ON THE PROBLEMS OF PPS SAMPLING IN MULTI-CHARACTER SURVEYS

A. C. AKPANTA

(Received 16 October 2008; Revision Accepted 27 March 2009)

ABSTRACT

This paper, which is on the problems of PPS sampling in multi-character surveys, compares the efficiency of some estimators used in PPSWR sampling for multiple characteristics. From a superpopulation model, we computed the expected variances of the different estimators for each of the first two finite populations considered, as well as the exact bias and variance of each of these estimators. The results obtained show that the estimators proposed by Rao (1966), Amahia et. al. (1989) and the alternative in Amahia et. al. (1989) are better than the conventional estimator. In population I, where the study variable and the ancillary variable are highly and positively correlated, results show that the estimator in Amahia et. al. (1989) fare better than the alternative estimator. On the other hand, the results obtained from our population II where the correlation between the study variable and the ancillary variable is poor, reveal that the alternative estimator in Amahia et. al. (1989) is more efficient. Several other finite populations whose ρ are neither too high as in population I nor too poor as in population II were considered and it was discovered that the competition for efficiency only rests with the two estimators suggested by Amahia et al (1989) and Rao (1966). These interesting comparative results are shown in Tables.

KEY WORDS: PPS Sampling, estimators, superpopulation, variances, biases.

INTRODUCTION

In sampling, the sampling units, as usually defined, are similar in size and structure. However, with some types of population it is convenient or necessary to use sampling units that differ in size, thus the farm is often the sampling unit for collecting agricultural data, though farms in the same region may vary in land acreage from a few acres to over 1,000 acres. Similarly, when obtaining information about sales or prices, the sampling unit may be a dealer or store, these ranging from small to large concerns.

Again, the total sample size can be subject to unduly large variation if it is based on random selection of clusters that differ greatly in size. If we subsample the selected clusters at a fixed rate, the expected sizes of the subsamples are proportional to the unequal cluster sizes. The total sample size depends on which clusters happen to fall into the sample.

In such cases as mentioned above the question arises: should differences between the sizes of the sampling units be ignored or taken into consideration in selecting the sample and in making estimates from the results of the sample? The differences should not be ignored otherwise there would be uncontrolled random sampling.

To account for the differences between the sizes of the sample units we have sampling with varying (unequal) probabilities. The commonest of this type of sampling is sampling with probability proportional to 'size' (PPS), the size being the value of the ancillary variable, (Cochran, 1946). This procedure uses the values of the ancillary variable in such a way that unequal probabilities of selection are assigned to the population units. Hence, if the values of an ancillary variable related to the study variable were known for all the N units, the information could be used in selecting the samples so as to provide estimators with greater efficiency than those from simple random sampling. PPS sampling can be with replacement (PPSWR) where any unit drawn is replaced before the next draw is made. It can also be without replacement (PPSWOR) where there is no replacement of any unit drawn before drawing the next.

A. C. Akpanta, Department of Statistics, Abia State University, PMB 2000, Uturu, Abia State, Nigeria

Nevertheless, in large scale surveys it will be quite uneconomical to carry out such surveys for the main purpose of estimating one parameter when in actual fact many other parameters could be estimated with little or no additional cost. Therefore, it is usually of interest to estimate parameters relating to several characteristics in such cases. Hence, only a single measure of size can be used in selecting the sample of primary units with PPS. In such PPS sampling in multi-character surveys, some characteristics may not be related to the size (the ancillary variable). This situation has led to the development of many alternative estimators which shall be extensively examined in this work.

Recognition of the value of sampling with probability proportional to size especially as prelude to subsampling and when stratification with respect to other characteristics is desired – is due to Hensen and Hurwitz (1943). These men introduced the use of primary unit with probability proportional to some measure of their size for sampling of one primary sampling unit per stratum. Lahiri, (1951) advanced a sampling scheme where the sample is selected with probability proportional to the total size of the ancillary variable. He also presented a method for actually drawing the sample, which avoids the need for listing all possible samples and finding their total size and their cumulative sizes. Grundy (1954) developed a practical method of drawing sampling units with probabilities exactly proportional to size, in which both preliminary calculations and the addition of a large number of sizes are avoided. This method is considered as an extension of Lahiri's method for samples of one.

