GLOBAL JOURNAL OF MATHEMATICAL SCIENCES VOL. 3, NO.2, 2004: 173 - 178
COPYRIGHT(C) BACHUDO SCIENCE CO. LTD. PRINTED IN NIGERIA. ISSN 1596-6208

ESTIMATION FOR SMALL DOMAINS IN DOUBLE SAMPLING FOR STRATIFICATION

GODWIN A. UDOFIA

(Received 9 July 2004; Revision accepted 23 August, 2004).

ABSTRACT

In this article, we investigate the effect of randomness of the size of a small domain on the precision of an estimator of mean for the domain under double sampling for stratification. The result shows that for a small domain that cuts across various strata with unknown weights, the sampling variance depends on the within domain variance, the variance of means of the domain units in each of the various strata and the variance of the proportion of units of the observed sample that belong to the domain.

KEYWORDS: Domain, Double sampling, stratification.

INTRODUCTION

In practical survey situations, efficiency of estimation often requires stratification of the population under study by a given criterion or by two or more criteria. In some situations, the distribution of the stratifying variable is not known before the start of the survey and double sampling for stratification has to be applied. It has been shown [Cochran (1977), p330] that the variance of an estimator of a global mean under double sampling for stratification is affected by variation within the strata and by variation among the strata means.

During analysis of the data, information may be required for certain small subpopulations or domains of study. Where a domain is one of the exhaustive strata of the population under study, no special method is required to calculate the variance of an estimator of the domain mean. The problem arises where a domain cuts across various strata of the population with unknown weights. Yates (1953) first considered in detail some of the problems associated with estimation of such domain totals, means and proportions in simple, random sampling. Extension of Yates' results by various authors dating up to 1988 are listed in Udofia (2002). Udofia (2004) also extends Yates' results further to ratio estimation with subsampling the non-respondents. Given the fact that the size of the domain in each stratum is a random variable with a variance, it is of interest also to see how Yates' results applies to estimation of domain mean in double sampling for stratification. This article is an attempt in this direction.

SAMPLE DESIGN

Let $\Pi = \{U_1, U_2, \ldots, U_N\}$ denote a population of size N under study and D_j denote domain j of size N_j. Let V denote the random variable that defines D_j, U the stratifying variable, N_h the size of stratum h in the population, h = 1, 2, ..., H N_{hj}; the number of units of D_j that are in stratum h, $W_h = N_h / N$, the proportion of the population elements that are in stratum h and $W_{hj} = N_{hj} / N$, the proportion of D_j in stratum h.

We assume that the distribution of U and that of V are unknown before the start of a survey. An initial sample, S_1, of size n_1 is drawn from Π by simple random sampling without replacement and U and V are measured on it. Based on the observed distribution of U, the n_1 sample units are divided into strata and the number of units of D_j that fall in each stratum is determined from the observed distribution of V. Let n_{1h} denote the number of S_1 units that fall in stratum h, $w_{1h} = n_{1h} / n_1$, $n_1 = \sum_h n_{1h}$, n_{1hj} the number of units of D_j that fall in the subsample of n_{1h} units, and $w_{1hj} = n_{1hj} / n_{1j}$where $n_{1j} = \sum_h n_{1hj}$. Then $E(w_{1h}) = W_h$ and $E(w_{1hj}) = W_{hj}$.

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

A subsample of size n_{2h} is drawn from the first subsample of n_{1h} units, h = 1, 2, ..., H. Let $w_{2h} = n_{2h}/n_2$ where $n_2 = \sum_h n_{2h}$ is the total size of S_2, the second sample. The study variable Y, is measured only on S_2. Let n_{2h_j} denote the number of units of D_j in the subsample of n_{2h} units.

