Journal of Computational and Applied Mathematics 233 (2010) 2265–2277

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Optimum calibration points estimating distribution functions

S. Martínez [a,*], M. Rueda [b], A. Arcos [b], H. Martínez [a]

[a] Department of Statistics and Applied Mathematics, University of Almería, Spain

[b] Department of Statistics and Operational Research, University of Granada, Spain

ARTICLE INFO

Article history:
Received 30 June 2009
Received in revised form 6 October 2009

Keywords:
Auxiliary information
Calibration technique
Distribution function estimates
Model-assisted estimation
Survey sampling

ABSTRACT

The calibration method has been widely discussed in the recent literature on survey sampling, and calibration estimators are routinely computed by many survey organizations. The calibration technique was introduced in [12] to estimate linear parameters as mean or total. Recently, some authors have applied the calibration technique to estimate the finite distribution function and the quantiles. The computationally simpler method in [14] is built by means of constraints that require the use of a fixed value t_0. The precision of the resulting calibration estimator changes with the selected point t_0. In the present paper, we study the problem of determining the optimal value t_0 that gives the best estimation under simple random sampling without replacement. A limited simulation study shows that the improvement of this optimal calibrated estimator over possible alternatives can be substantial.

1. Introduction

In survey sampling a common task is the estimation of the finite population distribution, as the finite population quantiles and some poverty measures can be obtained by means of this function.

In the presence of auxiliary information, various general estimation procedures can be used to obtain more efficient estimators of population means and totals. Work has been done to apply these general procedures directly to estimating the distribution function.

The design-based ratio $\widehat{F}_r(t)$ and difference $\widehat{F}_d(t)$ type estimators for the distribution function [1] suffer from several drawbacks, the obvious one being that they can take values outside [0, 1]. Furthermore, they are not always monotone functions and therefore are not recommended for estimating finite population quantiles. Other important papers on this topic [2,3] have assumed a superpopulation model, and have suggested model-based estimators. Careful model checking and diagnostics need to be carried out before these model-based estimators are used. Under simple random sampling, Wang and Dorfman [4] combined the Chambers–Dunstan estimator and the Rao–Kovar–Mantel estimator in a hybrid estimator, which under certain conditions is more efficient than either of the above. However, their estimator also inherits the drawbacks of the other two estimators and cannot be readily generalized to more complex designs. Other references related to estimating the distribution function are [5,6].

In the last decade, calibration estimation [7–11] has become an important field of research in survey sampling. Calibration is now an important methodological instrument in the production of statistics. It was introduced in [12] to estimate the population total, but this approach can be adapted to the estimation of more complex parameters than just a population total. Harms and Duchesne [13] and Rueda et al. [14] use different ways to implement the calibration approach in the estimation of the distribution function and the quantiles.

* Corresponding author. Fax: +34 958243267.
E-mail address: mrueda@ugr.es (S. Martínez).

0377-0427/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2009.10.011

Both methods give nearly design-unbiased estimation and compare favorably with earlier estimation methods for the distribution function, not based on the calibration approach but using the same auxiliary information (see [15]).

The computationally simpler method of Rueda et al. [14] is an application of model calibration, in that they calibrate with respect to the predicted y-values. The weights are obtained by minimizing the chi-square distance subject to calibration equations that require the use of an arbitrarily fixed value t_0. The precision of the resulting calibration estimator changes with the selected point t_0. Furthermore, this estimator undergoes a loss of efficiency when t_0 is far away from t, the point where the distribution function is being evaluated. Thus, the selection of the point t_0 is a serious problem and one not analyzed in the above mentioned work. In this paper, we study the problem of the optimal value t_0 that gives the best estimation under simple random sampling without replacement. Finally, a simulation study compares the method proposed with other conventional methods.

Section 2 summarizes Cumulative distribution function (CDF) calibration estimation. In Section 3 we derive an optimum calibration point when estimation by CDF calibration is based on one point. Section 4 is similar to the previous section but with estimation by CDF calibration based on two points. A brief simulation study is performed and our conclusions are reported in Section 5.

2. Calibration estimation of 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 π_k and π_{kl} assumed to be strictly positive. Let y_k be the value of the study variable y, for the kth population element, with which an auxiliary value x_k is also associated. The values $x_1, x_2, \ldots, x_N$ are known for the entire population but y_k is known only if the kth unit is selected in the sample s. The finite population distribution function of the study variable y is given by $F_y(t) = \sum_{k\in U} \Delta(t - y_k)/N$ with $\Delta(t - y_k) = 1$ if $t \geq y_k$ and $\Delta(t - y_k) = 0$ otherwise. A purely design-based estimator of the distribution function is the Horvitz–Thompson estimator, defined by $\widehat{F}_{YH}(t) = \sum_{k\in s} d_k \Delta(t - y_k)/N$ with $d_k = 1/\pi_k$ describing the basic design weights. The estimator $\widehat{F}_{YH}(t)$ is unbiased, but in general, it is not a distribution function ($\lim_{t\to+\infty} \widehat{F}_{YH}(t) \neq 1$) and does not use the auxiliary information provided by the variable x.

Rueda et al. [14] consider a calibration estimator by first defining a pseudo-variable $g_k = \widehat{\beta}'\mathbf{x}_k$ for $k = 1, 2, \ldots, N$, where

$$\widehat{\beta} = \left(\sum_{k\in s} d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{k\in s} d_k q_k \mathbf{x}_k y_k \tag{1}$$

is a weighted estimator of the multiple regression coefficient β between y and $\mathbf{x}$. The q_k are known positive constants unrelated to $d_k = 1/\pi_k$. They then define the calibration estimator

$$\widehat{F}_{yc}(t) = \frac{1}{N} \sum_{k\in s} \omega_k \Delta(t - y_k)$$

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

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

subject to the calibration equations

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

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 chosen arbitrarily, assuming that

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

The resulting estimator can be written as

$$\widehat{F}_{yc}(t) = \widehat{F}_{YH}(t) + \left(F_g(\mathbf{t}) - \widehat{F}_{GH}(\mathbf{t})\right)' \widehat{D} \tag{4}$$

where $\mathbf{t} = (t_1, \ldots, t_p)'$, $\widehat{F}_{GH}$ is the Horvitz–Thompson estimator of F_g and

$$\widehat{D} = T^{-1} \sum_{k\in s} d_k q_k \Delta(\mathbf{t} - g_k) \Delta(t - y_k)$$

assuming that the inverse of symmetric matrix T

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

exists.

This estimator is a distribution function (continuous on the right, monotone nondecreasing and $\lim_{t\to-\infty}\widehat{F}_{yc}(t)=0$; $\lim_{t\to+\infty}\widehat{F}_{yc}(t)=1$) and gives a nearly design-unbiased estimation. The estimator gives perfect estimates of the distribution function $F_g(t)$ in t_j; $j=1,2,\ldots,P$, but with these conditions we have assumed that the study variable y and the auxiliary vector $\mathbf{x}$ are linearly related.[1]

The calibrated estimator uses the same $\mathbf{t}$ for any t in $F(t)$. If different values $\mathbf{t}$ are used to estimate various arguments t, nonmonotonicity is unsatisfying from a theoretical point of view and causes complications in the inverse problem of quantile estimation. The solution to this issue is to consider the same calibration point. This optimum for the fixed point is obtained in Section 3.

