GLOBAL JOURNAL OF MATHEMATICAL SCIENCES, VOL. 1, NO.1 & 2, 2002: 67-74

ESTIMATION FOR DOMAINS IN DOUBLE SAMPLING WITH PROBABILITIES PROPORTIONAL TO KNOWN SIZE

GODWIN A. UDOFIA

(Received 3 July 2002; Revision accepted 5 August 2002)

ABSTRACT

Available publications show that the variance of an estimator of a domain parameter depends on the variance of the study variable for the domain elements and on the variance of the mean of that variable for elements of the domain in each constituent stratum. In this article, we show that the variance of an estimator of a domain total for a small domain of study in double sampling with probabilities proportional to size (pps) when the size, X_n, is known before the start of the survey depends only on the variance of the study variable for elements of the domain within each separate stratum. The reduction in variance of the estimator by knowledge of X is greater when both the proportion of non-domain elements and the deviation of mean of the study variable for the domain elements in each stratum from the population mean for the domain are large while the study variable varies little among the domain elements in each stratum.

Key Words: Domain, Double, Sampling, Unequal, Probabilities

INTRODUCTION

After sample survey data have been collected totals, means or proportions must be estimated. It is usually realised during the analysis of the data that these estimates should also refer to certain subpopulations that are also called domains of study. Estimation of domain parameters is straightforward if the domain of study constitutes a separate stratum of the population and no special theory is necessary: see Udofia (2002b). In most cases, a domain of study cuts across constituent strata of the population with unknown number of elements within each stratum.

The problem of estimation for such domains was first considered in detail by Yates (1953). Yates (1953) notes that the variance of an estimator of a domain parameter is increased by the fact that the number of the domain elements, and hence the number of those elements that can fall in a random sample of a fixed size, is unknown before the start of the survey. Durbin (1958) and Hartley (1959) give a derivation of Yates' results in multi-stage sampling. Scott and Smith (1971) extends Yates' results to stratified multistage sampling while Tin and Toe (1972) extends the same results to Bayes estimation for domains in stratified sampling. Udofia (2002a and 2002b) extend Yates' results to double sampling for pps when information on the size, X, of each sampling unit is unknown. Tripathi (1988) gives extension of Yates' results to estimation of proportions in sampling on two occasions using inverse equal probability sampling. The problem of allocation of resources when domains of study are of primary interest is discussed by Kish (1969). Cochran (1977) page 38 shows that knowledge of the size, N_j, of domain j that is of interest reduces the variance of the estimator of

GODWIN A. UDOFIA, Department of Mathematics, Statistics and Computer Science, University of Calabar, Calabar, Nigeria

domain mean in a single-phase simple random sample design. The reduction in variance is shown to be greater when the proportion of non-domain elements in the population is large and the study variable varies little among the domain elements.

It is well known that stratified random sampling and sampling with pps give a more precise result than equal probability sampling; see Udofia (2002b) for reference. It is also of interest therefore to determine the effect of knowledge of the size, X, of each sampling unit on the variance of an estimator of a domain total in double sampling using probabilities proportional to size. This article is an attempt to provide a mathematical expression for the variance of an estimator of the domain total in double sampling with pps and to determine the effect of knowledge of X on the variance of such an estimator.

SAMPLE DESIGN

The population under study consisting of N elements is divided into H strata with N_h elements in stratum h (h=1,2,...,H). Let D_{hj} denote the part of domain j (j =1,2,...,M) in stratum h. The number of elements, N_{hj}, in D_{hj} is not known. We assume that $0 < N_{hj} < N_h$. Information on X, the size of each element in the population, is known before the start of the survey. For example, in a survey to determine the average household expenditure on consumption items, the household size may be known from records of earlier surveys but whether the household head has the educational qualification required for a particular subclass of the population that is of interest may not be known before the start of the survey. Hence the number of household heads that fall in that particular class remains a random variable.