Although since 1934 a phenomenal number of learned papers have been written extolling the virtues of various modes of sampling with unequal probabilities, not much has been done in the area of PPS sampling in multi-character surveys. Rao (1966) who first looked into this area suggested an alternative estimator of the population total for characteristics which are poorly correlated with the selection probabilities in probability proportional to size sampling schemes for multi characteristics. He further compared these alternative estimators with the conventional estimators under a superpopulation model. It is shown that the average variance of the alternative estimators is smaller than the average variance of the conventional estimators under their superpopulation model. For making efficiency comparison between the usual estimators and the alternative estimators he proposed, Rao regarded the finite population as being drawn from an infinite superpopulation in which the study variable, *y*, and the ancillary variable, *x*, are independent. The results obtained do not apply to any single finite population but to the average of all finite populations that can be drawn from the superpopulation.

Bansal and Singh (1985) put forward another alternative estimator of the population total for characteristics that are poorly correlated with the selection probabilities. They suggested another alternative estimator of the population total for probability proportional to size with replacement sampling scheme which considers the rough value of the correlation coefficient between the study variable y, and the ancillary variable x.

Their action (suggesting another alternative estimator) is informed by the fact that the situation considered in the model in Rao (1966) is "not commonly encountered in practice, since hardly can the correlation in the population be exactly equal to zero". Though Bansal and Singh mention that the bias of their estimator is expected to be smaller than that of the corresponding estimator in Rao (1966), they did not derive any expressions for the bias and did not make the necessary comparison. However, the expressions and the condition needed to show that their proposed estimator is more efficient than Rao's estimator are, to put mildly, 'quite complicated and difficult' to handle algebraically.

So far, we see that irrespective of the several estimators proposed in this case of PPS sampling in multi-character surveys, none of these can be considered to be entirely satisfactory from the point of view of precision and also applicability in practice.

No wonder Amahia et al. (1989) in their work They also studied the efficiency of their estimators compared with other related estimators.

RELEVANT THEOREMS AND DATA SELECTION

THEOREM 1: If a sample of size n units is drawn with PPS of x_i and with replacement, then

$$\hat{Y}_c = \frac{1}{n}\sum_i^n \frac{y_i}{p_i} \qquad 1$$

is an unbiased estimate of the population total Y with variance

$$\hat{Y}_c = \frac{1}{n}\sum_i^n p_i\left(\frac{y_i}{p_i} - Y\right)^2 \qquad 2$$

Proof: Let t_i be the number of times that the i^{th} unit appears in a specific sample of size n, where t_i may have any of the values 0,1,2,…,n. Consider the joint frequency distribution of the t_i for all N units in the population. The method of drawing the sample is equivalent to the standard probability problem in which n balls are thrown into N boxes, the probability that a ball goes into the i^{th} box being p_i at every throw. Consequently the joint distribution of the t_i is the multinomial expression

$$\frac{n!}{t_1!t_2!...t_n!}p_1^{t_1}p_2^{t_2}...p_N^{t_N}$$

For the multinomial, the following properties of the distribution of t_i are well known:

$$E(t_i) = np_i \qquad 3$$

$$V(t_i) = np_i(1 - p_i) \qquad 4$$

$$Cov(t_i t_j) = -np_i p_j \qquad 5$$

we may therefore write (1) as

$$\hat{Y}_c = \frac{1}{n}\left(t_1\frac{y_1}{p_1} + t_2\frac{y_2}{p_2} + ... + t_N\frac{y_N}{p_N}\right) = \frac{1}{n}\sum_i^N t_i\frac{y_i}{p_i} \qquad 6$$

Where the sum extends over all units in the population. In repeated sampling the t_i are the random

Variables, whereas the y_i *and the* p_i are a set of fixed numbers.

Hence, since $E(t_i) = np_i$ by (3) we have,

$$E(\hat{Y}_c) = \frac{1}{n}\sum_{i=1}^N (np_i)\frac{y_i}{p_i} = \sum_{i=1}^N y_i = Y$$

Therefore $\hat{Y}$ is unbiased.