3. Optimal point for estimating the cumulative distribution function by one-point calibration

In this section, we study the optimal choice of the point in the calibration equations (3). We consider the case where one point $\mathbf{t}=t_0$ is used in the calibration estimator $\widehat{F}_{yc}(t)$ under simple random sampling. The following theorem summarizes the main result.

Theorem 1. *Let $g_k=\widehat{\beta}'\mathbf{x}_k$, $k=1,\ldots,N$ and F_g represent the finite population distribution function of g-values. Suppose that we wish to estimate F_y at point t. Let*

$$G(\gamma)=\frac{B}{N-1}\left[2NF_y(t)F_g(\gamma)-(1+F_g(\gamma))\sum_{k\in U}\Delta(t-y_k)\Delta(\gamma-g_k)\right]$$

where

$$B=\sum_{k\in U}\Delta(t-y_k)\Delta(\gamma-g_k)\Big/\sum_{k\in U}\Delta(\gamma-g_k).$$

y-values are denoted in ascending order by $y_{[1]},\ldots,y_{[N]}$ and g-values are arranged by the y-variable by $g_{[1]},\ldots,g_{[N]}$. Let $A_t=\{g_{[k]}\mid k\in U, y_{[k]}\le t\}$.

Then, the point at which the calibration estimator is optimum is given by

$$t_{\text{opt}}=\operatorname*{argmin}_{g_k\in A_t}G(g_k).$$

A proof of the above result is as follows. In a first step, the variance of the $\widehat{F}_{yc}(t)$ estimator is expressed as a real function $G(\gamma)$. In a second step, the $G(\gamma)$ properties are studied and, finally, we find the optimum t_0.

Variance of the $\widehat{F}_{yc}(t)$ estimator

The estimator $\widehat{F}_{yc}(t)$ is given by

$$\widehat{F}_{yc}(t)=\widehat{F}_{YH}(t)+\left(F_g(t_0)-\widehat{F}_{GH}(t_0)\right)B_s \tag{6}$$

where

$$B_s=\frac{\sum_{k\in s}d_kq_k\Delta(t-y_k)\Delta(t_0-g_k)}{\sum_{k\in s}d_kq_k\Delta(t_0-g_k)}.$$

Rueda et al. [14] show that the asymptotic behaviour of $\widehat{F}_{yc}(t)$ is the same as the estimator

$$F_{YH}(t)+\left(F_g(t_0)-\widehat{F}_{GH}(t_0)\right)\cdot B$$

[1] Note. If the relation between y and $\mathbf{x}$ is not linear, the pseudo-variable g_k and the condition (3) are inadequate and it is necessary to adapt or to modify these conditions for nonlinear models.

Rueda et al. [16] assume that the relationship between y and $\mathbf{x}$ can be described by the following superpopulation model (the linear or nonlinear regression model)

$$y_k=\mu(\mathbf{x}_k,\boldsymbol{\theta})+\nu_k\varepsilon_k\quad k=1,2,\ldots,N \tag{5}$$

where $\boldsymbol{\theta}=(\theta_0,\ldots,\theta_J)'$ and σ^2 are unknown superpopulation parameters, $\mu(\mathbf{x}_k,\boldsymbol{\theta})$ is a known function of $\mathbf{x}$ and $\boldsymbol{\theta}$, the $\nu_k=\nu(\mathbf{x}_k)$ is a strictly positive known function of $\mathbf{x}_k$, and the ε_k are independently and identically distributed random variables with

$$E_\xi(\varepsilon_k)=0\quad\text{and}\quad V_\xi(\varepsilon_k)=\sigma^2$$

where E_ξ and V_ξ denote the expectation and variance with respect to the superpopulation model. Under this model the authors propose an alternative estimator based on nonparametric regression.

with

$$B = \frac{\sum_{k\in U} q_k \Delta(t - y_k)\Delta(t_0 - g_k)}{\sum_{k\in U} q_k \Delta(t_0 - g_k)}.$$

Consequently, the asymptotic variance of $\widehat{F}_{yc}(t)$ is

$$V(\widehat{F}_{yc}(t)) = V(\widehat{F}_{YH}(t)) + B^2 V(\widehat{F}_{GH}(t_0)) - 2B\,\mathrm{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_0)).$$

Thus, we must choose the value of t_0 that minimizes the following expression

$$B^2 V(\widehat{F}_{GH}(t_0)) - 2B\,\mathrm{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_0)). \tag{7}$$

Expression (7) can be written in the following way

$$B\left[BV(\widehat{F}_{GH}(t_0)) - 2\mathrm{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_0))\right]. \tag{8}$$

Under simple random sampling, we have:

$$V(\widehat{F}_{GH}(t_0)) = \frac{N}{N-1} F_g(t_0)\left(1 - F_g(t_0)\right) \tag{9}$$

$$\mathrm{Cov}(\widehat{F}_{GH}(t_0), \widehat{F}_{YH}(t)) = \frac{\sum_{k\in U} \Delta(t - y_k)\Delta(t_0 - g_k) - NF_y(t)F_g(t_0)}{N-1}. \tag{10}$$

Let $H = \sum_{k\in U} \Delta(t - y_k)\Delta(t_0 - g_k)$. Expressions (9) and (10) can be replaced in (8)

$$B\left[B\left[\frac{N}{N-1} F_g(t_0)\left(1 - F_g(t_0)\right)\right] - \frac{2}{N-1}\left(H - NF_y(t)F_g(t_0)\right)\right]. \tag{11}$$

Now, we consider the term

$$\begin{aligned} &B\left[\frac{N}{N-1} F_g(t_0)\left(1 - F_g(t_0)\right)\right] - \frac{2}{N-1}\left(H - NF_y(t)F_g(t_0)\right) \\ &\quad = \frac{1}{N-1}\left[B\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right) - 2\left(H - NF_y(t)F_g(t_0)\right)\right] \end{aligned} \tag{12}$$

if we take into account that:

$$B\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right) = \frac{\sum_{k\in U} q_k \Delta(t - y_k)\Delta(t_0 - g_k)}{\sum_{k\in U} q_k \Delta(t_0 - g_k)}\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right)$$

and we choose $q_k = 1$ for all $k \in U$ (this choice guarantees that the calibration estimator is monotone nondecreasing, see [14]), the previous expression takes the form

$$\begin{aligned} \frac{\sum_{k\in U} q_k \Delta(t - y_k)\Delta(t_0 - g_k)}{\sum_{k\in U} q_k \Delta(t_0 - g_k)}\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right) &= \frac{\sum_{k\in U} \Delta(t - y_k)\Delta(t_0 - g_k)}{NF_g(t_0)}\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right) \\ &= \left(1 - F_g(t_0)\right) H \end{aligned}$$

and

$$B\left(NF_g(t_0)\left(1 - F_g(t_0)\right)\right) = \left(1 - F_g(t_0)\right) H. \tag{13}$$