An initial sample, $S(n_{1h})$, of size n_{1h} is drawn from stratum h(h=1,2,...H) with probabilities p_i (i=1,2,...,N_h) proportional to X and, for the reason given in Udofia (2002b), with replacement. We denote by n_{1hj} the number of units in $S(n_{1h})$ that fall in D_{hj}. For a fixed n_{1h}, the n_{1hj} units constitute an initial random sample from D_{hj}, and we assume that $0 < n_{1hj} < n_{1h}$. A second sample, $S(n_{2h})$, of n_{2h} units, $n_{2h} < n_{1h}$, is drawn from $S(n_{1h})$ by simple random sampling without replacement (SRSWOR) and the study variable, Y, is measured on it. We denote by n_{2hj} the number of units of $S(n_{2h})$ that fall in D_{hj}. These n_{2hj} units constitute a second random sample from D_{hj} and we assume that $0 < n_{2hj} < n_{2h}$. We shall refer to the above sample design as double sampling with probability proportional to size. See Raj (1968), page. 145.

ESTIMATOR OF THE TOTAL OF Y FOR DOMAINS

Let,

$$y'_{hi} = y_{hij} \quad \text{if the } i^{th} \text{ element is in } D_{hj}$$

$$= 0 \text{ if the } i^{th} \text{ element is not in } D_{hj} \qquad ...(1)$$

Then under the above sample design, an unbiased estimator of the population total of Y for domain j is

$$\hat{Y}_j = \sum_{h=1}^{H} \frac{1}{n_{2h}} \sum_{i=1}^{n_{2h}} \frac{y'_{hi}}{P_{hi}}, \qquad ...(2)$$

Proof of unbiasedness:

$$E(\hat{Y}_j) = E_1\left\{E_2(\hat{Y})_j)\right\}$$

From (2)

$$E_2(\hat{Y}_j)=\sum_{h=1}^{H}\frac{1}{n_{2h}}\sum_{i=1}^{n_{2h}}\left(\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{P_{hi}}\right)=\sum_{h=1}^{H}\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{P_{hi}} \qquad ...(3)$$

and hence

$$E(\hat{Y}_j)=E_1\left(\sum_{h=1}^{H}\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{P_{hi}}\right)=\sum_{h=1}^{H}\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\left(\sum_{i=1}^{N_h}P_{hi}\frac{y'_{hi}}{P_{hi}}\right)$$

$$=\sum_{h=1}^{H}\sum_{i=1}^{N_h}y'_{hi}$$

Substitution from (1) gives the result.

$$E(\hat{Y}_j)=\sum_{h=1}^{H}\sum_{i=1}^{N_{hj}}y_{hij}=\sum_{h=1}^{H}Y_{hj}=Y_j$$

where

$$Y_{hj}=\sum_{i=1}^{N_{hj}}y_{hij}$$

Sampling Variance of $\hat{Y}_j$

The sampling variance of $\hat{Y}_j$ is given by the conditional variance formula

$$V(\hat{Y}_j)=V_1E_2(\hat{Y}_j)+E_1V_2(\hat{Y}_j) \qquad ... \quad (4)$$

From (3),

$$V_1E_2(\hat{Y}_j)=V_1\left(\sum_{h=1}^{H}\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{p_{hi}}\right)=\sum_{h=1}^{H}\frac{1}{n_{1h}^2}\sum_{i=1}^{n_{1h}}V_1\left(\frac{y'_{hi}}{P_{hi}}\right)$$

The cross-product terms vanish for all h ≠ k and i ≠ j because of independence of selection and selection with replacement in the first phase of the survey within each stratum.

$$V_1E_2(\hat{Y}_j)=\sum_{h=1}^{H}\frac{1}{n_{1h}}\sum_{i=1}^{N_h}P_{hi}\left(\frac{y'_{hi}}{P_{hi}}-Y'_h\right)^2;Y'_h=\sum_{i=1}^{N_h}y'_{hi} \qquad ...(5)$$

