td>
<td>0.271518</td>
<td>0.000838</td>
<td>0.115906</td>
<td>-0.002253</td>
<td>0.048933</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>-0.002036</td>
<td>0.272810</td>
<td>0.002935</td>
<td>0.109373</td>
<td>-0.004694</td>
<td>0.066093</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.001768</td>
<td>0.212160</td>
<td>0.010487</td>
<td>0.089525</td>
<td>0.002575</td>
<td>0.037923</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>-0.001431</td>
<td>0.111640</td>
<td>-0.001481</td>
<td>0.059442</td>
<td>-0.000001</td>
<td>0.000001</td>
</tr>
<tr>
<td>$n = 30$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.168468</td>
<td>0.238171</td>
<td>0.144295</td>
<td>0.151880</td>
<td>0.045918</td>
<td>0.049346</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>-0.011503</td>
<td>0.276318</td>
<td>0.004444</td>
<td>0.108010</td>
<td>-0.001507</td>
<td>0.039559</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>-0.013671</td>
<td>0.223805</td>
<td>-0.002444</td>
<td>0.099612</td>
<td>-0.002448</td>
<td>0.041248</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>-0.015207</td>
<td>0.227757</td>
<td>-0.003475</td>
<td>0.095787</td>
<td>-0.002614</td>
<td>0.057302</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.017921</td>
<td>0.179677</td>
<td>0.003379</td>
<td>0.078160</td>
<td>0.001755</td>
<td>0.032837</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>-0.003355</td>
<td>0.098952</td>
<td>-0.001988</td>
<td>0.048112</td>
<td>0.000001</td>
<td>0.000001</td>
</tr>
<tr>
<td>$n = 35$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{CD}$</td>
<td>-0.142530</td>
<td>0.203590</td>
<td>0.131673</td>
<td>0.138624</td>
<td>0.041678</td>
<td>0.044724</td>
</tr>
<tr>
<td>$F_{d}$</td>
<td>0.007182</td>
<td>0.246550</td>
<td>-0.001043</td>
<td>0.087145</td>
<td>0.001121</td>
<td>0.035710</td>
</tr>
<tr>
<td>$F_{R}$</td>
<td>-0.001356</td>
<td>0.199984</td>
<td>-0.001023</td>
<td>0.087943</td>
<td>-0.000674</td>
<td>0.036955</td>
</tr>
<tr>
<td>$F_{ps}$</td>
<td>-0.001356</td>
<td>0.199984</td>
<td>-0.001023</td>
<td>0.087943</td>
<td>-0.000674</td>
<td>0.036955</td>
</tr>
<tr>
<td>$F_{RKM}$</td>
<td>-0.007557</td>
<td>0.158813</td>
<td>0.004122</td>
<td>0.067913</td>
<td>0.001318</td>
<td>0.028976</td>
</tr>
<tr>
<td>$F_{yc}$</td>
<td>-0.000584</td>
<td>0.084259</td>
<td>0.000147</td>
<td>0.041959</td>
<td>-0.000001</td>
<td>0.000001</td>
</tr>
</tbody>
</table>
The main finding is that both estimators, $\hat{F}_{RMK}$ and $\hat{F}_{Yc}$ estimators perform satisfactorily. Their biases are negligible and their RSME values are small. Note that the bias and RSME of the calibration estimator at the 75th percentile are really low for all sizes and sampling methods.

7. Conclusions

In this paper we have used the calibration technique to provide a new estimator for the distribution function. We have compared this estimator with several other estimators of the finite population distribution. The calibration estimator possesses a number of desirable properties, such as yielding a genuine distribution function, asymptotic unbiasedness, asymptotic normality, availability of variance estimator and simplicity of computation. Four other estimators (ratio, difference, poststratified and Rao, Kovar and Mantel estimators) possess some of these properties, but appear to be inferior in terms of efficiency. Chambers and Dunstan’s estimator can be very efficient when the model upon which it is based is appropriate, but as noted by Rao et al. (1990) and Dorfman and Hall (1993), this estimator can perform poorly under model misspecification. In many surveys detailed model checking cannot be performed, and thus the calibration estimator may be preferable.

In conclusion we suggest that in many standard survey settings, the calibration technique provides a simple and practical approach to incorporating auxiliary information into the estimation of distribution function. The calibration estimator has good performance, and is a valid alternative to other estimators of the distribution function especially if the model is not linear.

Acknowledgments

The authors would like to thank to associate editor and the referees for their many helpful comments and suggestions which helped improve the paper. Research partially supported by Ministerio de Educación y Ciencia (Spain) contract no. MTM2004-04034.

References