Computational Statistics and Data Analysis 69 (2014) 122–132

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Inclusion probabilities in partially rank ordered set sampling

Omer Ozturk[a,*], Mohammad Jafari Jozani[b]

[a] Department of Statistics, The Ohio State University, Columbus, OH, 43210, USA

[b] Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada

ARTICLE INFO

Article history:
Received 14 June 2012
Received in revised form 22 July 2013
Accepted 22 July 2013
Available online 3 August 2013

Keywords:
Finite population
Horvitz–Thompson estimator
Partial ranking
Coefficient of variation
Ranked set sampling

ABSTRACT

In a finite population setting, this paper considers a partially rank ordered set (PROS) sampling design. The PROS design selects a simple random sample (SRS) of M units without replacement from a finite population and creates a partially rank ordered judgment subsets by dividing the units in SRS into subsets of a pre-specified size. The subsetting process creates a partial ordering among units in which each unit in subset h is considered to be smaller than every unit in subset h' for $h' > h$. The PROS design then selects a unit for full measurement from one of these subsets. Remaining units are returned to the population based on three replacement policies. For each replacement policy, we compute the first and second order inclusion probabilities and use them to construct the Horvitz–Thompson estimator and its variance for the estimation of the population total and mean. It is shown that the replacement policy that does not return any of the M units, prior to selection of the next unit for full measurement, outperforms all other replacement policies.

1. Introduction

In many survey sampling studies, in addition to variable of interest, researchers often have additional auxiliary information to improve statistical inference. If this information is well-structured so that it can be turned into numerical covariates, it is usually incorporated into the model that explains data generating mechanism. Regression and ratio estimators are two such examples that use this kind of information. In many instances, this information may not be accurate, cannot be turned into a numerical covariate or may be even subjective. Nevertheless, it contains valuable information which could be used at the design stage of the survey. Use of subjective information has generated extensive research interests on ranked set sampling (RSS) which was originally developed to keep the overall cost of data collection minimal in estimating mean pasture yield in agricultural fields (McIntyre, 1952, 2005). In this case, use of simple random sampling (SRS) is time consuming and expensive since it requires close clipping, weighing and moving the yield of many quadrants. The SRS procedure ignores the substantial auxiliary information available to researcher before close clipping, such as the apparent vegetation density in each quadrant. RSS uses this information to create an artificially stratified sample by ranking the sampling units. Even though this ranking information is subjective and varies from ranker to ranker, the implementation of the ranked set sampling procedure does not require an established standard of measurement for the sampling units. It only requires knowing the relative positions of the units in a set. Auxiliary information that does not require an established standard of measurement, such as previous census outcomes, is very common in survey sampling studies and can be used for ranking purpose in RSS.

To construct a ranked set sample of size n from an infinite population, we first determine a set size M and then select nM units at random from the population. These units are randomly divided into n sets, each of size M. Units in each set are

[*] Corresponding author. Tel.: +1 614 2922866; fax: +1 614 2922096.
E-mail address: omer@stat.osu.edu (O. Ozturk).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.034

ranked from smallest to largest without actual measurement of the variable of interest. Ranking may be done based on all available auxiliary information and it does not have to be perfectly accurate. The i-th smallest ranked unit is selected for a full measurement in n_i sets in such a way that $\sum_{i=1}^{M} n_i = n$. The fully measured units $X_{[j]i}$, $i = 1, \ldots, n_j, j = 1, \ldots, M$, are called a ranked set sample. If $n_j \equiv n/M$, the ranked set sample is called balanced. The square brackets are used to indicate that the ranking process may not be perfect. If the ranking process is perfect, an RSS usually yields higher efficiency for many statistical procedures. On the other hand, the error in the ranking process not only reduces the efficiency, but it may also produce an invalid inference. For example, substantial amount of the ranking error, in certain settings, may produce tests with inflated type I error rates, confidence intervals with deflated coverage probabilities and estimators with large biases. In order to reduce the impact of the ranking error, Ozturk (2011) introduced a partially rank ordered set (PROS) sampling design. The PROS design does not require a full ranking of units in each set. Instead, the ranker assigns units in each set into subsets of pre-specified sizes. The units within each subset are not ranked, but each unit in subset h is considered to have smaller rank than the rank of each unit in subset h' for all $h' > h$. Ozturk (2012) and Gao and Ozturk (2012) used this design to develop nonparametric inference for one and two sample problems, respectively. Recently, Ozturk (2013) and Frey (2012) relaxed the assumption that number of subsets needs to be pre-specified. This provides a flexibility in that the ranker is allowed to declare as many subsets as desired depending on his/her ranking ability. They showed that this flexibility further improves the efficiency of PROS design.

