Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT

Journal of Statistical Planning and Inference 128 (2005) 569–591

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Variance approximation under balanced sampling

Jean-Claude Deville[a], Yves Tillé[b,*]

[a] *Laboratoire de Statistique d'Enquête, CREST-ENSAI, École Nationale de la Statistique et de l'Analyse de l'Information rue Blaise Pascal, Campus de Ker Lann, Bruz 35170, France*

[b] *Chaire de statistique appliquée, Université de Neuchâtel, Espace de l'Europe 4, Case Postale 827, Neuchâtel 2002, Switzerland*

Received 17 December 2002; accepted 26 November 2003

Abstract

A balanced sampling design has the interesting property that Horvitz–Thompson estimators of totals for a set of balancing variables are equal to the totals we want to estimate, therefore the variance of Horvitz–Thompson estimators of variables of interest are reduced in function of their correlations with the balancing variables. Since it is hard to derive an analytic expression for the joint inclusion probabilities, we derive a general approximation of variance based on a residual technique. This approximation is useful even in the particular case of unequal probability sampling with fixed sample size. Finally, a set of numerical studies with an original methodology allows to validate this approximation.

MSC: 62D05

Keywords: Sampling design; Balanced sampling; Unequal probability sampling; Variance approximation; Variance estimation; Sampling theory; Sample surveys; Sample; Sampling unequal probability

1. Introduction

A sample is said to be balanced on a set of balancing variables if the estimated totals of these variables are equal to the population totals. Several partial solutions had been presented in Deville et al. (1988), Deville (1992), Ardilly (1991), Hedayat and Majumdar (1995), and Valliant et al. (2000). The cube method proposed by Deville and Tillé (2002) provides a general non-enumerative solution to select balanced samples, and has been used at the *Institut National de la Statistique et des Études*

* Corresponding author. Tel.: 0041327181475; fax: 0041327181381.
E-mail addresses: deville@ensai.fr (J.-C. Deville), yves.tille@unine.ch (Y. Tillé).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.11.011

Économiques (INSEE, French Bureau of Statistics) for its most important statistical projects: the renovated continuous census and the master sample (on this topic see Dumais and Isnard, 2000; Wilms, 2000). The use of balanced sampling in several projects has dramatically improved efficiency, allowing the reduction of the variance from 20% to 90% comparatively to simple random sampling in some cases (see for instance Deville et al., 1988).

An important question is the derivation of the variance of total estimators under balanced sampling. In balanced sampling it is hard to derive an analytic expression for balanced sampling. Firstly, we propose a class of approximations of the variance derived from theoretical arguments. These approximations are based on a residual technique and are related to previous works: Yates (1949), Hájek (1964, Chapters 4, 7 and 14), and Brewer (2002, Chapter 9).

These approximations show that the variance depends only on the residuals for a particular regression of the interest variable against the balancing variables. The variance is then equal to zero under a linear deterministic model. Secondly, we check the approximation by means of a set of numerical studies. We examine several cases of balanced sampling, and we look for the worst interest variable, i.e. the variable for which the approximation gives the largest overevaluation or underevaluation. We show that in most cases, the error due to the approximation does not exceed 10% of the value of the variance.

The notation is defined in Section 2. A class of variance approximations is given in Section 3. An original and new methodology to evaluate the accuracy of balanced design is developed in Section 4. In Section 5, we present the results of the numerical study. The question of estimation of the variance is discussed in Section 6. Mathematical technicalities are given in the appendices.

2. Balanced sampling

Consider a finite population U of size N whose units can be identified by a label $k \in \{1,\dots,N\}$. The aim is to study the interest variable y that takes the values y_k, $k \in U$, on the units of the population. More precisely, we want to estimate the total $Y = \sum_{k\in U} y_k$. Suppose also that p balancing variables $x_1,\dots,x_p$ are available, i.e. that the vectors of values $\mathbf{x}_k = (x_{k1} \ \dots \ x_{kj} \ \dots \ x_{kp})'$ taken by the p balancing variables are known for all the units of the population. Moreover, without loss of generality, the p vectors $(x_{1j} \dots x_{kj} \dots x_{Nj})'$, $j = 1,\dots,p$, are assumed linearly independent.

A sampling design $p(.)$ is a probability distribution on the set Ω of all the subsets of U such that $\sum_{s\in\Omega} p(s) = 1$. The random sample S takes a value s with probability $\Pr(S = s) = p(s)$. The inclusion probability of unit k is the probability that unit k is in the sample $\pi_k = \Pr(k \in S)$ and the joint inclusion probability is the probability that two distinct units are together in the sample $\pi_{k\ell} = \Pr(k \text{ and } \ell \in S)$. The Horvitz–Thompson estimator given by $\hat{Y} = \sum_{k\in S} y_k/\pi_k$ is an unbiased estimator of Y. The Horvitz–Thompson estimator of the jth auxiliary balancing total $X_j = \sum_{k\in U} x_{kj}$ is $\hat{X}_{j\pi} = \sum_{k\in S} x_{kj}/\pi_k$. With a vectorial notation, the Horvitz–Thompson estimator vector $\hat{\mathbf{X}} = \sum_{k\in S} \mathbf{x}_k/\pi_k$, estimates without bias $\mathbf{X} = \sum_{k\in U} \mathbf{x}_k$.