Available online at www.sciencedirect.com
ScienceDirect
Mathematics and Computers in Simulation 81 (2011) 2317–2325
MATHEMATICS AND COMPUTERS IN SIMULATION
www.elsevier.com/locate/matcom

Original article

Indirect estimation of proportions in natural resource surveys

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

[a] Department of Statistics and O.R., University of Granada, C.P. 18071, Granada, Spain
[b] Department of Quantitative Methods in Economics and Business, University of Granada, C.P. 18071, Granada, Spain

Received 6 November 2009; received in revised form 22 November 2010; accepted 20 December 2010
Available online 1 January 2011

Abstract

Regression type estimators of a population proportion under a general sampling design and using auxiliary information are obtained. Confidence intervals based on various methods, involving auxiliary information, are also derived. An application of the proposed methods is illustrated by estimating the proportion of lakes at risk of acidification, based on data from the U.S. Environmental Protection Agency. Theoretical properties suggest that the proposed methods can outperform alternative methods, and the results derived from a Monte Carlo simulation study support this view.

Keywords: Auxiliary information; Qualitative variable; Regression estimator

1. Introduction

Many natural resource surveys contain auxiliary information at the population level in addition to sample data. This auxiliary information can come from satellite images, aerial photographs, GPS data, etc. A data set is then created for the statistical agency which reflects the knowledge of both the design and the auxiliary information.

In the presence of auxiliary information, there exist many design-based approaches (see [7–9]) to improve the precision of estimators in comparison with customary methods, which do not involve auxiliary information. However, techniques involving auxiliary information have been discussed for quantitative variables, and the extension to the estimation of proportions requires further investigation. The use of auxiliary information at the estimation stage could improve the precision of estimators in the context of qualitative variables.

In this paper, we propose to use interval and point regression type estimators for estimating the population proportion. The proposed estimators incorporate auxiliary information into the estimation stage and are based on a general sampling design, which implies that the proposed methods can be applied in many situations in practice. Theoretical properties are also derived. Using data published by the U.S. Environmental Protection Agency, we show that the proposed methods are straightforwardly applicable. Results derived from simulation studies indicate that the proposed regression type estimator can be more efficient than the traditional estimator, and that the proposed confidence intervals can outperform the alternative methods, especially in terms of average length.

* Corresponding author. Tel.: +34 958240494; Fax: +34 958243267.
E-mail addresses: mrueda@ugr.es (M. Rueda), jfmunoz@ugr.es (J.F. Muñoz), arcos@ugr.es (A. Arcos), encarniav@ugr.es (E. Álvarez).

0378-4754/$36.00
doi:10.1016/j.matcom.2010.12.028

2. Estimation of the proportion using auxiliary information

We consider the scenario of a finite population $U = \{1, \ldots, N\}$ containing N units. Let $A_1, \ldots, A_N$ denote the values of an attribute of interest A, where $A_i = 1$ if the i th unit possesses the attribute A and $A_i = 0$ otherwise. Let B denotes an auxiliary attribute associated with A and let the values be given by $B_1, \ldots, B_N$. We also assume that a sample s, of size n, is selected from U according to the well known simple random sampling without replacement (SRSWOR). The extension to a general sampling design is addressed in Section 5.

The aim is to estimate the population proportion of individuals that possess the attribute A, i.e., $P_A = N^{-1}\sum_{i=1}^{N} A_i$. Assuming a finite population, the naive estimator of P_A, which makes no use of the auxiliary information, is given by $\hat{p}_A = n^{-1}\sum_{i \in s} A_i$.

We propose to use regression type estimators to estimate P_A, as they make use of the auxiliary information at the estimation stage, and more efficient estimates are likely to be achieved. A regression type estimator of P_A is

$$\hat{p}_{reg} = \hat{p}_A + b(P_B - \hat{p}_B), \tag{1}$$

where b is a known constant and $\hat{p}_B$ is the naive estimator of P_B. We assume that the population proportion of individuals that possess the attribute B, P_B, is known from a census or is estimated without error. This is the usual requirement in the problem of the estimation of a mean. Estimator (1) belongs to a general class of estimators that can provide other known estimators. For example, the difference type estimator is obtained when $b = 1$.

3. Theoretical properties and the optimum regression estimator

In this section, we derive some theoretical properties of the proposed regression estimator, which are based upon the following framework. We assume a design-based approach, where $E(\cdot)$ and $V(\cdot)$ denote the expectation and the variance with respect to the sample design. The asymptotic properties of estimators are derived by assuming that the finite population embeds in a sequence of populations $\{U_\nu\}$, where n_ν and N_ν increase such that $n_\nu/N_\nu \to f$ when n_ν, $N_\nu \to \infty$.

Let A^c and B^c denote the complementary attributes of A and B, and consider the population two-way table given by

$$\begin{array}{c|cc|c} & B & B^c & \\ \hline A & N_{11} & N_{12} & N_{1\cdot} \\ A^c & N_{21} & N_{22} & N_{2\cdot} \\ \hline & N_{\cdot 1} & N_{\cdot 2} & N \end{array} \tag{2}$$

where $N_{1\cdot} = \sum_{i=1}^{N} A_i$ is the number of units in the population that possess the attribute A, $N_{2\cdot}$ is the number of units in the population that do not possess the attribute A, etc. Analogously, N_{11} is the number of units in the population that simultaneously possess the attributes A and B, etc. Cramer's V coefficient based on (2) is given by $\phi = (N_{11}N_{22} - N_{12}N_{21})(N_{1\cdot}N_{2\cdot}N_{\cdot 1}N_{\cdot 2})^{-1/2}$.

Classification (2) can also be defined at the sample level as

$$\begin{array}{c|cc|c} & B & B^c & \\ \hline A & n_{11} & n_{12} & n_{1\cdot} \\ A^c & n_{21} & n_{22} & n_{2\cdot} \\ \hline & n_{\cdot 1} & n_{\cdot 2} & n \end{array} \tag{3}$$

Proposition 1. *The proposed regression estimator $\hat{p}_{reg}$ is unbiased.*