ARTICLE IN PRESS

Journal of Computational and Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Distribution function estimates from dual frame context

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

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

[b] Department of Mathematics, University of Almería, Spain

ARTICLE INFO

Article history:
Received 6 June 2016
Received in revised form 20 September 2016

MSC:
62D05

Keywords:
Auxiliary information
Calibration technique
Distribution function estimates
Dual-frame surveys

ABSTRACT

The estimation of a finite population distribution function under a dual frame context is considered when auxiliary population information is available. Several procedures are defined and compared to various methods adapted from the literature. The asymptotic distribution of the proposed estimators is established, a brief simulation is implemented and an application to real data is included.

1. Introduction

The focus of this paper is on the estimation of the finite population distribution function, on the basis of a sample taken from a dual-frame context. The subject is important: the distribution function is a basic statistic underlying many others [1]; for purposes of assessing and comparing finite populations, it can be more revealing than means and totals [2].

Expressed in words, for a given value of t, the finite distribution function $F(t)$ is the proportion of y in the population of size N not exceeding t (usually defined using the indicator function $\Delta(u)$ defined by $\Delta(u) = 1$, if u is true, $\Delta(u) = 0$, otherwise). The following basic properties estimating the finite distribution function $F(t)$ are: it is monotonic nondecreasing in t, it is a step function with step size $1/N$ and its values are confined to [0, 1]. Estimating the finite population distribution function $F(t)$ is in some respects easier and in others more difficult than estimating a population total or mean. On the one hand, for fixed t, $F(t)$ is simply a mean of 0's and 1's. On the other hand, we typically want to estimate $F(t)$ for more than one value of t and these estimates need to be coordinated. When the main variable y is related to an auxiliary variable x, the issue arises of how to use this information, since we are now concerned with $\Delta(y \leq t)$ and not y itself, which is what is usually modelled on x. In most cases, the estimators of $F(t)$ are constructed with reference to a specific linear regression model with "slope" β which is estimated by weighted least squares using design weights. With the fitted values from the regression available, model calibrated estimators can be defined.

Multiple frame surveys were first introduced by [3] as a device for minimizing costs without reducing the accuracy of results with respect to the single frame surveys. Since then, the multiple frame sampling theory has experienced a noticeable development and several estimators for the total of a continuous variable have been proposed. First estimators were formulated in a dual frame context, i.e. for the case where two frames are available for sampling. [4,5] proposed dual frame estimators based on new techniques. [6,7] applied likelihood methods to compute estimators that perform well in

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

http://dx.doi.org/10.1016/j.cam.2016.09.027
0377-0427/

complex designs. More recently, [8] proposed a new class of estimators in dual frame survey sampling that makes use of a power transformation and [9,10] used calibration techniques to propose estimators in the dual frame context.

In a dual-frame context, there are several ways to obtain design weights and several situations in which the auxiliary information can be related from frames. In the present case, the main variable y is related to an auxiliary variable x in each frame, and this auxiliary variable may be different from the frames or it may be the same. To date, no comprehensive comparison has been made of the many available alternatives for estimating $F(t)$.

In this paper we adapt distribution function estimates in a dual-frame context in which there is no auxiliary information, with auxiliary information solely about the frame sizes and with more auxiliary information. In the latter case, a post-stratification estimator with partial auxiliary information (as in [11]) and a model calibration estimator with complete auxiliary information (following [12,13]), are defined. The asymptotic distribution of the proposed estimators is established. A brief simulation study is included in Section 3. The paper ends with an application to real data.

2. Distribution function estimates in dual frame context

Assume we have a finite set of N population units identified by integers, $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, and let A and B be two sampling frames, both may be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c$, $b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$, where ^{c} denotes the complementary of a set. Let N, N_A, N_B, N_a, N_b and N_{ab} be the number of population units in $\mathcal{U}$, $\mathcal{A}$, $\mathcal{B}$, a, b and ab, respectively.

Let y be a variable of interest in the population and let y_k be its value on unit k, for $k = 1, \ldots, N$. Our aim is then to estimate the finite population distribution function

$$F_y(t) = \frac{1}{N}\sum_{k\in U} \Delta(t - y_k) \quad \text{where } \Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k \end{cases} \tag{1}$$

which can be written as the population mean of Δ_t values. Let D_{t_k} be the Δ_t value on unit k, for $k = 1, \ldots, N$. Then, our aim is to estimate, for each t, the population mean, $\overline{D}_t = D_t/N$, where D_t denotes the population total of variable D_t. When the population size N is unknown, N can be viewed as the total of constant $\mathbf{1}_N$ or $N = D_\infty$. This reduces our goal that of estimating the population totals D_t.

This population total D_t can be written as $D_t = D_{t_a} + D_{t_{ab}} + D_{t_b}$, where $D_{t_a} = \sum_{k\in a} D_{t_k}$, $D_{t_{ab}} = \sum_{k\in ab} D_{t_k}$ and $D_{t_b} = \sum_{k\in b} D_{t_k}$. To this end, independent samples s_A and s_B are drawn from frame A and frame B of sizes n_A and n_B, respectively. Unit k in $\mathcal{A}$ has first-order inclusion probability $\pi_k^A = Pr(k \in s_A)$ and unit k in $\mathcal{B}$ has first-order inclusion probability $\pi_k^B = Pr(k \in s_B)$.

From data collected in s_A, it is possible to compute one unbiased estimator of the total for each domain in frame A, $\hat{D}_{t_a}$ and $\hat{D}^A_{t_{ab}}$, as described below:

$$\hat{D}_{t_a} = \sum_{k\in s_A} \delta_k(a) d_k^A D_{t_k}, \qquad \hat{D}^A_{t_{ab}} = \sum_{k\in s_A} \delta_k(ab) d_k^A D_{t_k},$$

where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise and d_k^A are the weights under the sampling design used in frame A, defined as the inverse of the first-order inclusion probabilities, $d_k^A = 1/\pi_k^A$. Similarly, using information included in s_B, we can obtain an unbiased estimator of the total for domain b and another one for domain ab, $\hat{D}_{t_b}$ and $\hat{D}^B_{t_{ab}}$, which can be expressed as

$$\hat{D}_{t_b} = \sum_{k\in s_B} \delta_k(b) d_k^B D_{t_k}, \qquad \hat{D}^B_{t_{ab}} = \sum_{k\in s_B} \delta_k(ab) d_k^B D_{t_k},$$

with $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise, and where d_k^B the weights under the sampling design used in frame B defined as the inverse of the first-order inclusion probabilities, $d_k^B = 1/\pi_k^B$.

Various approaches for estimating the population total from dual-frame surveys have been proposed. Below, we adapt some of them to the context of distribution function estimation.

2.1. No auxiliary information

Dual-frame approach

Under the dual-frame approach [3,14], a convex combination of the two overlap estimates is utilized to obtain an unbiased global estimator of the total population.

Using this idea, we consider the use of a fixed quantity η to weight $\hat{D}^A_{t_{ab}}$ and $\hat{D}^B_{t_{ab}}$, providing the estimator

$$\hat{D}_{t_\eta} = \hat{D}_{t_a} + (\eta)\hat{D}^A_{t_{ab}} + (1 - \eta)\hat{D}^B_{t_{ab}} + \hat{D}_{t_b}. \tag{2}$$

Please cite this article in press as: A. Arcos, et al., Distribution function estimates from dual frame context, Journal of Computational and Applied Mathematics (2016), http://dx.doi.org/10.1016/j.cam.2016.09.027