JORIND (9) 1, June, 2011 ISSN 1596 – 8308. www.transcampus.org./journals, www.ajol.info/journals/jorind

COMPARISON OF FOUR ESTIMATORS UNDER SAMPLING WITHOUT REPLACEMENT

L. A. Nafiu and M. D. Shehu
Department of Mathematics and Statistics, and
A. M. Saliua
Department of Computer Science, Federal University of Technology, Minna
E-mail:lanconserv@yahoo.com

Abstract
In some applications, it is cost efficient to sample data in more than one stage. In the first stage, a simple random sample is drawn and then stratified accordingly to some easily measured attributes. This project described four estimators for the treatment of samples drawn without replacement when equal and unequal probabilities are considered. It also compared their resulting standard error under sampling without replacement using data on diabetic patients in Niger state for the years 2000, 2001, 2002 and 2003. The results were obtained using a program written in Microsoft Visual C++ programming language. It was observed that the two-stage sampling under unequal probabilities without replacement is always better than the other three estimators considered.

Keywords: Unequal probability sampling, two-stage sampling, hansen-hurwitz estimator and horvitz-thompson estimator.

Introduction

In designing a study, it can be advantageous to sample units in more than one- stage. The criteria for selecting a unit at a given stage typically depend on attributes observed in the previous stages. Some types of units may be more informative than others, and it is better to sample them at a higher rate. If it costs little to determine the attributes that are necessary to classify the units, it can be cost efficient to stratify a large sample in stage one and then in stage two to subsample the strata at different rates (Chambers and Dunstan, 1986).

In sampling extensive populations, primary units that vary in size are encountered frequently. Moreover, considerations of cost often dictate the use of multistage sampling. Suppose that each unit in the population can be divided into a number of smaller units, or subunits. A sample of n units has been selected. If subunits within a selected unit give similar results, it seems uneconomical to measure them all (Cochran, 1977).

A common practice according to Durbin (1953) and Thompson (1992) is to select and measure a sample of the subunits in any chosen unit. This technique is called subsampling, since the unit is not measured completely but is itself sampled. Another name, due to Mahalanobis, is two-stage sampling, because the sample is taken in two steps. The first is to select a sample of units, often called the primary units, and the second is to select a sample of second stage units or subunits from each chosen primary unit. Subsampling has a great variety of applications, which go far beyond the immediate scope of sample surveys. Whenever any process involves chemical, physical or biological tests that can be performed on a small amount of material, it is likely to be drawn as a subsample from a

larger amount that is itself a sample (Sarndal and Wright, 1992).

Rao *et al* (1990) and Rao (1998) opined that some concentrated effort is required in order to obtain a good working knowledge of multistage sampling when the units vary in size, because the technique is flexible. The units may be either with equal probabilities or with probabilities proportional to size or to some estimate of size. Various rules can be devised to determine the sampling and subsampling fractions, and various methods of estimation are available. The advantages of the different methods depend on the nature of the population, on the field costs, and on the supplementary data that are at our disposal.

A common study design in health services research is the multistage cluster sample (Chambers *et al*, 1992) to abstract data from medical records. The typical example would be the sampling of physicians associated with those practices, followed by the sampling of patients treated by the sampled physicians. The sampling plan for the study must address which practices, physicians and patients to select. The ultimate goal is a sample of patients, but the human subject's approvals process and the costs of access to medical records compel following the hierarchy of the service delivery system to the patient.

John and Steven (2003) considered the simple and common two-stage problem of sampling patients within physician practices to estimate a population average or total. The reference design would be to sample practices proportional to size and then take equal numbers of patients per sampled practice up to the budget constraint. This design minimizes the design effect (the design effect specifies the loss or the gain in precision of a design relative to simple random sampling) by simultaneously minimizing cluster effects and the variability in the sampling weights.

Study objective

The main objective of this research work is to compare four different estimators when data arise from one-stage or two-stage design with equal or unequal probabilities (probabilities proportional to size, PPS) under sampling without replacement. It considers the study design decisions when sampling patients within health plans/health centers.

Type and source of data

The data used in this research work is of secondary type and was collected from Niger State Ministry of Health, Minna. We constructed a sample frame from all Local Government Areas in Niger State with diabetic patients.