Concerning the variance we have :

$$V(\hat{Y}_c) = \frac{1}{n^2}\left[\sum_i^N\left(\frac{y_i}{p_i}\right)^2 V(t_i) + 2\sum_{i=1}^N\sum_{j>1}^N \frac{y_i y_j}{pip_j}Cov(t_i t)\right]$$

$$= \frac{1}{n^2}\left[\sum_i^N\left(\frac{y_i}{p_i}\right)^2 np_i(1-p_i) - 2\sum_{i=1}^N\sum_{j>1}^N \frac{y_i y_j}{pip_j}np_i p_j\right]$$

$$= \frac{1}{n}\left(\sum_i^N \frac{y_i^{\,2}}{p_i} - Y^2\right)$$

$$\therefore V(\hat{Y}_c) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 \qquad 7$$

since $\sum_{i=1}^{N} p_i = 1.$ This completes the proof.

Theorem 2: If a sample of n units is drawn with PPS of x_i and with replacement, an unbiased sample estimate of $V(\hat{Y}_c)$ is, for any $n > 1$,

$$v(\hat{Y}_c) = \sum_{i}^{n} \frac{(y_i/p_i - \hat{Y}_c)^2}{n(n-1)} \qquad 8$$

Proof: By the usual algebraic identity,

$$\sum_{i}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_c\right)^2 = \sum_{i}^{n}\left[\left(\frac{y_i}{p_i} - Y\right) - (\hat{Y}_c - Y)\right]^2$$

$$= \sum_{i}^{n}\left[\left(\frac{y_i}{p_i} - Y\right)^2 - 2\left(\frac{y_i}{p_i} - Y\right)(\hat{Y}_c - Y) + (\hat{Y}_c - Y)^2\right]$$

$$= \sum_{i}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 - 2(\hat{Y}_c - Y)\sum_{i}^{n}\left(\frac{y_i}{p_i} - Y\right) + n(\hat{Y}_c - Y)^2 \qquad (*)$$

But

$$\sum_{i}^{n}\left(\frac{y_i}{p_i} - Y\right) = \sum_{i}^{n}\frac{y_i}{p_i} - nY$$

$$= n\sum_{i}^{n}\frac{y_i}{np_i} - nY$$

$$= n\hat{Y}_c - nY = n(\hat{Y}_c - Y)$$

$$= n\left(\frac{1}{n}\sum\frac{y_i}{p_i} - Y\right)$$

$\therefore (*)$ becomes

$$\sum_{i}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_c\right)^2 = \sum\left(\frac{y_i}{p_i} - Y\right)^2 - 2n(\hat{Y}_c - Y)^2 + n(\hat{Y}_c - Y)^2$$

$$= \sum_{i}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 - n(\hat{Y}_c - Y)^2 \qquad 9$$

$$= \sum_{i}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 - n\left(\frac{1}{n}\sum\frac{y_i}{p_i} - Y\right)^2$$

$\therefore$ from (8) we have

$$n(n-1)v(\hat{Y}_c)=\sum_i^n\left(\frac{y_i}{p_i}-Y\right)^2-n(\hat{Y}_c-Y)^2$$

$$E[\ n(n-1)v(\hat{Y}_c)\ ]=E[\ \sum_i^n\left(\frac{y_i}{p_i}-Y\right)^2-n(\hat{Y}_c-Y)^2\ ]$$

$$=E\left[\sum_{i=1}^n\left(\frac{y_i}{p_i}-Y\right)^2\right]-nE(\hat{Y}_c-Y)^2$$

$$=E\left[\sum\left(\frac{y_i}{p_i}-Y\right)^2\right]-nV(\hat{Y}_c) \qquad 10$$

By definition of $V(\hat{Y})$.