ESTIMATION OF MEAN FOR SMALL DOMAINS

Let Y_{hij} denote the value of Y for element *i* in Π that falls in both D_j and stratum *h*, and y_{hij} be similarly defined for unit *i* in the sample. The mean of Y for the part of D_j in stratum *h* is $\bar{Y}_{hj} = \frac{1}{N_{hj}}\sum_{i=1}^{N_{hj}} Y_{hij}$ and hence the mean of Y over D_j is

$$\bar{Y}_j = \frac{1}{N_j}\sum_{h=1}^{H}\sum_{i=1}^{N_{hj}} Y_{hij} = \sum_h W_{hj}\bar{Y}_{hj}\,. \qquad \ldots \qquad (1)$$

An unbiased estimator of $\bar{Y}_j$ is

$$\hat{\bar{y}}_{jst} = \sum_{h=1}^{H} w_{1hj}\bar{y}_{2hj}\,; \quad \bar{y}_{2hj} = \frac{1}{n_{2hj}}\sum_{i=1}^{n_{2hj}} y_{hij}\,.$$

Since Y was not measured on S_1, $\bar{y}_{2hj}$ is treated as the mean of a random sample from a subpopulation of N_{hj} units and hence,

$$E\left(\hat{\bar{y}}_{jst}\right) = E_1\left[E_2\left(\sum_h w_{1hj}\bar{y}_{2hj}\right)\right] = E_1\left(\sum_h w_{1hj}\bar{Y}_{hj}\right)$$
$$= \sum_h W_{hj}\bar{Y}_{hj} = \bar{Y}_j$$

where $E_2(\cdot) = E(\cdot \backslash S_1)$ and $E_1(\cdot) = E(\cdot \backslash \Pi)$.

The variance of $\hat{\bar{y}}_{jst}$ is given by the following conditional formula:

$$V\left(\hat{\bar{y}}_{jst}\right) = V_1E_2\left(\hat{\bar{y}}_{jst}\right) + E_1V_2\left(\hat{\bar{y}}_{jst}\right) \qquad \ldots \qquad (2)$$

We now assume [Cochran (1977), p328] that y_{hij} was measured on all the n_{1hj} first sample units. Then,

$$E_2\left(\hat{\bar{y}}_{jst}\right) = \sum_h w_{1hj}E_2\left(\bar{y}_{2hj}\right) = \sum_h w_{1hj}\bar{y}_{1hj} = \bar{y}_{1j}$$

and hence, if we ignore the randomness of n_{1hj} and consider N_{hj} and N_j as known quantities,

$$V_1E_2\left(\hat{\bar{y}}_{jst}\right) = V_1\left(\bar{y}_{1j}\right) = \left(\frac{1}{n_{1j}} - \frac{1}{N_j}\right)S_j^2 \qquad \ldots \qquad (3)$$

where

$$S_j^2 = \frac{1}{N_j - 1}\sum_h^H \sum_{i=1}^{N_{hj}}\left(Y_{hij} - \bar{Y}_j\right)^2 \qquad \ldots \qquad (4)$$

Now

$$V_2\left(\hat{\bar{y}}_{jst}\right) = V_2\left(\sum_h w_{1hj}\bar{y}_{2hj}\right) = \sum_h w_{1hj}^2 V_2\left(\bar{y}_{2hj}\right).$$

Again, if we ignore the randomness of n_{2hj}, we obtain

$$E_1 V_2\left(\hat{\bar{y}}_{jst}\right) = E_1 \sum_h w_{1hj}^2 \left(\frac{1}{n_{2hj}} - \frac{1}{n_{1hj}}\right) S_{hj}^2$$

$$= \frac{1}{n_{1j}} \sum_h W_{hj} \left(\frac{1}{v_{hj}} - 1\right) S_{hj}^2 \qquad \ldots \qquad (5)$$

where $v_{hj} = n_{2hj}/n_{1hj}$ and $S_{hj}^2 = \frac{1}{N_{hj}-1}\sum_{i=1}^{N_{hj}} \left(Y_{hij} - \bar{Y}_{hj}\right)^2$. Substitution of (3) and (5) in (2) gives the result

$$V\left(\hat{\bar{y}}_{jst}\right) = \left(\frac{1}{n_{1j}} - \frac{1}{N_j}\right) S_j^2 + \frac{1}{n_{1j}} \sum_h W_{hj} \left(\frac{1}{v_{hj}} - 1\right) S_{hj}^2 . \qquad \ldots \qquad (6)$$

This corresponds to the following result given by Cochran (1977), page 329:

$$V\left(\hat{\bar{y}}_{st}\right) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S^2 + \frac{1}{n_1} \sum_h W_h \left(\frac{1}{v_h} - 1\right) S_h^2 \qquad \ldots \qquad (7)$$

for the estimator, $\bar{y}_{st}$.