Next, if we replace (13) in (12), we obtain

$$\frac{1}{N-1}\left[\left(1 - F_g(t_0)\right) H - 2\left(H - NF_y(t)F_g(t_0)\right)\right] = \frac{1}{N-1}\left[2NF_y(t)F_g(t_0) - (1 + F_g(t_0))H\right]. \tag{14}$$

Finally, if we replace (14) in (11), we have

$$G(t_0) = \frac{B}{N-1}\left[2NF_y(t)F_g(t_0) - (1 + F_g(t_0))\sum_{k\in U} \Delta(t - y_k)\Delta(t_0 - g_k)\right]. \tag{15}$$

Properties of $G(\gamma)$

Therefore, we must obtain the value of t_0 that minimizes the function $G(t_0)$ given in (15). To do so, let us consider the population values of the study variable y in ascending order

$$y_{[1]} \leq y_{[2]} \leq \cdots \leq y_{[N]}. \tag{16}$$

Next, we consider the population values of the variable g, arranged by the variable y, that is

$$g_{[1]}; g_{[2]}; \ldots; g_{[N]} \tag{17}$$

where, for example, the value $g_{[1]}$ is not the smallest value of the variable g but the value of the variable g that corresponds to the value $y_{[1]}$, $g_{[2]}$ corresponds to the value $y_{[2]}$, . . ., that is

$$(y_{[1]}, g_{[1]});\ (y_{[2]}, g_{[2]}); \ldots;\ (y_{[N]}, g_{[N]}).$$

Thus, if we wish to estimate the distribution function F_y at the value t, we must prove that the function $G(t_0)$ reaches the minimum at a point in the set

$$A_t = \{g_{[k]} : k \in C_t\} \tag{18}$$

where C_t is the following set

$$C_t = \left\{k \in \{1, 2, \ldots, N\} : y_{[k]} \leq t\right\}. \tag{19}$$

The set A_t is finite and we suppose that $a_1 < a_2 < \cdots < a_P$ are the P different elements of A_t in ascending order.

Now, we study the choice of t_0 with the function $G(t_0)$, which can take various values:

(1) If $t_0 < a_1$ the function $G(t_0)$ takes the value 0, because

$$\sum_{k\in U} \Delta(t - y_{[k]})\Delta(t_0 - g_{[k]}) = 0.$$

(2) If $t_0 \geq a_P$, we have

$$\sum_{k\in U} \Delta(t - y_{[k]})\Delta(t_0 - g_{[k]}) = NF_y(t)$$

and

$$B = \frac{NF_y(t)}{NF_g(t_0)}.$$

Then, the function $G(t_0)$ is

$$\begin{aligned} G(t_0) &= \frac{B}{N-1}\left[2NF_y(t)F_g(t_0) - (1 + F_g(t_0))\sum_{k\in U}\Delta(t - y_{[k]})\Delta(t_0 - g_{[k]})\right] \\ &= \frac{NF_y(t)}{(N-1)NF_g(t_0)}\left[2NF_y(t)F_g(t_0) - (1 + F_g(t_0))NF_y(t)\right] \\ &= \frac{N^2(F_y(t))^2}{(N-1)NF_g(t_0)}\left[F_g(t_0) - 1\right] = \frac{N(F_y(t))^2}{(N-1)}\left[1 - \frac{1}{F_g(t_0)}\right] < 0. \end{aligned}$$

(3) Finally, we consider the case where $a_i \leq t_0 < a_{i+1}$. For $i \in \{1, \ldots, P-1\}$ we denote

$$k_i = \sum_{k\in U} \Delta(t - y_{[k]})\Delta(t_0 - g_{[k]}) = \sum_{k\in U} \Delta(t - y_{[k]})\Delta(a_i - g_{[k]}).$$

For all $i \in \{1, \ldots, P-1\}$ we have $k_i < NF_y(t)$ and $G(t_0)$ takes the value

$$\begin{aligned} G(t_0) &= \frac{k_i}{(N-1)NF_g(t_0)}\left[2NF_y(t)F_g(t_0) - (1 + F_g(t_0))k_i\right] \\ &= \frac{k_i}{(N-1)NF_g(t_0)}\left[(2NF_y(t) - k_i)F_g(t_0) - k_i\right] \\ &= \frac{k_i}{(N-1)}\left[\frac{2NF_y(t) - k_i}{N} - \frac{k_i}{NF_g(t_0)}\right]. \end{aligned}$$

Thus, $G(t_0)$ is a piecewise function given by

$$G(t_0) = \begin{cases} 0 & t_0 < a_1 \\ \dfrac{k_i}{(N-1)}\left[\dfrac{2NF_y(t) - k_i}{N} - \dfrac{k_i}{NF_g(t_0)}\right] & a_i \leq t_0 < a_{i+1} \\ \dfrac{N(F_y(t))^2}{(N-1)}\left[1 - \dfrac{1}{F_g(t_0)}\right] & a_P \leq t_0. \end{cases}$$

Now, we consider the interval $[a_P, +\infty)$ where the function $G(t_0)$ is

$$G(t_0) = \frac{N(F_y(t))^2}{(N-1)}\left[1 - \frac{1}{F_g(t_0)}\right] \quad a_P \le t_0.$$

It is apparent that $G(t_0)$ is monotone nondecreasing in $[a_P, +\infty)$. Consequently, the minimum point of $G(t_0)$ is at $t_0 = a_P$ and this minimum takes the value

$$\frac{N(F_y(t))^2}{(N-1)}\left[1 - \frac{1}{F_g(a_P)}\right] < 0.$$

If we consider the interval $[a_i, a_{i+1})$ with $i \in \{1, \ldots, P-1\}$, the function $G(t_0)$ is

$$G(t_0) = \frac{k_i}{(N-1)}\left[\frac{2NF_y(t) - k_i}{N} - \frac{k_i}{NF_g(t_0)}\right]$$

where k_i is constant for all $t_0 \in [a_i, a_{i+1})$ and equal to

$$k_i = \sum_{k \in U} \Delta(t - y_{[k]})\Delta(a_i - g_{[k]}).$$

Therefore, the function $G(t_0)$, in the interval $[a_i, a_{i+1})$, is monotone nondecreasing and its minimum point is at $t_0 = a_i$. This minimum is equal to

$$\frac{k_i}{(N-1)}\left[\frac{2NF_y(t) - k_i}{N} - \frac{k_i}{NF_g(t_0)}\right].$$

Thus, the local minimum values of $G(t_0)$ are:

- 0 at $t_0 < a_1$
- $\frac{k_i}{(N-1)}\left[\frac{2NF_y(t)-k_i}{N} - \frac{k_i}{NF_g(a_i)}\right]$ at $t_0 = a_i, i = 1, \ldots, P-1$
- $\frac{N(F_y(t))^2}{(N-1)}\left[1 - \frac{1}{F_g(a_P)}\right] < 0$ at $t_0 = a_P$.

Because the last local minimum value is negative, the value 0 of the first interval cannot be the global minimum of $G(t_0)$ and then the global minimum of the function $G(t_0)$ is at one point of $A_t = \{a_i : i = 1, \ldots P\}$.