Introducing the variable t_i, we have

$$n(n-1)E[v(\hat{Y}_c)]=E\left[\sum_{i=1}^N t_i\left(\frac{y_i}{p_i}-Y\right)^2\right]-nV(\hat{Y}_c)$$

$$=n\sum_{i=1}^N p_i\left(\frac{y_i}{p_i}-Y\right)^2-nV(\hat{Y}_c) \qquad 11$$

But from (7), we have

$$nV(\hat{Y}_c)=\sum_{i=1}^N p_i\left(\frac{y_i}{p_i}-Y\right)^2$$

Hence (11) becomes

$$n(n-1)E[v(\hat{Y}_c)]=n^2V(\hat{Y}_c)-nV(\hat{Y}_c)$$

$$=n(n-1)V(\hat{Y}) \qquad 12$$

Using (2) from the first theorem

$$\therefore E[v(\hat{Y}_c)]=V(\hat{Y}_c).$$

Hence, an unbiased sample estimate of $V(\hat{Y}_c)$ is given by

$$v(\hat{Y}_c)=\sum_{i=1}^n\left(\frac{y_i}{p_i}-\hat{Y}_c\right)^2\Big/n(n-1). \qquad 13$$

The above result will be smaller than the corresponding result for

$$\hat{Y}=\frac{N}{n}\sum_i^n y_i$$

in equal probability sampling if and only if the correlation between $y \text{ and } x$ is high and positive. On the other hand, if $y \text{ and } x$ are unrelated $\hat{Y}$ would have a smaller variance than $\hat{Y}_c$. According to Rao (1966) if y is unrelated to x and the sample is drawn with PPS of x_i and with replacement, an alternative estimator to (1) becomes

$$\hat{Y}_R=\frac{1}{n}\sum_i^n\frac{Ny_i p_i}{p_i}$$

$$= \frac{N}{n}\sum_{i}^{n} y_i \qquad 14$$

Which although bias is more efficient than the conventional unbiased estimator $\hat{Y}_c$ in (1)

Note that we obtained (14) by replacing y_i in (1) with $Ny_i p_i$. The estimator $\hat{Y}_R$ is biased since

$$E\left(\hat{Y}_R\right) = N\sum_{i}^{N} y_i p_i \qquad 15$$

The bias of $\hat{Y}_R$ becomes

$$B\left(\hat{Y}_R\right) = N\sum_{i}^{N} y_i p_i - Y$$

$$= \sum_{i}^{N} y_i \left(Np_i - 1\right) \qquad 16$$

An unbiased estimator of $V\left(\hat{Y}_R\right)$ is obtained by replacing y_i in (8) with $Ny_i p_i$ to get

$$\hat{V}\left(\hat{Y}_R\right) = \frac{N^2}{n\left(n-1\right)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \qquad 17$$

where $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$.

Observe that the estimator $\hat{Y}_R$ in (13) and its variance in (16) have the form as the estimator $\hat{Y}$ and its variance estimator in equal probability sampling.

Since the correlation in the population is never exactly equal to zero (a condition implied by Rao's estimator),Bansal and Singh (1985) developed a new estimator of the population total for characteristics that are poorly correlated with the selection probabilities as

$$\hat{Y}_{BS} = \frac{1}{n}\sum_{i}^{n} \frac{y_i}{p_i^{BS}} \qquad 18$$

where

$$p_i^{BS} = \left(1 + \frac{1}{N}\right)^{1-\rho}\left(1 + p_i\right)^{\rho} - 1 \qquad 19$$

and $\rho =$ the correlation coefficient between $y_i \text{ and } x_i$.

Note that (19) becomes $\frac{1}{N}$ if $\rho = 0$ and $p_i \text{ if } \rho = 1$. Consequently (18) reduces to $\hat{Y}_R \text{ at } \rho = 0 \text{ and to } \hat{Y}_c \text{ at } \rho = 1$.

Amahia et el. (1989) give a much simpler alternative estimator for the situations discussed by Bansal and Singh which is intuitively simple and easy to interpret as

$$\hat{Y}_p = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i p_i / p^*_i}{p_i} \qquad 20$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p^*_i} \qquad 21$$

The result is obtained by replacing $y_i \text{ by } y_i p_i / p_i^* \text{ in}$ (1) where

$$p_i^* = (1-\rho)\frac{1}{N} + \rho p_i \qquad 22$$

If $\rho = 0$ (22) becomes $\frac{1}{N}$, thereby reducing the estimator $\hat{Y}_p$ in (21) to Rao's estimator $\hat{Y}_R$ in (14).