From (4), we obtain

$$\frac{N_j - n_{1j}}{N_j n_{1j}} S_j^2 = \frac{g_j}{n_{1j}} \sum_{h=1}^{H} \left(W_{hj} - \frac{1}{N_j}\right) S_{hj}^2 + \frac{g_j}{n_{1j}} \sum_{h=1}^{H} W_{hj} \left(\bar{Y}_{hj} - \bar{Y}_j\right)^2$$

where $g_j = (N_j - n_{1j})/(N_j - 1)$.

Substitution of this result in (6) gives, after rearrangement,

$$V\left(\hat{\bar{y}}_{jst}\right) = \frac{1}{n_{1j}} \sum_h W_{hj} \left(\frac{1}{v_{hj}} - 1\right) S_{hj}^2 + \frac{g_j}{n_{1j}} \sum_h \left(W_{hj} - \frac{1}{N_j}\right) S_{hj}^2 + \frac{g_j}{n_{1j}} \sum_h W_{hj} \left(\bar{Y}_{hj} - \bar{Y}_j\right)^2 \ldots \qquad (8)$$

which corresponds to

$$V\left(\hat{\bar{y}}_{st}\right) = \sum_h \frac{W_h S_h^2}{n_1} \left(\frac{1}{v_h} - 1\right) S_{hj}^2 + \frac{g'}{n_1} \sum_h \left(W_h - \frac{1}{N}\right) S_h^2 + \frac{g'}{n_1} \sum_h W_h \left(\bar{Y}_h - \bar{Y}\right)^2 \qquad \ldots \qquad (9)$$

where $g' = (N - n_1)/(N-1)$. The only difference between (8) and (9) is that whereas $V\left(\bar{y}_{jst}\right)$ depends on variation among units of D_j in stratum *h* and on variation of means of those units over the various strata, $V\left(\bar{y}_{st}\right)$ depends on variation among all units within the various strata and variation of the different strata means. In other words, $V\left(\bar{y}_{jst}\right)$ depends only on domain parameters while $V\left(\bar{y}_{st}\right)$ depends on parameters of the strata.

4. ESTIMATION OF VARIANCE OF ESIMATOR OF DOMAIN MEANS

We now recognize the fact that n_{1hj} and n_{2hj} are random variables. Let $y'_{hi} = y_{hij}$ if unit *i* in stratum *h* is also in D_j and $x'_{hi} = 1$ if unit *i* of stratum *h* is in D_j and zero otherwise. Then,

$$\bar{y}'_{2h} = \frac{n_{2hj}}{n_{2h}} \bar{y}_{2hj} \; ; \quad \bar{x}'_{2h} = \frac{n_{2hj}}{n_{2h}} \; ;$$

$$\bar{y}'_{1h} = \frac{n_{1hj}}{n_{1h}} \bar{y}_{1hj} \; ; \quad \bar{x}'_{1h} = \frac{n_{1hj}}{n_{1h}} \; ;$$

and hence, $\bar{y}_{2hj}$ and $\bar{y}_{1hj}$ can be expressed as ratios

$$\bar{y}_{2hj} = \frac{\bar{y}'_{2h}}{\bar{x}'_{2h}} \quad and \quad \bar{y}_{1hj} = \frac{\bar{y}'_{1h}}{\bar{x}'_{1h}}.$$

We also note that $\bar{y}_{1j} = \sum_h w_{1hj}\bar{y}_{1hj}$. Thus an estimator of $E_1V_2\left(\hat{\bar{y}}_{jst}\right)$ is

$$E_1\hat{V}_2\left(\hat{\bar{y}}_{jst}\right) = \hat{V}\left(\sum_h w_{1hj}\bar{y}_{2hj} \mid w_{1hj}\right) = \sum_h w_{1hj}^2 \hat{V}\left(\frac{\bar{y}'_{2h}}{\bar{x}'_{2h}}\right)$$

