Available online at www.sciencedirect.com

ScienceDirect

Mathematics and Computers in Simulation 102 (2014) 51–61

MATHEMATICS AND COMPUTERS IN SIMULATION

www.elsevier.com/locate/matcom

Original article

New ratio and difference estimators of the finite population distribution function

J.F. Muñoz [a], A. Arcos [b], E. Álvarez [a], M. Rueda [b,*]

[a] *University of Granada, Department of Quantitative Methods in Economics and Business, Spain*
[b] *Department of Statistics and Operational Research, Spain*

Received 21 October 2011; received in revised form 17 December 2012; accepted 5 April 2013
Available online 28 August 2013

Abstract

New design-based ratio and difference estimators of the distribution function are defined by minimizing the mean square error of a class of estimators. Proposed estimators do not assume a superpopulation model between the variable of interest and the auxiliary variable. Results derived from simulation studies indicate that proposed estimators can be more accurate than existing estimators, especially when alternative estimators suffer from model misspecifications.

Keywords: Auxiliary information; Distribution function; Ratio-type estimators

1. Introduction

Estimation of the distribution function is an important objective in sample surveys that has received much attention in the last years, especially since [3] proposed the first method for estimating the distribution function using auxiliary information at the estimation stage. Thenceforth there has been much work on the use of auxiliary information, as such information can increase the precision of estimators of the distribution function.

Given a finite population U containing N units, the purpose of this paper is to estimate the distribution function $F(t)$ for a variable of interest y, which is defined as the proportion of units in U for which the value of y is less than or equal to t, i.e. $F(t)=N^{-1}\sum_{i\in U}\delta(y_i\le t)$, where $\delta(a)$ takes the value 1 if its argument is true and 0 otherwise.

To estimate $F(t)$, we assume that a sample s, of size n, is selected from U according to a sampling design, where π_i is the inclusion probability for the ith unit and π_{ij} is the joint inclusion probability for the units i and j. The customary design-based estimator of $F(t)$ is $\widehat{F}(t)=\widehat{N}^{-1}\sum_{i\in s}\pi_i^{-1}\delta(y_i\le t)$, where $\widehat{N}=\sum_{i\in s}\pi_i^{-1}$. We observe that $\widehat{F}(t)$ makes no use of the auxiliary information. However, it is common to assume that there exists an auxiliary variable x associated to y, which can be used at the estimation stage to improve the estimation of $F(t)$. Various estimators that incorporate auxiliary information at the estimation stage are now described.

* Corresponding author at: Department of Statistics and Operational Research, University of Granada, Granada, Spain. Tel.: +34 629199659.
E-mail addresses: jfmunoz@ugr.es (J.F. Muñoz), arcos@ugr.es (A. Arcos), encarniav@ugr.es (E. Álvarez), mrueda@ugr.es (M. Rueda).

0378-4754/$36.00
http://dx.doi.org/10.1016/j.matcom.2013.04.027

Ref. [14] used standard results for totals or means to define the design-based ratio and difference estimators

$$\widehat{F}_r(t) = \frac{1}{N}\frac{\sum_{i\in s}\pi_i^{-1}\delta(y_i \le t)}{\sum_{i\in s}\pi_i^{-1}\delta(\widehat{R}x_i \le t)}\sum_{i\in U}\delta(\widehat{R}x_i \le t) \tag{1}$$

and

$$\widehat{F}_d(t) = \frac{1}{N}\left\{\sum_{i\in s}\pi_i^{-1}\delta(y_i \le t) + \sum_{i\in U}\delta(\widehat{R}x_i \le t) - \sum_{i\in s}\pi_i^{-1}\delta(\widehat{R}x_i \le t)\right\}, \tag{2}$$

where $\widehat{R} = (\sum_{i\in s}\pi_i^{-1}y_i)(\sum_{i\in s}\pi_i^{-1}x_i)^{-1}$ is the customary design-consistent estimator of the population ratio $R = Y/X$, where Y and X denote respectively the population totals of the interest and auxiliary variables. In other words, they defined ratio and difference estimators for $F(t)$ by treating $\delta(y_i \le t)$ and $\delta(\widehat{R}x_i \le t)$ as "y- and x-variables". As noted by [14], the design-based estimators, $\widehat{F}_r(t)$ and $\widehat{F}_d(t)$, can lead to considerable gains in efficiency when y is approximately proportional to x.

Ref. [3] proposed a model-based estimator, $\widehat{F}_m(t)$, which is based on the superpopulation model

$$y_i = \beta x_i + \upsilon(x_i)u_i \quad (i = 1, \ldots, N), \tag{3}$$

where β is an unknown parameter, $\upsilon(x) = x^{1/2}$ and the u_i's are independent and identically distributed random variables with zero mean. Assuming the model (3), [14] also defined the estimator $\widehat{F}_{dm}(t)$, which is asymptotically both design-unbiased and model-unbiased under the model (3), where

$$\widehat{F}_{dm}(t) = \frac{1}{N}\left\{\sum_{i\in s}\pi_i^{-1}\delta(y_i \le t) + \left(\sum_{i\in U}\widehat{G}_i - \sum_{i\in s}\pi_i^{-1}\widehat{G}_{ic}\right)\right\} \tag{4}$$

where

$$\widehat{G}_i = \left(\sum_{j\in s}\frac{1}{\pi_j}\right)^{-1}\left[\sum_{j\in s}\pi_j^{-1}\delta\{\widehat{u}_j \le x_i^{-1/2}(t - \widehat{R}x_i)\}\right],$$

$$\widehat{G}_{ic} = \left(\sum_{j\in s}\frac{\pi_i}{\pi_{ij}}\right)^{-1}\left[\sum_{j\in s}(\pi_i/\pi_{ij})\delta\{\widehat{u}_j \le x_i^{-1/2}(t - \widehat{R}x_i)\}\right],$$

$\widehat{u}_j = x_j^{-1/2}(y_j - \widehat{R}x_j)$. From [14] it is clear that $\widehat{F}_{dm}(t)$ is more efficient than estimators $\widehat{F}_r(t)$ and $\widehat{F}_d(t)$. As noted by [9], the efficiencies of the model-based estimators, $\widehat{F}_m(t)$ and $\widehat{F}_{dm}(t)$, depend on an implicit linearity assumption between y and x. This issue implies that such estimators require the initial specification of a superpopulation model and they can have a poor performance under misspecification (see also, [16]). Ref. [14] also studied the advantages of the design-based estimators over the model-based estimator proposed by [3] under model misspecifications.

Non-parametric estimators of the distribution function [11,9] do not require the initial specification of a superpopulation model. However, such estimators require the specification of a bandwidth parameter to control the amount of smoothing, with Kuk's estimator also requiring appropriate scaling of the response variable. As noted by [18], the performance of the Kuk's estimator depends on the choice of the transformation, and poor results can be achieved if the transformation is not appropriate. A recent review of existing estimators of the distribution function in the literature can be seen in [6]).

The idea of this paper is to define, in Section 2, new design-based estimators for $F(t)$ without assuming a superpopulation model between y and x. Proposed ratio and difference estimators are defined by substituting $\widehat{R}$ in (1) and (2) by an appropriate value λ that minimizes the mean square error of a proposed class of estimators. This implies that proposed estimators should be more efficient than (1) and (2), and proposed estimators should be more efficient than model-based estimators under model misspecifications. The problem of finding an analytical expression for the optimum value of λ is not a simple issue. For this reason, in Section 3 we describe a grid search method to find the optimum value of λ. In Section 4, results derived from various simulation studies support our findings.