Estimation of population total

In a one-stage design such as cluster sampling, the variability of the estimator occurs because different samples of primary units will give different values of the estimate. With two-stage designs, the estimator has variability even for a given selection of primary units, because different subsamples of secondary units will give rise to different values of the estimator. Let N denote the number of primary units in the population and n the number of primary units in the sample. Let M_i be the number of secondary units in the ith primary unit. The total number of secondary units in the population is $M = \sum_{i=1}^{N} M_i$. Let y_{ij} denote the value of the variable of interest of the jth secondary unit in the ith primary unit. The total of the y-values in the ith primary unit is denoted as y_i, that is,

$y_i = \sum_{j=1}^{M_i} y_{ij}$. The population total is

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} \qquad 1$$

Suppose that the population was stratified into L independent strata, indexed by s = 1,2, …, L and that the members of the sth stratum was grouped into N_{1s} primary sampling unit's (PSU's), indexed by i =1,2,…, N_{1s} and the ith PSU included N_{2si} members, indexed by j = 1,2,…, N_{2si}. The corresponding symbols for sample are denoted by the lower case n with the same subscripts as shown in table 1 below.

Table 1: Symbols for two-stage clustered sample data.

	Population	Sample
One-stage Unit Two-stage Unit	N_1 N_{2i}	n_1 n_{2i}
One-stage index Two-stage index	$i = 1,2,\ldots,N_1$ $j = 1,2,\ldots,N_{2i}$	$i = 1,2,\ldots,n_1$ $j = 1,2,\ldots,n_{2i}$
Total Cell Count Cell Proportion	$N = \sum_{i=1}^{N_1} N_{2i}$ $Y = \sum_{i=1}^{N_1}\sum_{j=1}^{N_{2i}} y_{ij}$ $\frac{Y}{N}$	$n = \sum_{i=1}^{n_1} n_{2i}$ $\hat{Y} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_{2i}} y_{ij}$ $\frac{\hat{Y}}{n}$

Also, table 2 below shows the variances for the four estimators under sampling without replacement.

Table 2: Variances for the Estimated Population Total.

Variances for the Estimators		
Equal Probability Unequal Probability	$V(\hat{Y}_{1ewor})$ $V(\hat{Y}_{1uwor})$	$V(\hat{Y}_{2ewor})$ $V(\hat{Y}_{2uwor})$

Methodology

Primary units selected using equal probabilities without replacement

When primary units are selected by simple random sampling without replacement, an unbiased estimator of the population total is

$$\hat{Y}_{1ewor} = \frac{N}{n}\sum_{i=1}^{n} y_i \qquad 2$$

The variance of this estimator is

$$V(\hat{Y}_{1ewor}) = N(N-n)\frac{\sigma_u^2}{n} \qquad 3$$

where σ_u^2 is the finite population variance of the primary unit totals,

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_1)^2 \quad \text{and} \quad \mu_1 = \frac{1}{N}\sum_{i=1}^{N} y_i$$

An unbiased estimate of the variance is

$$\hat{V}(\hat{Y}_{1ewor}) = N(N-n)\frac{s_u^2}{n} \qquad 4$$

where s_u^2 is the sample variance of the primary unit totals,

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Primary units selected using unequal probabilities without replacement

Suppose that two units are to be drawn from a stratum. The first unit is drawn with probabilities p_i, proportional to some measure of size. Let the ith unit be selected. If we follow the most natural method, at the second draw one of remaining units is selected with assigned probabilities $p_j/(1-p_i)$. Hence the total probability Π_i that the ith unit will be selected at either the first or the second draw is

$$\Pi_i = p_i\left(1 + A - \frac{p_i}{1-p_i}\right)$$

where $A = \sum_{j=1}^{N}\frac{p_j}{(1-p_j)}$

Also, the probability Π_{ij} that units i and j are in the sample is

$$\Pi_{ij} = p_i p_j\left(\frac{1}{1-p_i} + \frac{1}{1-p_j}\right)$$

$$\sum_{j\neq i}^{N}\Pi_{ij} = (n-1)\Pi_i \quad \text{and} \quad \sum_{i=1}^{N}\sum_{j>i}^{N}\Pi_{ij} = \frac{1}{2}n(n-1).$$

These give

$$\sum_{j\neq i}^{N}(\Pi_{ij} - \Pi_i\Pi_j) = (n-1)\Pi_i - \Pi_i(n-\Pi_i) = -\Pi_i(1-\Pi_i)$$

Thus, the variance of $\hat{Y}_{1uwor}$ is given as:

An unbiased estimator of the population total under sampling without replacement with probabilities proportional to size, based on the Horvitz and Thompson (1952), is

$$\hat{Y}_{1uwor} = \sum_{i=1}^{n}\frac{y_i}{\Pi_i} \qquad 5$$

Let $P(s)$ denote the probability of a sample consisting of n specified units. Then $\Pi_{ij} = \sum P(s)$ over all samples containing the ith and jth units, and $\Pi_i = \sum P(s)$ over all samples containing the ith unit. When we take $\sum\Pi_{ij}$ for $j \neq i$, every $P(s)$ for a sample containing the ith unit is counted $(n-1)$ times in the sum, since there are $(n-1)$ other values of j in the sample. Consequently, $\sum_{i=1}^{N}\Pi_i =$ n,

$$V(\hat{Y}_{1uwor}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\Pi_i\Pi_j - \Pi_{ij})(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j})^2 \quad 6$$

An unbiased sample estimator of $V(\hat{Y}_{1uwor})$ is given as:

$$\hat{V}(\hat{Y}_{1uwor}) = \sum_{i=1}^{n}\sum_{j>i}^{n}(\Pi_i\Pi_j - \Pi_{ij})(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j})^2 \quad 7$$

Provided that none of the Π_i in population vanishes.

Variances for various designs in two-stage sampling

Units selected with equal probabilities without replacement

The m_i sample subunits in the ith unit are chosen by simple random sampling. The unbiased estimator of the population total is:

$$\hat{Y}_{2ewor} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{M_i}{m_i}y_{ij} \quad 8$$

The variance of $\hat{Y}_{2ewor}$ as given by Okafor (2002) is:

$$V(\hat{Y}_{2ewor}) = \frac{N^2(1-f_1)}{n}S_1^2 + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})}{m_i}S_{wi}^2 \quad 9$$

where $S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$, $S_{wi}^2 = \frac{1}{M_i-1}\sum_{i=1}^{M_i}(Y_{ij} - \overline{Y_i})^2$, $f_1 = \frac{n}{N}$ and $f_{2i} = \frac{m_i}{M_i}$

Suppose that we have an unbiased estimate $\hat{S}_{wi}^2$ of the second-stage variance S_{wi}^2 of $\overline{Y_i}$ and an unbiased sample estimate $\hat{S}_1^2$ of S_1^2 from one-stage sampling. Then, an unbiased sample estimator of $V(\hat{Y}_{2ewor})$ is:

$$\hat{V}(\hat{Y}_{2ewor}) = \frac{N^2(1-f_1)}{n}\hat{S}_1^2 + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})}{m_i}\hat{S}_{wi}^2 \quad 10$$

where $\hat{S}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$, $\hat{S}_{wi}^2 = \frac{1}{m_i-1}\sum_{i=1}^{m_i}(y_{ij} - \overline{y_i})^2$, $f_1 = \frac{n}{N}$ and $f_{2i} = \frac{m_i}{M_i}$

Units selected with unequal probabilities without replacement

Let n first sampling units be selected from the total population N by probability proportional to size. Let Π_i be the probability of selecting the ith first sampling unit. In each selected first sampling, m_i second sampling units are selected from M_i by simple random sampling without replacement.

The estimator of the population total is

$$\hat{Y}_{2uwor} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\Pi_i} \qquad 11$$

where $\Pi_i = p_i(1 + A - \frac{p_i}{1-p_i})$

where $A = \sum_{j=1}^{N} \frac{p_j}{(1-p_j)}$

Also, the probability Π_{ij} that units i and j are in the sample is

$$\Pi_{ij} = p_i p_j (\frac{1}{1-p_i} + \frac{1}{1-p_j})$$

The variance of $\hat{Y}_{2uwor}$ as given by Thompson (1992) is:

$$V(\hat{Y}_{2uwor}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\Pi_i\Pi_j - \Pi_{ij})(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j})^2 + \sum_{i=1}^{N}\frac{M_i^2(1-f_{2i})S_{wi}^2}{m_i\Pi_i} \qquad 12$$

where $S_{wi}^2 = \frac{1}{M_i - 1}\sum_{i=1}^{M_i}(Y_{ij} - \overline{Y}_i)^2$ and $f_{2i} = \frac{m_i}{M_i}$

Its unbiased sample estimator is

$$\hat{V}(\hat{Y}_{2uwor}) = \sum_{i=1}^{n}\sum_{j>i}^{n}(\Pi_i\Pi_j - \Pi_{ij})(\frac{y_i}{\Pi_i} - \frac{y_j}{\Pi_j})^2 + \sum_{i=1}^{n}\frac{M_i^2(1-f_{2i})\hat{S}_{wi}^2}{m_i\Pi_i} \qquad 13$$

where $\hat{S}_{wi}^2 = \frac{1}{m_i - 1}\sum_{i=1}^{m_i}(Y_{ij} - \overline{Y}_i)^2$ and $f_{2i} = \frac{m_i}{M_i}$

Provided that none of the Π_i in population vanishes.