In recent years, many researchers considered some variation of RSS designs in finite population setting, e.g., Patil et al. (1995), Deshpande et al. (2006), Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2008), Gokpinar and Ozdemir (2010), Jafari Jozani and Johnson (2011, 2012) and Nourmohammadi et al. (submitted for publication). In finite population setting, the construction of a ranked set sample can be done in different ways depending on the replacement policy for measured and ranked units in a set. Deshpande et al. (2006) considered three different designs, Level 0, Level 1 and Level 2 designs. The Level 0 sampling design requires that units in a given set are selected without replacement, but all units in the set including the measured unit are replaced back into the population prior to selection of the next set. The Level 1 design has the same replacement policy as the Level 0 design except that the unit selected for full measurement is not returned into the population. The Level 2 design requires that none of the units in a set, regardless of whether they were measured or not, is replaced back into the population. These designs have similar properties for large population sizes, but they have different behaviors when the population size is small. Level 2 design induces a stronger negative correlation among the sample membership indicators of population units than the other two designs and is usually more efficient. The efficiency improvement is the largest when the ranking process is perfect and diminishes as the ranking information becomes poor. The quality of ranking is a decreasing function of the set size in the sense that larger set sizes lead to poor ranking quality and reduce efficiency. In practice, it is recommended that the set size should not be more than five or six to control the ranking error.

One of the advantages of using RSS (or PROS) sampling design is its feature to reduce the cost of the data collection process in settings where the cost of the measurement of a unit is higher than the ranking of a small set. A data set for such a setting is provided in Ozturk et al. (2005) at the research farm of Ataturk University, Erzurum, Turkey. The research focus at this farm is on improving meat quality and production as well as other traits of local sheep population. This requires periodically sampling the population to monitor the biological characteristic of the population. The measurement process is labor intensive and costly due to active nature of animals. It usually requires two to three people to hold the animal firm during a measurement process. The farm has available auxiliary variables in its data base, such as mothers weight at mating and birth weight. These auxiliary variables are highly correlated with the variable of interest, weight at the seventh month. The PROS design provides a natural setting for this type of research, where birth weight can be used as an auxiliary variable to create judgment classes. Since the auxiliary variable is available with no additional cost and highly correlated with the variable of interest, set size can be increased and required sample size can be reduced to minimize the total cost of the sampling process.

In this paper, we use PROS design for finite populations to increase the set size without causing too much ranking error, and hence improving the information content of the sample and reducing the total cost. The PROS design performs fewer ranking than the RSS of a comparable size since it only partitions sample units to subsets and units within subsets are not ranked. Information loss due to fewer ranking is compensated by the increase in the set size. The PROS design is still subjected to the misplacement error where a unit whose rank is in subset s_i could be misplaced to some other subsets due to improper ranking. On the other hand, this misplacement error is much smaller than a ranking error in a comparable RSS design.

In finite population setting, it is important to know the probability that each particular unit appears in the sample. This probability is called the first order inclusion probability and defined to be $\pi_i = P(\text{unit } i \text{ appears in the sample})$. This probability in RSS setting depends on the replacement level. Inclusion probabilities for Level 0 RSS design have been considered by Jafari Jozani and Johnson (2011) and for Level 1 RSS design by Al-Saleh and Samawi (2007), and Ozdemir and Gokpinar (2008). Recently, Frey (2011) provided recursive algorithms to calculate inclusion probabilities for all Levels 0–2 RSS designs that can be used for relatively large sample sizes.

In this paper, in Section 2, we introduce PROS sampling design and consider three different ways of constructing a PROS design depending on the replacement policies, Levels 0–2. Section 3 provides recursive algorithms to compute the first and second order inclusion probabilities which are necessary to construct design based estimators of population parameters and their variances. Section 4 uses the results in Section 3 to construct the Horvitz–Thompson estimator for population total