The proposed estimator

The optimal value of t_0, (t_{opt}) depends on some unknown values, so the optimal estimator $\widehat{F}_{ycop} = \widehat{F}_{YH}(t) + (F_g(t_{\text{opt}}) - \widehat{F}_{GH}(t_{\text{opt}}))B_s$ cannot be calculated. In the absence of good a priori knowledge these characteristics, we go to replace the optimal value t_{opt} by sample-based estimates.

First, we have to estimate the function $G(t_0)$ by

$$\begin{aligned}\widehat{G}(t_0) &= \frac{\widehat{B}}{N-1}\left[2N\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_0) - (1 + \widehat{F}_{GH}(t_0))\sum_{k \in s} d_k \Delta(t - y_k)\Delta(t_0 - g_k)\right] \\ &= \frac{\widehat{B}}{N-1}\left[2N\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_0) - (1 + \widehat{F}_{GH}(t_0))\frac{N}{n}\sum_{k \in s} \Delta(t - y_k)\Delta(t_0 - g_k)\right] \\ &= \frac{\widehat{B}N}{N-1}\left[2\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_0) - (1 + \widehat{F}_{GH}(t_0))\frac{1}{n}\sum_{k \in s} \Delta(t - y_k)\Delta(t_0 - g_k)\right]\end{aligned}$$

where

$$\widehat{B} = \frac{\sum_{k \in s} \Delta(t - y_k)\Delta(t_0 - g_k)}{\sum_{k \in s} \Delta(t_0 - g_k)}.$$

In a similar way, we can prove that the global minimum of the function $\widehat{G}(t_0)$ is at one point of $A_{st} = \{g_k : k \in C_{st}\}$ with

$$C_{st} = \{k \in s : y_k \le t\}.$$

Then, with the sample s we can find the global minimum of $\widehat{G}(t_0)$; to do this, we have to calculate the image of the points in the set $A_{st} = \{g_k : k \in A_{st}\}$ under $\widehat{G}(t_0)$ and take the point whose image is less than the others.

The value which minimizes $\widehat{G}(t_0)$ is an estimator of t_{opt} and after replacing t_{opt} by this estimator $\widehat{t}_{\text{opt}}$ we obtain the proposed estimator:

$$\widehat{F}_{ycprop} = \widehat{F}_{YH}(t) + \left(F_g(\widehat{t}_{\text{opt}}) - \widehat{F}_{GH}(\widehat{t}_{\text{opt}})\right)\widehat{B}_s \tag{20}$$

being

$$\widehat{B}_s = \frac{\sum_{k\in s} \Delta(t - y_k)\Delta(\widehat{t}_{\text{opt}} - g_k)}{\sum_{k\in s} \Delta(\widehat{t}_{\text{opt}} - g_k)}.$$

Thus we consider the estimator $\widehat{F}_{ycprop}$ which is obtained by replacing the unknown values with theirs sample estimators. This estimator does not coincide with the theoretical optimum estimator $\widehat{F}_{ycop}$ but we can derive the limit distribution for such statistics using the results of Randles [17]. We embed our finite population in a sequence of populations where the sample sizes and the population sizes both increase without bound.

Following Randles' notation, we denote the estimator $\widehat{F}_{ycprop}$ as $T_n(\widehat{\gamma}) = T_n(\widehat{\gamma}_1, \widehat{\gamma}_2)$ with $\widehat{\gamma} = (\widehat{t}_{\text{opt}}, \widehat{B}_s)$ an consistent estimator of $\gamma = (t_{\text{opt}}, B_s)$. We embed our finite population in a sequence of populations where the sample sizes and the population sizes both increase without bound.

We replace $\widehat{\gamma}$ in $T_n(\cdot)$ with a variable ς. Now we calculate the limit of the expectation of the statistic $T_n(\varsigma)$ when the current value of the parameter is γ:

$$\mu(\varsigma) = \lim_{n\to+\infty} E_\gamma[T_n(\varsigma)] = \lim_{n\to+\infty} E_\gamma\left[\widehat{F}_{YH}(t) + \left(F_g(\varsigma_1) - \widehat{F}_{GH}(\varsigma_1)\right)\cdot\varsigma_2\right] = \widetilde{F}_y(t)$$

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

Then

$$\left.\frac{\partial\mu(\varsigma)}{\partial\varsigma}\right|_{\varsigma=(t_{\text{opt}},B_s)} = (0, 0).$$

We conclude that the asymptotic distribution of $T_n(\widehat{\gamma})$ (the proposed estimator, $\widehat{F}_{ycprop}$) is the same as that of $T_n(\gamma)$ (the estimator based on optimum point, $\widehat{F}_{ycop}$).

4. Optimal point for estimating the cumulative distribution function by two-point calibration

We now consider the case in which two points $\mathbf{t} = (t_1, t_2)'$ are used in the calibration estimator $\widehat{F}_{yc}(t)$ under simple random sampling. The first point t_1 is an arbitrary chosen point and the second point is $t_2 = \max_{k\in U} g_k$. There is a specific reason for choosing the point t_2. Since the estimator $\widehat{F}_{yc}(t)$ is an estimator for the distribution function $F_y(t)$, it is desirable for the estimator $\widehat{F}_{yc}(t)$ to be a genuine distribution function. Therefore, the following property must be satisfied:

$$\lim_{t\to+\infty} \widehat{F}_{yc}(t) = 1. \tag{21}$$

Following Rueda et al. [14], by taking a sufficiently large t_2, the condition (21) is satisfied. Therefore, it is logical to take $t_2 = \max_{k\in U} g_k$.

As in Section 3, Theorem 2 summarizes the main result.

Theorem 2. *Let $g_k = \widehat{\beta}'\mathbf{x}_k$, $k = 1, \ldots, N$ and F_g be the finite population distribution function of g-values. Assume that we wish to estimate F_y at the point t. Let*

$$G_1(\gamma) = \frac{-1}{N-1}\frac{\left[\sum_{k\in U}\Delta(\gamma - g_k)\Delta(t - y_k) - NF_y(t)F_g(\gamma)\right]^2}{NF_g(\gamma)(1 - F_g(\gamma))}.$$

Assume $A_t = \{a_1 \le \cdots \le a_P\}$ in Theorem 1. Let $B_t = \{b_k : k = 1, \ldots, P\}$ defined as

$$b_1 = \max_{l\in U_1}\{g_l\} \quad \text{with } U_1 = \{l \in U : g_l < a_1\}$$

$$b_k = \max_{l\in U_k}\{g_l\} \quad \text{with } U_k = \{l \in U : a_{k-1} \le g_l < a_k\};\ k = 2, \ldots, P.$$

Then, the point at which the calibration estimator is optimum is given by

$$t_{\text{opt}} = \operatorname*{argmin}_{g_k\in A_t\cup B_t} G_1(g_k).$$

As in Section 3, the proof is established in several steps.