But if $\rho = 1$, (22) reduces to p_i and hence, $\hat{Y}_p$ becomes equal to the conventional estimator $\hat{Y}_C$ in (1).

The unbiased estimator of the variance $V\left(\hat{Y}_p\right)$ is obtained by replacing y_i by $y_i p_i / p_i^*$ in (13). Therefore,

$$\hat{V}\left(\hat{Y}_p\right) = \sum_{i=1}^{n}\left\{\frac{y_i}{p^*_i} - \left(\sum \frac{y_i}{np_i^*}\right)\right\}^2 / n(n-1) \qquad 23$$

The referee to Amahia et el. (1989) suggested an alternative estimator worth-mentioning, i e

$$\hat{Y}'_p = (1-p)\hat{Y}_R + \rho\hat{Y}c \qquad 24$$

This estimator $\hat{Y}'_p$ is easy to construct and also is motivated by the fact that

$$B\left(\hat{Y}'_p\right) < B\left(\hat{Y}_R\right)$$

Note that (24) can be written as

$$\hat{Y}'_p = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i p_i / p'_i}{p_i} \qquad 25$$

$$\therefore \quad \hat{Y}'_p = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p'_i} \qquad 26$$

where

$$p'_i = \left[(1-\rho)N + \frac{\rho}{p_i}\right]^{-1} \qquad 27$$

The variance estimate of $V\left(\hat{Y}'_p\right)$ could be obtained by replacing y_i *by* $y_i p_i / p'_i$ in (13), i.e.

$$\hat{V}\left(\hat{Y}'_p\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{\frac{y_i}{p'_i} - \left(\sum_{i=1}^{n}\frac{y_i}{np'_i}\right)\right\}^2 \qquad 28$$

the biases of the proposed estimators by amahia et al.(1989) and their referee are respectively given as:

$$B\left(\hat{Y}_p\right) = \sum_{i=1}^{n} y_i\left(\frac{p_i}{p_i^*} - 1\right) \qquad 29$$

and

$$B\left(\hat{Y}'_p\right) = (1-\rho)B\left(\hat{Y}_R\right) \qquad 30$$

From (30) it becomes clear that

$$\left|B\left(\hat{Y}_p\right)\right| < \left|B\left(\hat{Y}_R\right)\right| \qquad 31$$

Recall (16):

$$B\left(\hat{Y}_R\right) = \sum_{i-1}^{N} y_i(Np_i - 1)$$

$$\therefore \; B\left(\hat{Y}_R\right) - B\left(\hat{Y}_p\right) = \sum_{i=1}^{N} y_i(Np_i - 1) - \sum_{i=1}^{n} y_i\left(\frac{p_i}{p_i^*} - 1\right)$$

$$\sum_{i=1}^{N} y_i\left(Np_i - \frac{p_i}{p_i^*}\right) > 0 \qquad 32$$

METHODS OF PPSWR SAMPLING

Among the methods of PPSWR include: Cumulative method; Grundy's method (1954); method of selecting from a map; PPS systematic sampling and the Lahiri's method (1951); which we shall adopt in this paper.

In using this method we let α denote $X_{\max}$(obtained by inspection),then choose a random number r in the range $0 < r \leq \alpha$, and a random integer S in the range $1 \ to \ N$. If $r \leq x_s$,accept unit S as a member of the sample, otherwise, try another pair of random numbers. Continue in such a manner until the required number of sample units are obtained .Naturally, this method involves the fewest rejections when x_i do not differ too much in size (Cochran 1977, pp.251).

From a hypothetical census population of 20 villages, we select PPS sample of 10 villages and picking the accepted actual census figures which form our ancillary variable x_i and obtaining their corresponding monthly household income (our study variable) y_{1i} we generate Table 1. Similarly, a sample of monthly household expenditure y_{2i}, a sample of monthly household phone calls made y_{3i}, a sample of monthly household visitors entertained y_{4i}, and a sample of monthly household mails received y_{5i}, each corresponding to the same ancillary variable x_i are shown in Tables 2, 3, 4, and 5 respectively.