Now

$$V_2(\hat{Y}_j)=V_2\left(\sum_{h=1}^{H}\frac{1}{n_{2h}}\sum_{i=1}^{n_{2h}}\frac{y'_{hi}}{P_{hi}}\right)=\sum_{h=1}^{H}V_2\left(\frac{1}{n_{2h}}\sum_{i=1}^{n_{2h}}\frac{y'_{hi}}{P_{hi}}\right)$$

and since selection during the second phase of the survey was by SRSWOR

$$V_2(\hat{Y}_j)=\sum_{h=1}^{H}\left(\frac{1}{n_{2h}}-\frac{1}{n_{1h}}\right)\frac{1}{n_{1h}-1}\sum_{i=1}^{n_{1h}}\left(\frac{y'_{hi}}{P_{hi}}-\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{P_{hi}}\right)^2$$

and hence

$$EV_2(\hat{Y}_j)=\sum_{h=1}^{H}\left(\frac{1}{n_{2h}}-\frac{1}{n_{1h}}\right)\sum_{i=1}^{N_h}P_{hi}\left(\frac{y'_{hi}}{P_{hi}}-Y'_n\right)^2 \qquad ...(6)$$

since

$$E\left[\frac{1}{n_{1h}-1}\sum_{i=1}^{n_{1h}}\left(\frac{y'_{hi}}{P_{hi}}-\frac{1}{n_{1h}}\sum_{i=1}^{n_{1h}}\frac{y'_{hi}}{P_{hi}}\right)^2\right]=\sum_{i=1}^{N_h}P_{hi}\left(\frac{y'_{hi}}{P_{hi}}-Y'_h\right)^2$$

See Raj (1968), page 146.

Substitution of (5) and (6) in (4) gives the result

$$V(\hat{Y}_j)=\sum_{h=1}^{H}\frac{1}{n_{2h}}\sum_{i=1}^{N_h}P_{hi}\left(\frac{y'_{hi}}{P_{hi}}-Y'_h\right)^2 \qquad ...(7)$$

By substituting for y_{hi}' from (1) in (7) we obtain

$$V(\hat{Y}_j)=\sum_{h=1}^{H}\frac{1}{n_{2h}}\sum_{i=1}^{N_{hj}}P_{hij}\left(\frac{y_{hij}}{P_{hij}}-Y_{hj}\right)^2 \qquad ...(8)$$

where

$$P_{hij}=\frac{N_{hij}}{\sum_{i=1}^{N_{hj}}N_{hij}}$$

is the probability that the i^{th} element falls in D_{hj}.

ESTIMATION OF V ($\hat{Y}_j$)

Since the values of Y for the domain are available only for the subsample of n_{2hj} units of $S(n_{2h})$ and

$$E\left[\frac{1}{n_{2hj}-1}\sum_{i=1}^{n_{2hj}}\left(\frac{y_{hij}}{P_{hij}}-\frac{1}{n_{2hj}}\sum_{i=1}^{n_{2hj}}\frac{y_{hij}}{P_{hij}}\right)^2\right]=E_1\left\{E_2\left[\frac{1}{n_{2hj}-1}\sum_{i=1}^{n_{2hj}}\left(\frac{y_{hij}}{P_{hij}}-\frac{1}{n_{2hj}}\sum_{i=1}^{n_{2hj}}\frac{y_{hij}}{P_{hij}}\right)^2\right]\right\}$$

$$=E_1\left[\frac{1}{n_{1hj}-1}\sum_{i=1}^{n_{1hj}}\left(\frac{y_{hij}}{P_{hij}}-\frac{1}{n_{1hj}}\sum_{i=1}^{n_{1hj}}\frac{y_{hij}}{P_{hij}}\right)^2\right]$$