Variance of the $\widehat{F}_{yc}(t)$ estimator

The calibration estimator $\widehat{F}_{yc}(t)$ with two points t_1 and $t_2 = \max_{k\in U} g_k$ is given by

$$\widehat{F}_{yc}(t) = \widehat{F}_{YH}(t) + \left(F_g(t_1) - \widehat{F}_{GH}(t_1)\right) B_{1s} + \left(1 - \frac{1}{N}\sum_{k\in s} d_k\right) B_{2s} \tag{22}$$

where

$$B_{1s} = \frac{\sum_{k\in s} d_k q_k \delta_{y_k}(t)\delta_{g_k}(t_1) \sum_{k\in s} d_k q_k - \sum_{k\in s} d_k q_k \delta_{y_k}(t) \sum_{k\in s} d_k q_k \delta_{g_k}(t_1)}{\sum_{k\in s} d_k q_k \delta_{g_k}(t_1) \left(\sum_{k\in s} d_k q_k - \sum_{k\in s} d_k q_k \delta_{g_k}(t_1)\right)}$$

$$B_{2s} = \frac{\sum_{k\in s} d_k q_k \delta_{y_k}(t) - \sum_{k\in s} d_k q_k \delta_{y_k}(t)\delta_{g_k}(t_1)}{\sum_{k\in s} d_k q_k - \sum_{k\in s} d_k q_k \delta_{g_k}(t_1)}$$

with $\delta_{y_k}(t) = \Delta(t - y_k)$ and $\delta_{g_k}(t_1) = \Delta(t_1 - g_k)$.

Under simple random sampling, we have

$$\frac{1}{N}\sum_{k\in s} d_k = 1.$$

Consequently, the estimator $\widehat{F}_{yc}(t)$ is

$$\widehat{F}_{yc}(t) = \widehat{F}_{YH}(t) + \left(F_g(t_1) - \widehat{F}_{GH}(t_1)\right) B_{1s}.$$

If we choose $q_k = 1$ for all $k \in U$, the asymptotic variance of $\widehat{F}_{yc}(t)$ takes the following expression:

$$V(\widehat{F}_{yc}(t)) = V(\widehat{F}_{YH}(t)) + B_1^2 V(\widehat{F}_{GH}(t_1)) - 2B_1 \operatorname{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_1)) \tag{23}$$

with

$$B_1 = \frac{\sum_{k\in U} \Delta(t - y_k)\Delta(t_1 - g_k) - NF_g(t_1)F_y(t)}{NF_g(t_1)(1 - F_g(t_1))}.$$

Then, we must find the value of t_1 that minimizes

$$B_1^2 V(\widehat{F}_{GH}(t_1)) - 2B_1 \operatorname{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_1)) = B_1 \left[B_1 V(\widehat{F}_{GH}(t_1)) - 2\operatorname{Cov}(\widehat{F}_{YH}(t), \widehat{F}_{GH}(t_1))\right]. \tag{24}$$

Under simple random sampling

$$V(\widehat{F}_{GH}(t_1)) = \frac{N}{N-1} F_g(t_1)\left(1 - F_g(t_1)\right) \tag{25}$$

$$\operatorname{Cov}(\widehat{F}_{GH}(t_1), \widehat{F}_{YH}(t)) = \frac{\sum_{k\in U} \Delta(t - y_k)\Delta(t_1 - g_k) - NF_y(t)F_g(t_1)}{N-1} \tag{26}$$

and we have

$$B_1 V(\widehat{F}_{GH}(t_1)) = \frac{1}{N-1}\left[\sum_{k\in U} \Delta(t_1 - g_k)\Delta(t - y_k) - NF_y(t)F_g(t_1)\right]$$

$$2\operatorname{Cov}(\widehat{F}_{GH}(t_1), \widehat{F}_{YH}(t)) = \frac{2}{N-1}\left[\sum_{k\in U} \Delta(t_1 - g_k)\Delta(t - y_k) - NF_y(t)F_g(t_1)\right]$$

and

$$B_1 V(\widehat{F}_{GH}(t_1)) - 2\operatorname{Cov}(\widehat{F}_{GH}(t_1), \widehat{F}_{YH}(t)) = \frac{-1}{N-1}\left[\sum_{k\in U} \Delta(t_1 - g_k)\Delta(t - y_k) - NF_y(t)F_g(t_1)\right]. \tag{27}$$

Now, if we replace (27) in (24), we obtain the following function

$$G_1(t_1) = B_1 \left[B_1 V(\widehat{F}_{GH}(t_1)) - 2\operatorname{Cov}(\widehat{F}_{GH}(t_1), \widehat{F}_{YH}(t))\right]$$

$$= \frac{-1}{N-1} \frac{\left[\sum_{k\in U} \Delta(t_1 - g_k)\Delta(t - y_k) - NF_y(t)F_g(t_1)\right]^2}{NF_g(t_1)(1 - F_g(t_1))}. \tag{28}$$

The properties of $G_1(\gamma)$ are given in Appendix.

Now, if we wish to estimate the distribution function F_y at point t, we must prove that the global minimum $G_1(t_1)$ is at a point of the set A_t or B_t, by studying the values of $G_1(t_1)$:

$$G_1(t_1) = \begin{cases} \dfrac{N(F_y(t))^2}{N-1}\dfrac{F_g(t_1)}{(F_g(t_1)-1)} & t_1 < a_1 \\ \dfrac{-1}{N(N-1)}\dfrac{\left[k_i - NF_y(t)F_g(t_1)\right]^2}{F_g(t_1)(1-F_g(t_1))} & a_i \le t_1 < a_{i+1} \\ \dfrac{N\left[F_y(t)\right]^2}{(N-1)}\left(1-\dfrac{1}{F_g(t_1)}\right) & a_P \le t_1. \end{cases}$$

In the interval $(-\infty, a_1)$, the function

$$G_1(x) = \frac{N(F_y(t))^2}{N-1}\frac{x}{(x-1)}$$

is monotone nondecreasing, and because $F_g(t_1)$ is monotone nondecreasing, the function $G_1(t_1)$ is monotone nondecreasing and its minimum point is at $t_1 = b_1$, and this minimum takes the value

$$\frac{N(F_y(t))^2}{N-1}\frac{F_g(b_1)}{(F_g(b_1)-1)}.$$

It is now apparent that $G_1(t_1)$ is monotone nondecreasing in $[a_P, +\infty)$ and its local minimum is at $t_1 = a_P$. This minimum is given by

$$\frac{N\left[F_y(t)\right]^2}{(N-1)}\left(1-\frac{1}{F_g(a_P)}\right).$$

Finally, if we wish to study the monotonicity of $G_1(t_1)$ in the interval $[a_i, a_{i+1})$ with $i = 1, 2, \ldots, P-1$, we must consider the function

$$f(x) = \frac{-1}{N(N-1)}\frac{\left[k_i - NF_y(t)x\right]^2}{x(1-x)}$$

where k_i is constant for all $[a_i, a_{i+1})$ given by

$$k_i = \sum_{k\in U} \Delta(a_i - g_{[k]})\Delta(t - y_{[k]}) < NF_y(t).$$

The derivative of $f(x)$ is equal to

$$f'(x) = \frac{1}{N(N-1)}\frac{NF_y(t)(2k_i - F_y(t))x^2 - 2k_i^2 x + k_i^2}{(x-x^2)^2}$$

and the equation $f'(x) = 0$ has two solutions

$$x_1 = \frac{k_i}{NF_y(t)}; \qquad x_2 = \frac{k_i}{2k_i - NF_y(t)}.$$

It is clear that the solution $x_1 \in (0, 1)$ because $k_i < NF_y(t)$ and the solution $x_2 \in (-\infty, 0)$ if $(2k_i - NF_y(t)) < 0$ or $x_2 \in (1, +\infty)$ if $(2k_i - NF_y(t)) > 0$.
Thus, if $(2k_i - NF_y(t)) < 0$ we have

$$f'(x) < 0 \quad \text{for } x \in (-\infty, x_2) \cup (x_1, 1) \cup (1, +\infty)$$
$$f'(x) > 0 \quad \text{for } x \in (x_2, 0) \cup (0, x_1).$$

