Combining Probability and Non-Probability Samples in Finite Population Inference

G. Cicchitelli, G.E. Montanari
Dipartimento di Scienze Statistiche
Via A. Pascoli, Perugia - Italy
pino@stat.unipg.it

1. Introduction

As is known, finite population inference can be design-based or model-based. In the first approach probability distribution comes from the random mechanism by means of which the sample is drawn; in the second approach, uncertainty arises from the model we assume to describe the structure of the population.

Model-based inference is generally more efficient, particularly in small samples, but in large samples uncontrolled biases due to misspecified models might yield an invalid inference. On the other hand, probability samples allow us to make inferences which are objective, that is independent of subjective decisions about the choice of sample units, and on subjective assumptions on the structure of the population.

In what follows we will discuss a two-phase sampling strategy where, in the first phase, a non-probability sample is selected from the target population aimed at the modelling of the relationship between the study variable and a set of auxiliary variables. The estimated model is then used, at the second phase, for designing the probability sample.

2. The strategy

Consider a finite population $U = \{1, 2, \text{K}\ , N\}$ and the problem of estimating the mean of the study variable y. To this end assume that the value x_i of the scalar auxiliary variable x is known for each $i \in U$ (the procedure can be obviously extended to the case where x is a vector). Suppose that the study variable y is related to the auxiliary variable x through the model

$$y = m(x) + \varepsilon . \tag{1}$$

where ε is a random variable with mean zero and variance $v(x)$, $m(x)$ is a smooth function of x, and $v(x)$ is smooth and strictly positive (this model has been studied in a different context by

Breidt and Opsomer, 2000). Let s_1 be the first-phase sample $(y_1, y_2, \ldots, y_{n_1})$ drawn purposively from U, where y_i is assumed to be a realisation of a random variable with mean $m(x_i)$ and variance $v(x_i)$ according to model (1). Let $\hat{y}_i = \hat{m}(x_i), \quad i = 1, 2, .., N,$ where $\hat{y} = \hat{m}(x)$ denotes a convenient estimator of $m(x)$ based on s_1.

Now, the variable $\hat{y}$ is a new auxiliary variable, which can be employed for designing the second-phase sample s_2 of size n_2. For example, $\hat{y}$ could be used for stratification purposes and for allocating the sample among the strata by means of Neyman optimum criterion.

The rationale for the above procedure is represented by its robustness property. In fact, the inference procedure relies on the probability sample s_2 which yields unbiased or nearly unbiased and consistent estimates for large n_2. So, model misspecification does not affect the validity of the inference.

In this paper we compare the efficiency of the above procedure with that of classical one-phase sampling strategies based on the auxiliary variable x. Conditions under which the two-phase strategy is more efficient are highlighted, subject to cost constraints.

The technique can be advantageously used also in environmental studies. Here, the model is expressed in terms of expected value and covariance of regionalised variables $y(\mathbf{x})$, where $\mathbf{x}$ is the location at which the observation is made. In this case, $\hat{y}_i \quad (i = 1, 2, \ldots, N)$ are the population values computed using the kriging predictor based on the variogram estimated by means of s_1.

REFERENCE

Breidt F.J. and Opsomer J.D. (2000), Local polynomial regression estimators in survey sampling, *The Annals of Statistics*, 28, 1026-1053.

RESUME

Dans ce travail, on propose une stratégie d'échantillonnage en deux phases. Dans la première phase, on tire de la population finie étudiée un échantillon de taille n_1 pour estimer la relation de dépendence entre la variable d'intérêt et une ou plusieures variables auxiliaires connues pour toutes les unités de la population. Dans la seconde phase, on tire un échantillon de taille n_2 de la même population par un plan d'échantillonnage qui se fonde sur les valeurs prédites par le modèle estimé dans la première phase. On étudie le problème d'efficacité des estimateurs comparant cette stratégie avec les stratégies usuelles.