From Durbin (1958) and Udofia (2002), this can be calculated in terms of the second sample information as

$$E_1\hat{V}_2\left(\hat{\bar{y}}_{jst}\right) = \sum_h w_{1hj}^2\left(\frac{1}{n_{2h}} - \frac{1}{n_{1h}}\right)\frac{1}{n_{2h}-1}\sum_{i=1}^{n_{2h}}\left\{y'_{hi} - \hat{\bar{y}}_{jst}x_i - \left(\bar{y}'_{2h} - \hat{\bar{y}}_{jst}\bar{x}'_{2h}\right)\right\}^2$$

$$= \frac{1}{n_{1j}}\sum_h w_{1hj}\left(\frac{1}{v_{2h}} - 1\right)n_{1hj}\left\{\frac{n_{2hj}-1}{n_{1h}(n_{2h}-1)}s_{2hj}^2 + \frac{n_{2h}p_{2hj}q_{2hj}}{n_{1h}(n_{2h}-1)}\left(\bar{y}_{2hj} - \hat{\bar{y}}_{jst}\right)^2\right\} \qquad \ldots \qquad (10)$$

where $p_{2hj} = n_{2hj}/n_{2h}\, q_{2hj} = 1 - p_{2hj}$, $v_{2h} = n_{2h}/n_{1h}$, and $s_{2hj}^2 = \sum_{i=1}^{n_{2hj}}\left(y_{hij} - \bar{y}_{2hj}\right)^2 \Big/ (n_{2hj}-1)$.

Similarly,

$$\hat{V}_1E_2\left(\hat{\bar{y}}_{jst}\right) = \hat{V}_1\left(\sum_h w_{1hj}\bar{y}_{1hj}\right) = \sum_h w_{1hj}^2 \cdot \hat{V}_1\left(\frac{\bar{y}'_{1h}}{\bar{x}'_{1h}}\right)$$

$$= \sum_h w_{1hj}^2\left(\frac{1}{n_{1h}} - \frac{1}{N_h}\right)\frac{1}{n_{2h}-1}\sum_{i=1}^{n_{2h}}\left\{y'_{hi} - \hat{\bar{y}}_j x_i - \left(\bar{y}'_{2h} - \hat{\bar{y}}_{jst}\bar{x}'_{2h}\right)\right\}^2$$

$$= \frac{1}{n_{1j}}\sum_h w_{1hj}\left(1 - f_{1h}\right)n_{1hj}\left\{\frac{n_{2hj}-1}{n_{1h}(n_{2h}-1)}s_{2hj}^2 + \frac{n_{2h}p_{2hj}q_{2hj}}{n_{1h}(n_{2h}-1)}\left(\bar{y}_{2hj} - \hat{\bar{y}}_{jst}\right)^2\right\}$$

where $f_{1h} = n_{1h}/N$.

... (11)

Substitution of (10) and (11) in (2) gives the result

$$\hat{V}\left(\hat{\bar{y}}_{jst}\right) = \frac{1}{n_{1j}}\sum_h w_{1hj}\left(\frac{1}{v_{2h}} - 1\right)n_{1hj}\frac{n_{2hj}-1}{n_{1h}(n_{2h}-1)}s_{2hj}^2 + \frac{1}{n_{1j}}\sum_h w_{1hj}\left(1 - f_{1h}\right)n_{1hj}\frac{n_{2hj}-1}{n_{1h}(n_{2h}-1)}s_{2hj}^2 +$$

$$+ \frac{1}{n_{1j}}\sum_h w_{1hj}\left[\left(\frac{1}{v_{2h}} - 1\right) + \left(1 - f_{1h}\right)\right]n_{1hj}\frac{n_{2h}p_{2hj}q_{2hj}}{n_{1h}(n_{2h}-1)}\left(\bar{y}_{2hj} - \hat{\bar{y}}_{jst}\right)^2 \qquad \ldots \qquad (12)$$

This is not the same as the corresponding expression for a global estimator, $\bar{y}_{jst}$, as given in Theorem 12.3 in Cochran (1977). The difference arises from the dependence of $V\left(\bar{y}_{jst}\right)$ in (12) on the variability of the number of units from the domain in the H subsamples with n_{kh} units in the k^{th} phase of the survey, h =

1, 2, ..., H. We note that

$$\sum_h w_{1hj}^2 \frac{p_{2hj} q_{2hj}}{n_{2hj}-1} = \hat{Var}(p_{2hj})$$

where p_{2hj} denotes the proportion of units of stratum *h* in the sample that also belong to domain *j*. Thus, the randomness of n_{1hj} and that of n_{2hj} increase the sampling variance of the estimator of the domain mean by multiplying the positive contribution of variation of $\bar{y}_{hj}$ over *h* by $\hat{Var}(P_{hj})$ which is positive. This phenomenon is not an attribute of the estimator of $V(\bar{y}_{st})$ as given by Cochran (1977) on page 329. The estimator $\hat{\bar{y}}_{jst}$ is consistent and unbiased but it requires a much larger initial sample than the corresponding global estimator.