If $(2k_i - NF_y(t)) > 0$

$$f'(x) > 0 \quad \text{for } x \in (-\infty, 0) \cup (0, x_1) \cup (x_2, +\infty)$$
$$f'(x) < 0 \quad \text{for } x \in (x_1, 1) \cup (1, x_2).$$

In both cases $f'(x) > 0$ for $x \in (0, x_1)$ and $f'(x) < 0$ for $x \in (x_1, 1)$ and consequently

$$f(x) \quad \text{is monotone nondecreasing for } x \in (0, x_1)$$
$$f(x) \quad \text{is monotone nondecreasing for } x \in (x_1, 1).$$

Because $F_g(t_1) \in (0, 1)$ and $F_g(t_1)$ is monotone nondecreasing, the local minimum of the function $G_1(t_1)$ in the interval $[a_i, a_{i+1})$ is at the point $t_1 = a_i$ or $t_1 = b_{i+1}$.

Therefore, the global minimum of the function $G_1(t_1)$ is at one point of $A_t = \{a_i : i = 1, 2, \ldots, P\}$ or $B_t = \{b_i : i = 1, 2, \ldots, P\}$.

The proposed calibration estimator

We need the entire population value of the study variable y in order to obtain the function $G_1(t_1)$. This problem can be solved by estimating $G_1(t_1)$ by

$$\begin{aligned}\widehat{G}_1(t_1) &= \frac{-1}{N-1}\frac{\left[\sum_{k\in s} d_k \Delta(t_1 - g_k)\Delta(t - y_k) - N\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_1)\right]^2}{N\widehat{F}_{GH}(t_1)(1-\widehat{F}_{GH}(t_1))}\\ &= \frac{-1}{N-1}\frac{\left[\sum_{k\in s} \frac{N}{n}\Delta(t_1 - g_k)\Delta(t - y_k) - N\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_1)\right]^2}{N\widehat{F}_{GH}(t_1)(1-\widehat{F}_{GH}(t_1))}\\ &= \frac{-N^2}{n^2(N-1)}\frac{\left[\sum_{k\in s} \Delta(t_1 - g_k)\Delta(t - y_k) - n\widehat{F}_{YH}(t)\widehat{F}_{GH}(t_1)\right]^2}{N\widehat{F}_{GH}(t_1)(1-\widehat{F}_{GH}(t_1))}.\end{aligned} \tag{29}$$

In a similar way, if we consider the set

$$A_{st} = \{g_k : k \in C_{st}\} \tag{30}$$

with $C_{st} = \{k \in s : y_k \le t\}$ and assume that A_{st} has p points, that is

$$A_{st} = \{a_i : i = 1, 2, \ldots, p\}$$

and we define the set

$$B_{st} = \{b_i : i = 1, 2, \ldots, p\} \tag{31}$$

by

- $b_1 = \max_{l\in s_1}\{g_l\}$ with $s_1 = \{l \in s : g_l < a_1\}$
- $b_k = \max_{l\in s_k}\{g_l\}$ with $s_k = \{l \in s : a_{k-1} \le g_l < a_k\}$; $k = 2, \ldots, p$

the global minimum of the function $\widehat{G}_1(t_1)$ is at one point of A_{st} or B_{st}.

Now we define the calibration estimator obtained with the values that minimize the function $\widehat{G}_1(t_1)$. This estimator is denoted by $\widehat{F}_{ymprop}(t)$. In the same way that it was done in the Section 3, one can prove that $\widehat{F}_{ymprop}(t)$ and the obtained calibration estimator using the optimal value of t_1 have the same asymptotic behaviour.

5. Simulation study

The optimum calibration points are determined by minimizing the asymptotic variance. In this section, a limited study has been conducted to investigate the design-based finite sample performance of the proposed estimators in comparison with that of conventional calibration estimators.

We compare the precision of the proposed calibration estimators $\widehat{F}_{ycprop}(t)$ and $\widehat{F}_{ycmprop}(t)$ with the following estimators: $\widehat{F}_{CD}(t)$ [2], $\widehat{F}_{RKM}(t)$ [1] the difference estimator $\widehat{F}_d(t)$, the usual calibration estimator $\widehat{F}_{yc}(t)$ from (6) with $t_0 = Q_g(0.5)$ and finally the calibration estimator $\widehat{F}_{ycm}(t)$, from (22) with two points $t_1 = Q_g(0.5)$ and $t_2 = \max_{k\in U}\{g_k\}$.

Several simulated and natural populations with different relationships between the study variable and the auxiliary variable are used in this study.

First, two natural populations are considered. The CARS population consists of the number of cars in the Spanish region of Andalusia in 2003 (variable of interest) and the number of cars in 2002 (the auxiliary variable). This population is available from the Andalusian Statistics Institute web site: http://www.juntadeandalucia.es/institutodeestadistica.

In this case, the study variable y and the auxiliary variable x present a good linear relation. The MURTHY population consists of 80 factories where the variable of interest is the output and the auxiliary variable is the number of workers. This has been studied in [18–20]. Examination of the scatter plots in Fig. 1 reveals that the linearity assumption is no longer valid in this population.

A finite population of size $N = 1000$ (BUMP, population) was also generated from a regression model $y_k = 2 + 2(x_k - 0.5) + \exp(-200(x_k - 0.5)^2) + \epsilon_k$ (see [21]) where x is uniformly distributed over [0, 1] and the ϵ_i's are i.i.d random variables from $N(0, 0.1)$. This model is used because we wish to observe the effect of model misspecification on the estimators and to provide an indication of the robustness of the estimators.

We selected 1000 samples for three different sample sizes under simple random sampling without replacement (SRSWOR). The considered sample sizes were 50, 75 and 100 for the CARS and BUMP populations but the population size of the MURTHY population was small ($N = 80$), and so samples of size 25, 30 and 35 were considered for this population.