TABLE 1: SAMPLE ON MONTHLY HOUSEHOLD INCOME y_{1i} CORRESPONDING TO x_i

x_i	65	96	108	106	123	68	92	102	118	98
y_{1i}	77	104	111	120	130	86	118	116	125	108

TABLE 2: SAMPLE ON MONTHLY HOUSEHOLD EXPENDITURE y_{2i} CORRESPONDING TO x_i

x_i	65	96	108	106	123	68	92	102	118	98
y_{2i}	76	104	92	108	70	62	116	99	60	100

TABLE 3: SAMPLE ON MONTHLY HOUSEHOLD PHONE CALLS MADE y_{3i} CORRESPONDING TO x_i

x_i	65	96	108	106	123	68	92	102	118	98
y_{3i}	20	18	25	20	30	15	20	28	18	20

TABLE 4: SAMPLE ON MONTHLY HOUSEHOLD VISITORS ENTERTAINED y_{4i} CORRESPONDING TO x_i

x_i	65	96	108	106	123	68	92	102	118	98
y_{4i}	18	18	20	20	19	20	18	21	22	18

TABLE 5: SAMPLE ON MONTHLY HOUSEHOLD MAILS RECEIVED y_{5i} CORRESPONDING TO x_i

x_i	65	96	108	106	123	68	92	102	118	98
y_{5i}	40	30	25	30	48	20	35	32	46	15

All the computations below would be based on these samples.

THE SUPERPOPULATION MODEL AND COMPARISON OF ESTIMATORS

Given that the samples obtained in the Tables above are finite populations drawn from a superpopulation, and let $y_i \ and \ x_i$ denote respectively the values of a characteristic y and the

measure of size x for the i^{th} unit in the population $(i = 1,2,...,N)$. If y is unrelated to x, a reasonable model for the superpopulation assumed by Rao (1966) is

$$y_i = \mu + e_i \qquad 33$$

where $E(e_i \mid x_i) = 0$; $E(e^2{}_i \mid x_i) = a$, $a > 0$; $E(e_i e_j \mid x_i x_j) = 0$; and E denotes the average over all finite populations that can be drawn from the superpopulation.

Applying (33) in making efficiency comparisons between the usual estimators and the alternative estimators of Rao we have

And 34

$$n\left[EV\left(\hat{Y}_c\right)\right] = nV_c = \left(a\sum_i^N \frac{1}{p_i} - aN\right) + \mu^2\left(\sum_i^N \frac{1}{p_i} - N^2\right)$$

$$n\left[EV\left(\hat{Y}_R\right)\right] = nV_R = aN^2 - aN^2\sum_i^N p_i{}^2 \qquad 35$$

$$\Rightarrow nV_c - nV_R = \left(a + \mu^2\right)\left(\sum_i^N \frac{1}{p_i} - N^2\right) + aN^2\left(\sum p_i - \frac{1}{N}\right) \qquad 36$$

Since $\frac{1}{N}\sum p_i > \frac{N}{\sum 1/p_i}$ and $\sum p_i = 1$ *then* $\sum \frac{1}{p_i} > N^2$ *and* $\sum p_i{}^2 > \frac{1}{N}$

$\Rightarrow nV_c > nV_R$.

Following Rao (1966), we then have the variances of the other estimators obtained from $V\left(\hat{Y}_c\right)$.