5. APPLICATION

In 1993, the above survey design was applied in a survey of fishing communities in Umon Island in the Central region of Cross River State, Nigeria. The purpose of the survey was to determine the average value of fish caught per commercial fisher-folk per month. The estimated population was 2,500 and this was to be divided into two strata according to the number of years of experience in commercial fishing. Those with 1 to 3 years of experience were to be put in stratum 1 while those with more than 3 years of experience were to be put in stratum 2.

The number of years spent in fishing by fisher-folks on the Island was not known before the start of the survey. An initial sample of 230 fisher-folks was drawn by simple random sampling without replacement and information on number of years spent in fishing was obtained from 223 of the sample units. The other 7 persons in the sample did not respond. Of the 223 persons who responded, 145 belonged to stratum 1 while the other 78 belonged to stratum 2. A second sample of 108 persons was drawn by simple random sampling without replacement from the 223 persons in the initial sample who responded. The value of fish caught in the month that immediately preceded the survey was obtained from each person in the second sample.

During analysis of the survey data, a special demand was made for an estimate of the average value of fish caught per month per native commercial fisher-folk on the Island. The community of native commercial fisher-folks in the study area thus formed a domain of study which cut across the above two strata with unknown weights. A summary of the survey results is given below:

SUMMARY OF RESULT OF A SURVEY OF FISHING COMMUNITIES IN UMON ISLAND, NIGERIA, 1993.

Stratum	First Sample				Second Sample						
	n_{1h}	n_{1hj}	w_{1hj}	f_{1h}	n_{2h}	n_{2hj}	$\bar{y}_{2hj}$	v_{2h}	s_{2hj}^2	p_{2hj}	q_{2hj}
1	145	68	0.819	.058	70	13	22	.48	19.67	.186	.814
2	78	15	.181	.031	38	8	38	.49	23.14	.211	.789
Total	223	83	1.00	-	108	21	-	-	-	-	-

$$\hat{\bar{y}}_{jst} = 24.89$$

For convenience of computation, equation 12 can be expressed as

$$\hat{V}(\hat{\bar{y}}_{jst}) = \frac{1}{n_{1j}} \sum_{h=1}^{H} w_{1hj}^2 \left(\frac{1}{v_{2h}} - f_{1h} \right) \frac{1}{n_{2h}-1} \left[(n_{2hj}-1) s_{2hj}^2 + n_{2h}\, p_{2hj}\, q_{2hj} (\bar{y}_{2hj} - \hat{\bar{y}}_{jst})^2 \right] \qquad \ldots (13)$$

$$= \frac{1}{n_{1j}} \sum_{h=1}^{H} w_{1hj}^2 \left(\frac{1}{v_{2h}} - f_{1h} \right) \frac{n_{2h}(n_{2hj}-1)}{n_{2h}-1} \left[\frac{s_{2hj}^2}{n_{2h}} + \frac{p_{2hj}\, q_{2hj}}{n_{2hj}-1} (\bar{y}_{2hj} - \hat{\bar{y}}_{jst})^2 \right] \qquad \ldots (14)$$

Substitution of the above data in equation 13 gives the result $\hat{V}(\hat{\bar{y}}_{jst}) = 8.6/83 = 0.1036$

REFERENCES

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

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

Udofia, G. A., 2002. On the precision of an estimator of mean for domains in double sampling for inclusion probabilities. Global Journ. of Mathematical Sciences, 18(2): 49-58.

Udofia, G. A., 2004. Ratio estimation for small domains with subsampling the non-respondents: An application of Rao strategy. Statistics in Transition, 6(5): 713-724.

Yates, F., 2002. Sampling Methods for Censuses and Surveys. Charles Griffin and Co., London.