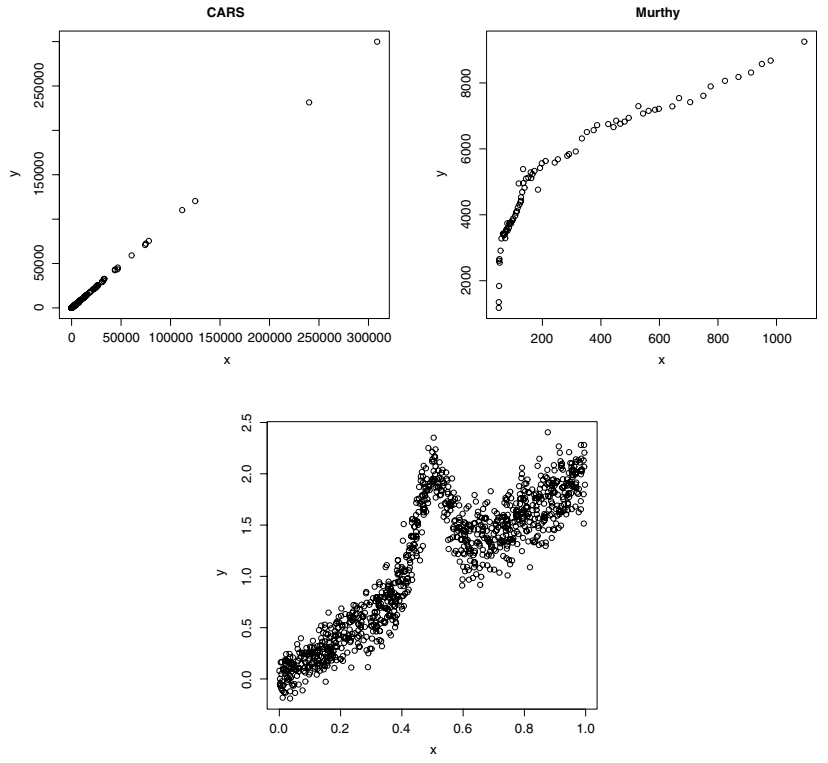

Fig. 1. Scatter plots for CARS, MURTHY and BUMP populations.

For each sample and for each estimator, estimates of the distribution function $F(t)$ were calculated for 11 different values of t, namely the quantiles $Q_y(\alpha)$ for $\alpha = 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8$ and 0.9.

The performance of all the estimators is measured by means of the average relative bias (AVRB) and the average relative efficiency (AVRE), given respectively by

$$\text{AVRB}(t) = \frac{1}{11}\sum_{q=1}^{11} |\text{RB}(t_q)|, \qquad \text{AVRE}(t) = \frac{1}{11}\sum_{q=1}^{11} \text{RE}(t_q)$$

where RB and RE are defined as

$$\text{RB}(t) = \frac{1}{B}\sum_{b=1}^{B} \frac{\widehat{F}(t)_b - F_y(t)}{F_y(t)} \quad \text{and} \quad \text{RE}(t) = \frac{MSE[\widehat{F}(t)]}{MSE[\widehat{F}_{YH}(t)]}, \tag{32}$$

where b indexes the bth simulation run, $\widehat{F}(t)$ is an estimator for the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical Mean Square Error for $\widehat{F}(t)$ and $MSE[\widehat{F}_{YH}(t)]$ is similarly defined for the Horvitz–Thompson estimator.

Note that the measures AVRB and AVRE have also been used in [22] and reveal the average behaviour of the various estimators at different values of t. Table 1 shows our results for all the populations.

Table 1 gives the values of the bias and the efficiency for all populations.

It can be seen in Table 1 that:

- In the CARS population, the $\widehat{F}_{CD}(t)$, $\widehat{F}_d$ and $\widehat{F}_{RKM}$ estimators have small biases, compared with the calibration estimators $\widehat{F}_{yc}$ and $\widehat{F}_{ycm}$. However, when the estimators based on the optimal points are considered, the bias is reduced considerably. Thus the $\widehat{F}_{ycprop}$ and $\widehat{F}_{ycmprop}$ estimators have the lowest AVRB of all the estimators considered.
- The $\widehat{F}_{CD}(t)$, $\widehat{F}_d$, $\widehat{F}_{RKM}$, $\widehat{F}_{yc}(t)$ and $\widehat{F}_{ycm}(t)$ estimators provide estimates with a large AVRB in the populations where the linear model does not fit. The estimated biases deviate significantly from zero for most quantiles, especially for the lower quantiles. This is not surprising given the extreme nonlinearity of these populations. In these populations, the $\widehat{F}_{ycprop}$ and $\widehat{F}_{ycmopt}$ estimators reduce the bias compared to the corresponding calibration estimators based on the median. This reduction is very significant in the MURTHY population.
- In all populations and all sizes the estimator $\widehat{F}_{ycmprop}$ produces the lowest AVRB.

Table 1
Average relative bias (AVRB) and the average relative efficiency (AVRE).

	MURTHY		CARS		BUMP	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 25$		$n = 50$		$n = 50$	
$\widehat{F}_{CD}$	0.169	0.4233	0.0503	0.2179	0.1374	0.3675
$\widehat{F}_{d}$	0.1549	0.401	0.0222	0.1734	0.1043	0.3345
$\widehat{F}_{RKM}$	0.1400	0.3755	0.0359	0.1944	0.1033	0.3299
$\widehat{F}_{yc}$	0.1597	0.4106	0.1285	0.3691	0.1407	0.3878
$\widehat{F}_{ycm}$	0.0875	0.4004	0.1231	0.3593	0.1344	0.3774
$\widehat{F}_{ycprop}$	0.0348	0.2305	0.022	0.1707	0.1337	0.3729
$\widehat{F}_{ycmprop}$	0.0291	0.2228	0.0184	0.1638	0.1003	0.3219
	$n = 30$		$n = 75$		$n = 75$	
$\widehat{F}_{CD}$	0.1494	0.3998	0.0506	0.2157	0.1336	0.3561
$\widehat{F}_{d}$	0.1318	0.3756	0.0187	0.1526	0.086	0.3014
$\widehat{F}_{RKM}$	0.1173	0.3481	0.0294	0.1721	0.0818	0.2938
$\widehat{F}_{yc}$	0.1339	0.3784	0.1012	0.3285	0.1123	0.3475
$\widehat{F}_{ycm}$	0.1293	0.3692	0.0964	0.3192	0.1066	0.3367
$\widehat{F}_{ycprop}$	0.0256	0.2009	0.018	0.1473	0.1062	0.3329
$\widehat{F}_{ycmprop}$	0.0206	0.1871	0.0157	0.1401	0.0785	0.286
	$n = 35$		$n = 100$		$n = 100$	
$\widehat{F}_{CD}$	0.1309	0.3744	0.0523	0.2148	0.1291	0.3478
$\widehat{F}_{d}$	0.115	0.3454	0.0166	0.1406	0.0739	0.2788
$\widehat{F}_{RKM}$	0.1036	0.3225	0.0262	0.1616	0.0699	0.271
$\widehat{F}_{yc}$	0.1183	0.3522	0.0859	0.3024	0.095	0.3196
$\widehat{F}_{ycm}$	0.1135	0.3437	0.0821	0.2944	0.0913	0.3103
$\widehat{F}_{ycprop}$	0.0194	0.1739	0.0158	0.1327	0.0894	0.3061
$\widehat{F}_{ycmprop}$	0.0162	0.1599	0.0139	0.1252	0.0666	0.2634

- In terms of average relative efficiency, we observe that $\widehat{F}_{ycmprop}$ performs better than $\widehat{F}_{ycprop}$, and that both estimators are more efficient than the $\widehat{F}_{yc}(t)$ and $\widehat{F}_{ycm}(t)$ estimators.
- In all cases, the estimator $\widehat{F}_{ycmprop}$ is the most efficient. This was expected in the case of the CARS population, where the linear model describes the data very well. However, in the case of the MURTHY population, this result is more impressive.