$$=\sum_{i=1}^{N_{hj}}P_{hij}\left(\frac{y_{hij}}{P_{hij}}-Y_{hj}\right)^2$$

an unbiased estimator of $V(\hat{Y}_j)$ in (8) is given by

$$\hat{V}(\hat{Y}_j) = \sum_{h=1}^{H} \frac{1}{n_{2h}(n_{2hj}-1)} \sum_{i=1}^{n_{2hj}} \left(\frac{y_{hij}}{P_{hij}} - \frac{1}{n_{2hj}} \sum_{i=1}^{n_{2hj}} \frac{y_{hij}}{P_{hij}} \right)^2$$

COMPARISON OF V($\hat{Y}_j$) WITH VARIANCE OF A CORRESPONDING ESTIMATOR IN DOUBLE SAMPLING FOR PPS

When information on the size, X, of each element of the population is not available before the start of the survey, the double sample strategy for estimation of domain total is given by Udofia (2000b) and is called double sampling for pps, see also Raj (1968) page 142. The variance of the estimator of the domain total in double sampling for pps is given by Udofia (2002b) as

$$V(\hat{Y}_j) = \sum_{h=1}^{H} \frac{N_h}{N_h - 1} \frac{n_{1h}-1}{n_{2h} n_{1h}} V_p(y)_{hj} + \sum_{h=1}^{L} \frac{N_h'(N_h - n_{1h})}{(N_h - 1) n_{1h}} \left\{ \left(N_{hj} - 1\right) S^2_{y(hj)} + N_{hj} \left(1 - \frac{N_{hj}}{N_h} \right) \left(\bar{Y}_{hj} - \bar{Y}_j \right)^2 \right\}$$

For large N_h, this can be written as

$$V(\hat{Y}_j) = \sum_{h=j}^{H} \frac{n_{1h}-1}{n_{2h} n_{1h}} V_p(y)_{hj} + \sum_{h=1}^{H} \frac{1}{n_{1h}} N_h^2 P_{hj} \left[S^2_{y(hj)} + Q_{hj} (\bar{Y}_{hj} - \bar{Y}_j)^2 \right] \quad \text{....(9)}$$

where $P_{hj} = N_{hj}/N_h$ is the proportion of elements of domain j in stratum h and $Q_{hj} = 1 - P_{hj}$ is the proportion of non-domain j elements in stratum h. We note that equation (8) can be written as

$$V(\hat{Y}_j) = \sum_{h=1}^{H} \frac{1}{n_{2h}} V_p(y)_{hj} \quad \text{...(10)}$$

where

$$V_p(y)_{hj} = \sum_{i<k} x_{hij} x_{hkj} \left(\frac{y_{hij}}{x_{hij}} - \frac{y_{hkj}}{x_{hkj}} \right)^2$$

corresponds to the same expression in equation 9, see Raj (1968), page. 133.

Whereas, when X is known, the variance of $\hat{Y}_j$ as given in (8) depends only on components of variance of the study variable for elements of the domain in each constituent stratum of the population, when

X is unknown the variance of $\hat{Y}_j$, as given in equation (9), depends on both the variability of the study variable for elements of the domain in each stratum and the variability of the mean of the study variable for subsets of the domain elements among the constituent strata. It is important to determine the major factors that lead to the reduction in variance by a knowledge of X.

Let $V_{1p}(\hat{Y}_j)$ denote the variance in (9) and $V_{2p}\hat{Y}_j$ the variance in (10). Then following the approach in Cochran (1977) page 38,

$$\frac{V_{2p}(\hat{Y}_j)}{V_{1p}(\hat{Y}_j)} = \frac{1}{\sum_{h=1}^{H}\left(1-\frac{1}{n_{1h}}\right)+\sum_{h=1}^{H}(N_h N_{hj} n_{2h}/n_{1h})[1+Q_{hj}(\bar{Y}_{hj}-\bar{Y}_j)^2/V_p(y)_{hj}]}$$