Estimated standard errors

The results in tables 3 – 6 below for standard errors were obtained with the help of computer program written in Microsoft Visual C++ programming language (Hubbard, 2000).

Table 3: Standard Errors for the estimated population totals using data in year 2000.

Standard errors for the estimates

Equal Probability	21561	14425
Unequal Probability	19401	13567

Table 4: Standard errors for the estimated population totals using data in year 2001.

	Standard Errors for the estimates	
Equal Probability	25057	24106
Unequal Probability	24553	23538

Table 5: Standard errors for the estimated population totals using data in year 2002.

	Standard Errors for the estimates	
Equal Probability	34189	31963
Unequal Probability	32371	31456

Table 6: Standard errors for the estimated population totals using data in year 2003.

	Standard Errors for the estimates	
Equal Probability	30438	29699
Unequal Probability	30303	29642

Findings

The quantity under estimate is the total number of diabetics' patients by local government areas in Niger state for years 2000, 2001, 2002 and 2003. The standard error of the estimate was investigated using an empirical method with different estimators under one-stage sampling and two-stage sampling. Both tables 3 - 6 show that the standard error in the case of sampling under unequal probability is less than the

standard error of sampling under equal probability without replacement.

Discussion

The data presented in tables 3 - 6 indicate that substantial reductions in the standard error can be obtained through the use of unequal probabilities without forfeiting an unbiased estimate of the sampling standard error. Two-stage sampling also gives better estimation compared with one-stage sampling.

Conclusion and recommendations

It was observed that two-stage sampling under unequal probabilities without replacement is always better. When an unbiased estimator of high precision and an unbiased sample estimate of its standard error are required, the sampling system employing unequal probabilities at each stage of sampling is particularly appropriate. The ultimate aim of this study as stated earlier is to compare different estimators when data arise from one-stage or two-stage with equal or unequal probabilities (probabilities proportional to size, PPS) under sampling without replacement. Importantly, two-stage sampling under unequal probabilities (probabilities proportional to size, PPS) without replacement was found to be the most efficient.

References

Chambers, R., Dorfman, A. and Hall, P. (1992). "Properties of Estimators of the Finite Population Distribution Function". *Biometrika. 79: 577-582.*

Chambers, R. and Dunstan, P.S. (1986). "Model-Based Finite Population Correction for the Horvitz-Thompson Estimator. *Biometrika. 74: 797-799.*

Cochran, W.G. (1977). *Sampling Techniques*. Third edition. New York: John Wiley and Sons.

Durbin, J. (1953). "Some Results in Sampling Theory When the Units are Selected with Unequal Probabilities". *Journal of the Royal Statistical Society. 15: 262-274.*

Hansen, M. and Hurwitz, W. (1943). "On the Theory of Sampling From Finite Populations". *Annals of Mathematical Statistics. 14: 333-362.*

Horvitz, D.G. and Thompson D. J. (1952). "A Generalization of Sampling Without Replacement From a Finite Universe". *Journal of American Statistical Association. 47: 663-685.*

Hubbard, J.R. (2000). *Programming with C++*. Second Edition. Schaum's Outlines, New Delhi: Tata McGraw-Hill Publishing Company Limited.

John, L.A and Steven, L.W. (2003). "Sampling Patients within Physician Practices and Health plans". *Journal for Improving Health Care Delivery and Policy.38: 1625-1640.*

Rao, J., Kovar, J. and Mantel, H. (1990). "On Estimating Distribution Functions and Quantiles From Survey Data Using Auxiliary Information". *Biometrika. 77: 365-375.*

Rao, J. (1998). "On the Efficiency of Sampling with Various Probabilities and the Selection of Units with Replacement". *Biometrika. 83: 34-42.*

Sarndal, C.E. and Wright, R.L. (1992). "Design-Based and Model-Based Inference in Survey Sampling". *Scandinavian Journal of Statistics. 5: 27-52.*

Thompson, S.K. (1992). *Sampling*. New York: John Wily and Sons.