The main finding of this study is that both of the proposed calibration estimators obtained perform satisfactorily. The sizes of their biases decline as n increases and a gain in precision is obtained in comparison with alternative calibrated estimators. They are also robust against model misspecification.

In conclusion, we suggest that the research of optimum points for calibration provides a practical approach to estimating distribution functions, which offers useful gains in efficiency.

Acknowledgments

This research was partially supported by Ministerio de Educación y Ciencia (grant no. MTM2006-04809).

Appendix. Properties of $G_1(\gamma)$

The function $G_1(t_1)$ always takes a negative value for any value t_1 selected and we must obtain the point t_1 that minimizes $G_1(t_1)$. Let us again consider the population values of y in ascending order as in (16), the population values of the variable g, arranged by the variable y as in (17) and the set A_t given by (18), but in this case we consider another set of values

$$B_t = \{b_k : k = 1, 2, \ldots, P\} \tag{A.1}$$

where

$$b_1 = \max_{l \in U_1}\{g_l\} \quad \text{with } U_1 = \{l \in U : g_l < a_1\}$$

$$b_k = \max_{l \in U_k}\{g_l\} \quad \text{with } U_k = \{l \in U : a_{k-1} \le g_l < a_k\};\ k = 2, \ldots, P$$

and the values $a_1 < a_2 < \cdots < a_P$ are the P elements of A_t. The value b_1 does not exist when $a_1 = \min_{k \in U}\{g_k\}$, but in this case, it makes no sense for us to consider $t_1 < a_1$ in the calibration equation because $F_g(t_1) = 0$.

Next we observe that the function $G_1(t_1)$ is defined by:

(1) If $t_1 < a_1$ the function $G_1(t_1)$ is

$$G_1(t_1) = \frac{-N}{N-1}\frac{-\left(F_y(t)F_g(t_1)\right)^2}{F_g(t_1)(1-F_g(t_1))} = \frac{N(F_y(t))^2}{N-1}\frac{F_g(t_1)}{(F_g(t_1)-1)}.$$

(2) If $t_1 \geq a_P$

$$\begin{aligned} G_1(t_1) &= \frac{-1}{N(N-1)}\frac{\left[NF_y(t) - NF_y(t)F_g(t_1)\right]^2}{F_g(t_1)(1-F_g(t_1))} \\ &= \frac{N\left[F_y(t)\right]^2}{(N-1)}\frac{(F_g(t_1)-1)}{F_g(t_1)} = \frac{N\left[F_y(t)\right]^2}{(N-1)}\left(1-\frac{1}{F_g(t_1)}\right). \end{aligned}$$

(3) If $a_i \leq t_1 < a_{i+1}$ with $i = 1, 2, \ldots, P-1$

$$G_1(t_1) = \frac{-1}{N(N-1)}\frac{\left[k_i - NF_y(t)F_g(t_1)\right]^2}{F_g(t_1)(1-F_g(t_1))}$$

where

$$k_i = \sum_{k\in U}\Delta(t_1 - g_{[k]})\Delta(t - y_{[k]}) = \sum_{k\in U}\Delta(a_i - g_{[k]})\Delta(t - y_{[k]}).$$

Consequently, $G_1(t_1)$ is a piecewise function.

References

[1] J.N.K. Rao, J.G. Kovar, H.J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, Biometrika 77 (1990) 365–375.
[2] R.L. Chambers, R. Dunstan, Estimating distribution functions from survey data, Biometrika 73 (1986) 597–604.
[3] A.H. Dorfman, P. Hall, Estimator of the finite population distribution function using nonparametric regression, The Annals of Statistics 16 (3) (1993) 1452–1475.
[4] S. Wang, A.H. Dorfman, A new estimator for the finite population distribution function, Biometrika 83 (1996) 639–652.
[5] J. Chen, C. Wu, Estimation of distribution function and quantiles using the model-calibrated pesudo empirical likelihood method, Statistica Sinica 12 (2002) 1223–1239.
[6] M. Rueda, J.F. Muñoz, New model-assisted estimators for the distribution function using the pseudo empirical likelihood method, Statistica Neerlandica 63 (2) (2009) 227–244.
[7] V.M. Estevao, C.E. Särndal, Survey estimates by calibration on complex auxiliary information, International Statistical Review 42 (2006) 127–147.
[8] G.E. Montanari, M.G. Ranalli, Nonparametric model calibration estimation in survey sampling, Journal of the American Statistical Association 100 (2005) 1429–1442.
[9] Y. Tillé, Unbiased estimation by calibration on distribution in simple sampling designs without replacement, Survey Methodology 28 (2002) 77–85.
[10] S. Singh, Generalized calibration approach for estimating variance in survey sampling, Annals of the Institute of Statistical Mathematics 53 (2) (2001) 404–417.
[11] S. Singh, S. Horn, F. Yu, Estimation of variance of general regression estimaro: Higher level calibration approach, Survey Methodology 24 (1998) 41–50.
[12] J.C. Deville, C.E. Särndal, Calibration estimators in survey sampling, Journal of the American Statistical Association 87 (1992) 376–382.
[13] T. Harms, P. Duchesne, On calibration estimation for quantiles, Survey Methodology 32 (2006) 37–52.
[14] M. Rueda, S. Martínez, H. Martínez, A. Arcos, Estimation of the distribution function with calibration methods, Journal of Statistical Planning and Inference 137 (2) (2007) 435–448.
[15] C.E. Särndal, The calibration approach in survey theory and practice, Survey Methodology 33 (2) (2007) 99–119.
[16] M. Rueda, S. Martínez, I. Sánchez-Borrego, Model-calibration estimation of the distribution function using nonparametric regression, Metrika (2009), in press (doi:10.1007/s00184-008-0199-y).
[17] R.H. Randles, On the asymptotic normality of statistics with estimated parameters, The Annals of Statistics 10 (1982) 462–474.
[18] M.N. Murthy, Sampling Theory and Method, Statistical Publishing Society, Calcutta, 1967.
[19] A. Kuk, T.K. Mak, Median estimation in the presence of auxiliary information, Journal of the Royal Statistical Society. Series B 51 (2) (1989) 261–269.
[20] A. Kuk, T.K. Mak, A functional approach to estimating finite population distribution functions, Communications in Statistics—Theory and Methods 23 (3) (1994) 883–896.
[21] F.J. Breidt, J.D. Opsomer, Local polynomial regression estimators in survey sampling, The Annals of Statistics 28 (4) (2000) 1026–1053.
[22] P.L.D. Silva, C.J. Skinner, Estimation distribution functions with auxiliary information using poststratification, Journal of Official Statistics 11 (3) (1995) 277–294.