$$\frac{V_{2p}(\hat{Y}_j)}{V_{1p}(\hat{Y}_j)} = \frac{1}{\sum_{h=1}^{H}\left(1-\frac{1}{n_{1h}}\right)+\sum_{h=1}^{H}N_h^2 w_{2h} P_{hj}[1+Q_{hj}(\bar{Y}_{hj}-\bar{Y}_j)^2/V_p(y)_{hj}]} \quad ...(11)$$

where $P_{hj} = N_{hj} / N_h$ and $w_{2h} = n_{2h} / n_{1h}$

REMARK

The estimator of $\hat{Y}_j$ in double sampling with probability proportional to size when X is known is more precise than the corresponding estimator in double sampling for probability proportoional to size when X is not known. Equation (9) has all the attributes proposed by Yates (1953) and proved by all the other results quoted above. Equation (8) is different from all these other results only in the sense that it is not affected by variability of the mean of the study variable for subsets of the domain elements in the different strata. When $P_{hj} = 0$, the domain coincides with a stratum and the reduction in variance of the estimator due to knowledge of X is equal to $\sum_{h=1}^{H}\left(1-\frac{1}{n_{1h}}\right)$ which is the same as in global estimation. In this case, no new theory is necessary as earlier concluded by Udofia (2002b).

In equation (11), N_h, n_{1h} and w_{2h} are fixed numbers. By our assumption $0 \leq Q_{hj} \leq 1$. The reduction in the variance of $\hat{Y}_j$ by knowledge of X and hence double sampling with probability proportional to X is greater when both Q_{hj}, the proportion of non-domain elements in stratum h (h =1, 2, ..., H), and $\bar{Y}_{hj} - \bar{Y}_j$ (the deviation of the mean of the study variable for D_{hj} from the population mean of the study variable for the domain) are large while $V_p(y)_{hj}$, the variance of the study variable for the subclass

D_{hj}, is small. Since strata are usually formed to be internally homogeneous and externally heterogeneous, $V_p(y)_{hj}$ and hence $V_{2p}(\hat{Y}_j)$ will always be small and

$\bar{Y}_{hj} - \bar{Y}_j$ will be large. The magnitude of the reduction in the variance of $\hat{Y}_j$ by knowledge of X therefore depends on efficiency of the stratification procedure.

REFERENCES

Cochran, W. G., 1977. Sampling Techniques. John Wiley and Sons Inc., New York.

Durbin, J., 1958. Sampling Theory for estimates based on fewer individuals than the number selected. Bull. Int. Stat. Inst., 36(3): 113-119.

Hartley, H.O., 1959. Analytical Studies of Survey Data. Instituto di Statistica. Rome.

Kish, L., 1969. Design and estimation for subclasses, comparisons and analytical statistics, in, Johnson, N. L. and Smith, and Harry Jr eds., (Editors), New Developments in Survey Sampling, John Wiley and Sons Inc., New York.

Raj, D., 1968. Sampling Theory. McGraw Hill Book Co., New York.

Scott, A. and Smith, T.M.F., 1971. Bayes estimates for subclasses in stratified sampling. Jour. Amer. Stat. Assoc., 66 : 834 - 8.

Tin, M. and Toe, T., 1972. Estimation for domains in multistage sampling. Jour. Amer. Stat. Assoc., 67:913-916.

Tripathi, T. P., 1988. Estimation for domains in sampling on two occasions. Sankhyā, 50: 103 - 110.

Udofia, G., 2002a. Estimation for domains in double sampling for probabilities proportional to size. Sankhyā, B, in print.

Udofia, G., 2002b. On the precision of an estimator of mean for domains in double sampling for inclusion probabilities. Global Journ. of Pure and Appl. Sciences, 8(4) in print.

Yates, F., 1953. Sampling Methods for Censuses and Surveys. Charles Griffin and Co., London.