Since $nV\left(\hat{Y}_c\right) = \sum_i^N \frac{y_i{}^2}{p_i} - Y^2$ 37

Where $Y = \sum_{i=1}^N y_i$, it follows that

$$nV\left(\hat{Y}_R\right) = N^2\left[\sum_i^N y_i p_i - \left(\sum_i^N y_i p_i\right)^2\right]; \qquad 38$$

$$nV\left(\hat{Y}_p\right) = \sum \frac{y^2{}_i p_i}{p^{*2}{}_i} - \left(\sum \frac{y_i p_i}{p^*{}_i}\right)^2; \qquad 39$$

$$nV\left(\hat{Y}_p{}'\right) = \sum \frac{y^2{}_i p_i}{p'^2{}_i} - \left(\sum \frac{y_i p_i}{p'{}_i}\right)^2; \qquad 40$$

After simplifying some algebra Amahia et al (1989) obtained the following result with respect to comparison of the estimators.

$$n\left[EV\left(\hat{Y}_c\right) - EV\left(\hat{Y}_p\right)\right] = aA + \mu^2 B \qquad 41$$

Where $A = \sum \frac{\left(p_i{}^{*2} - p_i{}^2\right)}{p_i{}^{*2}}\left(\frac{1 - p_i}{p_i}\right)$ and

$$B = \sum p_{i\,i}\left(\frac{1}{p_i} - N\right)^2 - \sum p_i\left(\frac{1}{p_i^*} - \sum \frac{p_i}{p^*}\right)^2$$

$= V\left(\frac{1}{p_i}\right) - V\left(\frac{1}{p^*_i}\right)$ where p_i are used as weights for computing the variances. Therefore from the fore going we can obtain $n\left[EV\left(\hat{Y}_p\right)\right]$ as

$$n\left[EV\left(\hat{Y}_p\right)\right] = a\left(\sum \frac{1}{p_i} - N - A\right) + \mu^2\left(\sum \frac{1}{p_i} - N^2 - B\right) \qquad 42$$

Similarly from

$n\left[EV\left(\hat{Y}_c\right) - EV\left(\hat{Y}'_p\right)\right] = aA' + \mu^2 B'$

where

$$A' = \sum \frac{\left(p_i'^2 - p_i^2\right)}{p_i'^2}\left(\frac{1-p_i}{p_i}\right) \quad \text{and} \quad B' = V\left(\frac{1}{p_i}\right) - V\left(\frac{1}{p'_i}\right)$$

we can obtain an expression for $n\left[EV\left(\hat{Y}'_p\right)\right]$ since we have $n\left[EV\left(\hat{Y}'_c\right)\right]$ in (34) i.e.,

$$n\left[EV\left(\hat{Y}'_p\right)\right] = a\left(\sum \frac{1}{p_i} - N - A'\right) + \mu^2\left(\sum \frac{1}{p_i} - N^2 - B'\right) \qquad 43$$

COMPUTATIONS OF THE EXPECTED VARIAVCES, EXACT VARIANCES AND BIASES OF DIFFERENT ESTIMATORS UNDER THE SUPERPOPULATION MODEL OF (32) FOR POPULATIONS I - V

Although 100 different populations were considered ,Table 6 only shows the computed results obtained for the expected variances of different estimators already discussed in Equations 33, 34, 41, and 42 under the superpopulation model for 5 different populations .

Table 6 EXPECTED VARIANCES OF DIFFERENT ESTIMATORS FOR POPULATIONS I-V.

Estimator	$\rho = 0.939$ Pop. I	$\rho = 0.062$ Pop. II	$\rho = 0.558$ Pop. III	$\rho = 0.404$ Pop. IV	$\rho = 0.373$ Pop. V
$n\left[EV\left(\hat{Y}_c\right)\right]$	$94.122a + 4.122\mu^2$	$94.122a + 4.122\mu^2$	94.122a + 4.122 μ^2	94.122a + 4.122 μ^2	94.122a + 4.122 μ^2
$n\left[EV\left(\hat{Y}_R\right)\right]$	$89.661a$	$89.661a$	89.661a	89.661a	89.661a
$n\left[EV\left(\hat{Y}_p\right)\right]$	$93.139a + 3.509\mu^2$	$89.351a + 0.011\mu^2$	89.325a + 1.038 μ^2	88.831a +0.516 μ^2	88.7918a + 0.4376 μ^2
$n\left[EV\left(\hat{Y}'_p\right)\right]$	$93.633a + 3.635\mu^2$	$89.717a + 0.016\mu^2$	91.215a + 1.281 μ^2	90.55a +0.671 μ^2	90.4401a + 0.5735 μ^2

Considering the exact variance and bias of each of the estimators under the 100 populations, we summarize the computed results in Table 7 for only the 5 populations.

