Statistics and Probability Letters 104 (2015) 163–168

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Cross-classified sampling: Some estimation theory

C.J. Skinner

Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK

ARTICLE INFO

Article history:
Received 2 May 2015
Received in revised form 1 June 2015
Accepted 1 June 2015
Available online 6 June 2015

Keywords:
Plane sampling
Survey sampling
Two dimensional sampling

ABSTRACT

For a population represented as a two-way array, we consider sampling via the product of independent row and column samples. Theory is presented for the estimation of a population total under alternative methods of sampling the rows and columns.

1. Introduction

In some surveys the population can be represented by the elements of a two-way array, $(i, j),\ i = 1, \ldots, N,\ j = 1, \ldots, M$, and it is natural to take a sample as a Cartesian product $S = \{(i, j) : i \in S^R,\ j \in S^C\}$, where S^R and S^C are samples selected from the rows $\{1, \ldots, N\}$ and columns $\{1, \ldots, M\}$, respectively. A procedure in which S^R and S^C are selected independently by probability sampling schemes is called *cross-classified sampling*, following Ohlsson (1996).

A typical application of cross-classified sampling is to a survey of businesses which each handle a large number of products. Data is then collected from a sample of businesses on a sample of products.

We take the inferential objective to be to estimate the total of a variable y across units in the finite population, given only data on the values of y for units in the sample. In the business application, y might denote the value of the sale or purchase of a product in some time period and there may be interest in the total sales in the population. In some applications, there might be subunits, such as transactions, in which case y may be the sum across such subunits. This framework can also allow for cases where a product is not handled by a business in the time period by setting $y = 0$. Dalén and Ohlsson (1995) present an application of the use of cross-classified sampling in the construction of the Swedish Consumer Price Index.

Although cross-classified sampling can be treated within the general framework of finite population sampling, so that, for example, a Horvitz–Thompson estimator can be defined (Ohlsson, 1996), many specific aspects of this method are not covered by standard theory. For example, the latter typically treats sample quantities in different strata as independent, whereas we shall have to allow for dependence induced between strata in one dimension by sampling in the other dimension. Apart from the two key papers cited so far, the literature on the theory of cross-classified sampling is very limited. Vos (1964) provides some results for simple random sampling. There is a rather more extensive literature on the special case when the row and columns are ordered, typically in space, but possibly in time. This is usually called two-dimensional sampling or plane sampling. See e.g. Quenouille (1949), Bellhouse (1977), Iachan (1985) and Stevens and Olsen (2004). We shall, however, not assume an ordering of rows or columns and shall not refer further to this literature.

In this paper, we extend the theory in Ohlsson (1996) in a number of ways. First, we provide more explicit results on stratified sampling both from design-based and model-based perspectives. Second, we present results for with replacement

E-mail address: c.j.skinner@lse.ac.uk.

http://dx.doi.org/10.1016/j.spl.2015.06.001
0167-7152/

unequal probability sampling. These results may be of interest in their own right, since it is common in business surveys for either businesses or the volume of product sales to vary considerably by size and for probability proportional to size sampling to be employed. However, in addition, we shall make use of the theory for with replacement sampling to construct bootstrap procedures for variance estimation. Such procedures may prove simpler to implement in practice than the more direct procedures we describe first.

2. Estimation for simple random and stratified sampling

We consider the estimation of the finite population total

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij},$$

where y_{ij} denotes the value of y for population unit (i, j) and is observed only for units $(i, j) \in S$. We consider two particular sampling schemes.

2.1. Simple random sampling

We first consider the prototypical case where S^R and S^C are selected by simple random sampling without replacement (SRSWOR). The natural unbiased estimator of Y here is the Horvitz–Thompson estimator given by

$$\hat{Y}_{srs} = \frac{NM}{nm} \sum_{i \in S^R} \sum_{j \in S^C} y_{ij},$$

where n and m are the sizes of S^R and S^C respectively. The (design-based) variance of this estimator is now presented together with an unbiased estimator of this variance. The first of these results is given in Ohlsson (1996).

Theorem 2.1. *Under the SRSWOR design above, the estimator $\hat{Y}_{srs}$ is unbiased for Y, with variance*

$$var(\hat{Y}_{srs}) = N^2 M^2 \left\{ (1 - f_R) \frac{\sigma_R^2}{n} + (1 - f_C) \frac{\sigma_C^2}{m} + (1 - f_C)(1 - f_R) \frac{\sigma_{RC}^2}{nm} \right\}, \tag{1}$$

where

$$\sigma_R^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{Y}_{i.} - \bar{Y}_{..})^2, \qquad \sigma_C^2 = \frac{1}{M-1} \sum_{j=1}^{M} (\bar{Y}_{.j} - \bar{Y}_{..})^2,$$

$$\sigma_{RC}^2 = \frac{1}{N-1} \frac{1}{M-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2,$$

$f_R = n/N, f_C = m/M, \bar{Y}_{i.} = \sum_{j=1}^{M} y_{ij}/M, \bar{Y}_{.j} = \sum_{i=1}^{N} y_{ij}/N$ *and* $\bar{Y}_{..} = \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}/NM$.

An unbiased estimator of $var(\hat{Y}_{srs})$ *is obtained by replacing* σ_R^2, σ_C^2 *and* σ_{RC}^2 *in (1) by*

$$\hat{\sigma}_R^2 = \frac{1}{n-1} \sum_{i \in S^R} (\bar{y}_{i.} - \bar{y}_{..})^2 - (1 - f_C) \frac{\hat{\sigma}_{RC}^2}{m},$$

$$\hat{\sigma}_C^2 = \frac{1}{m-1} \sum_{j \in S^C} (\bar{y}_{.j} - \bar{y}_{..})^2 - (1 - f_R) \frac{\hat{\sigma}_{RC}^2}{n},$$

$$\hat{\sigma}_{RC}^2 = \frac{1}{(n-1)(m-1)} \sum_{i \in S^R} \sum_{j \in S^C} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2,$$

where $\bar{y}_{i.} = \sum_{j \in S^C} y_{ij}/m, \bar{y}_{.j} = \sum_{i \in S^R} y_{ij}/n$ *and* $\bar{y}_{..} = \sum_{i \in S^R} \sum_{j \in S^C} y_{ij}/nm$.

Proof. The variance expression in (1) is given by Ohlsson (1996), with the basis of its proof indicated. The unbiasedness of $\hat{\sigma}_R^2$, $\hat{\sigma}_C^2$ and $\hat{\sigma}_{RC}^2$ may be shown by taking successive expectations with respect to the two sampling schemes and using standard results in sampling theory (Cochran, 1977, Theorem 2.4). □

2.2. Stratified random sampling

As noted in the introduction, variance expressions are less straightforward to obtain under stratification in cross-classified sampling than in standard theory. We now suppose that the rows and columns are stratified into G and H strata respectively and relabel the elements of the population as quadruples (g, h, i, j), where $g = 1, \ldots, G, h = 1, \ldots, H, i = 1, \ldots, N_g$ and $j = 1, \ldots, M_h$. The values N_g and M_h denote the stratum sizes, with $\sum N_g = N$ and $\sum M_h = M$. We suppose