TABLE 7: EXACT VARIANCE AND BIAS OF THE ESTIMATORS UNDER POPULATIONS I -V

Estimator	Population I		Population II		Population III		Population IV		Population V	
	n(variance)	$\lvert Bias \rvert$	n(variance)	$\lvert Bias \rvert$	n(variance)	$\lvert Bias \rvert$	n(variance)	$\lvert Bias \rvert$	n(variance)	$\lvert Bias \rvert$
$\hat{Y}_c$	6025.196	0	53763.68	0	1615	0	1314	0	10628	0
$\hat{Y}_R$	20991.137	27.510	36286.42	2.180	2118.92	4.627	190.10	1.010	10571.915	6.930
$\hat{Y}_p$	4718.188	0.490	36588.71	0.639	1375.79	0.215	226.02	1.068	8856.136	1.380
$\hat{Y}_p^1$	4836.336	1.678	36146.64	2.045	1425.40	2.045	276.84	0.602	9339.37	4.345

From Table 7, one can observe that $\hat{Y}_P$ beats all other estimators in populations I, III and V, since its variance in each case has a minimum value. However, $\hat{Y}_R$ fares better than all others in population IV whereas $\hat{Y}_P^1$ fares better than the rest in population II where ρ =0.062. The Table also depicts that $\hat{Y}_C$, $\hat{Y}_P$ *and* $\hat{Y}_P^1$ are better estimators than $\hat{Y}_R$ when the study variable has a high and positive correlation (ρ ≥ 0.5) with the ancillary variable. On the other hand, the table also shows that $\hat{Y}_R$, $\hat{Y}_P$ *and* $\hat{Y}_P^1$ are better than $\hat{Y}_C$ when ρ< 0.5

CONCLUSION

From the fore-going and within the limitation of the superpopulation model as assumed by Rao (1966), one can empirically conclude in multiple character survey, under PPS sampling with replacement, that if the correlation between each of the study variables and the ancillary variable is not in the extreme case of either 0 or 1

then : (1) $\hat{Y}_C$, $\hat{Y}_P$ *and* $\hat{Y}_P$, are better estimators than $\hat{Y}_R$ when the study variable has a high and positive correlation (ρ ≥0.5) with the ancillary variable;

(2) $\hat{Y}_R$, $\hat{Y}_P$ *and* $\hat{Y}_P^1$ are better estimators than $\hat{Y}_C$ when the study variable has a poor and positive correlation (ρ < 0.5) with the ancillary variable; and

(3) $\hat{Y}_R$ *and* $\hat{Y}_P$ are better estimators when the correlation coefficient is neither too high as in Population I nor too poor as in Population II.

REFERENCES

Amahia, G. N., Y. P.Chaubey, et al., 1989. Efficiency of a new estimator in PPS Sampling for multiple characteristics. J. Stat. Plan. Inf.erence. 21, 75-84,

Bansal, M. L. and Singh, R., 1985. An alternative estimator for multiple characteristics in PPS sampling. J. Stat. Plan. Inf. 11, 313-320,

Cochran, W. G., 1946. Sampling Theory when the sampling units are of unequal sizes .J. Amer. Stat. Assoc.,37,199-212.

Cochran, W. G., 1977. Sampling Techniques. 3rd ed.,New Delhi; Willey Eastern Ltd.,ISBN

Grundy, P. M., 1954. A method of Sampling with probability exactly proportional to size. J.Roy. Stat. Soc. B16, 236-238

Hansen, M. H. and Hurwitz, W. N., 1943. On the theory of sampling from finite populations. Ann. Math. Stat. 14, 333-362

Lahiri, D. B., 1951. A method for sample selection providing unbiased ratio estimators. Bull. Int. Stat. Inst., 33, 2,133-140.

Rao, J. N. K., 1966. Alternative estimators in PPS sampling for multiple characteristics. Sankhya , A 23, 47-60. Sankhya.