GLOBAL JOURNAL OF MATHEMAT.CAL SCIENNCES VOL. 6, N0. 2, 2007: 75 - 80
COPYRIGHT (C) BACHUDO SCIENCE CO. LTD. PRINTED IN NIGERIA. ISSN 1596-6208

THE UNIVARIATE DIFFERENCE ESTIMATOR OF THE POPULATION TOTAL UNDER DOUBLE SAMPLING FOR INCLUSION PROBABILITIES WHEN THE TOTAL OF THE AUXILIARY VARIABLE IS KNOWN

G. A. UDOFIA

(Received 21 November 2006; Revision Accepted 10 July 2007)

ABSTRACT

A double sampling strategy for inclusion probabilities with a difference estimator of the population total that depends on a known population total of an auxiliary variable is proposed. The unbiasedness and sampling variance of the proposed estimator are also determined for a situation where the two-phase samples are independent. The condition under which the proposed estimator can be preferred to the difference estimator which depends on the initial sample total of an auxiliary variable as discussed by Udofia (1995) is specified.

KEY WORDS: Double Sampling for Inclusion Probabilities, Difference Estimator With Population Total of An Auxiliary Variable.

1. INTRODUCTION

Udofia (1995) gives a comprehensive list of different double sampling strategies developed by different researchers for estimation of the total of a random variable Y and then proposes, as an additional contribution, an unbiased difference estimator of the total of Y under double sampling for inclusion probabilities. The paper discusses the proposed estimator in detail and proves that under a certain condition based on the cost per unit of information in both phases of the survey the proposed difference estimator gives a more precise result than a corresponding estimator based on a sample drawn in a single phase.

In the double sampling strategy discussed by Udofia (1995), an initial sample of size n_1 denoted here by $S(n_1)$ is drawn by simple random sampling without replacement (SRSWOR) and information on variables X and Z is obtained from the sample. Information on X is used to calculate the inclusion probability for each element of the pcpulation while the information on Z is used to improve the precision of the proposed estimator. A second sample of size n_2, $n_2 < n_1$, denoted here by $S(n_2)$ is drawn from the initial sample with probability $p_i=x_i/X_1$ proportional to size, X_1, and with replacement (PPSWR) where $X_1 = \sum_{i=1}^{n_1} x_i$. Information on Y, the variable of interest, is obtained from the second sample. The proposed difference estimator of the population total is

$$Y_1 = \frac{N}{n_1}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{y_i}{p_i} - k\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{z_i}{p_i} - Z_1\right)\right] \qquad \ldots(1)$$

where $Z_1 = \sum_{i=1}^{n_1} Z_i$ and k is a suitably chosen constant. The paper [Udofia (1995)] proves that this estimator is unbiased with variance

$$V(\hat{Y}_1) = \frac{N^2(1-f_1)}{n_1}S_y^2 + \frac{N}{N-1}\frac{n_1-1}{n_1 n_2}\left[V_p(y) + k^2V_p(z) - 2k\delta_{yz}\sigma_p(y)\sigma_p(z)\right] \qquad \ldots..(2)$$

where

$$\delta_{yz} = \frac{\operatorname{cov} p(y,z)}{\sigma_p(y)\sigma_p(z)}, \sigma_p(y) = \sqrt{V_p(y)}, \sigma_p(z) = \sqrt{V_p(z)}, \text{ and } f_1 = n_1/N.$$

The paper also shows that the value of k that minimizes $V(\hat{Y})$ is $k^* = \delta_{yz}\frac{\sigma_p(y)}{\sigma_p(z)}$.

An unbiased estimator of $V(\hat{Y}_1)$ which is missing in Udofia (1995) can be obtained as

G. A. Udofia, Department of Mathematics, University of Calabar, Calabar, Nigeria.

$$\hat{V}(\hat{Y}_1) = \frac{N(N-n_1)}{n_1 n_2 (n_1 - 1)}\left\{\sum_{i=1}^{n_2}\frac{y_i^2}{x_i} - \frac{1}{n_1(n_2-1)}\left[\left(\sum_{i=1}^{n_2}\frac{y_i}{p_i}\right)^2 - \sum_{i=1}^{n_2}\frac{y_i^2}{p_i^2}\right]\right\} +$$

$$+\frac{N}{N-1}\frac{n_1-1}{n_1 n_2(n_2-1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i}\right)^2 ;\ d_i = y_i - kz_i \qquad \text{....(3)}$$

In many practical survey situations, the value of the auxiliary variable for individual elements of the population may not be available but the population total of the variable can be easily obtained from official sources. It is therefore of interest to see how the above sampling strategy and all the related calculations can be modified if the initial sample total, Z₁, in equation (1) can be replaced by available population total, $Z_0 = \sum_{i=1}^{N} Z_i$, of the auxiliary variable Z and to determine the gain, if any, that such a modification can provide. This paper is an attempt in this direction.

2. THE UNIVARIATE DIFFERENCE ESTIMATOR WHEN THE POPULATION TOTAL IS KNOWN

We assume that $S(n_2)$ is drawn by PPSWR from $S(n_1)$ and that Z_0 is known before the start of the survey. Thus, unlike in Udofia (1995), information is obtained on X alone from the initial sample and information on Y and Z is obtained from the second sample. Then an estimator of the population total of a random variable Y is

$$\hat{Y}_2 = \frac{N}{n_1}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{y_i}{p_i} - k\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{z_i}{p_i} - \frac{n_1}{N}Z_0\right)\right] \qquad \text{.....(4)}$$

We now prove that this estimator is unbiased for the population total $Y = \sum_{i=1}^{N} Y_i$ and then find its sampling variance.

Now

$$E_2(\hat{Y}_2 \mid n_1) = \frac{N}{n_1}\left(\sum_{i=1}^{n_1} y_i - k\sum_{i=1}^{n_1} Z_i\right) + kZ_0 = N\bar{y}_1 - kN\bar{z}_1 + kN\bar{Z}.$$

Then

$$E(\hat{Y}_2) = E_1 E_2(\hat{Y}_2 \mid n_1) = N\bar{Y} - kN\bar{Z} + kN\bar{Z} = N\bar{Y} = Y$$

where $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$, $\bar{z}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} z_i$, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$.

This completes the proof of the unbiasedness of the estimator, $\hat{Y}_2$.

The sampling variance of the estimator can be obtained from the conditional variance formula [Raj (1956)].

$$V(\hat{Y}_2) = V_1 E_2(\hat{Y}_2 | n_1) + E_1 V_2(\hat{Y}_2 | n_1). \qquad \text{...(5)}$$

Now

$$E_2(\hat{Y}_2 \mid n_1) = N(\bar{y}_1 - k\bar{Z}_1) + kZ_0.$$

Let $d_i = y_i - kz_i$ so that $\bar{d}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} d_i = \bar{y}_1 - k\bar{z}_1$.

Then

$$E_2(\hat{Y}_2 \mid n_1) = N\bar{d}_1 + kZ_0$$

and hence

$$V_1 E_2(\hat{Y}_2 \mid n_1) = V_1(N\bar{d}_1) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)S_d^2 \qquad \text{...(6)}$$

where

$$S_d^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{D})^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i - \bar{Y}) - k(Z_i - \bar{Z})\right]^2$$

$$= S_y^2 + k^2 S_Z^2 - 2k\rho_{yz} S_z S_y. \qquad \text{...(7)}$$

Substitution of (7) in (6) gives the result

$$V_1 E_2(\hat{Y}_2 \mid n_1) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(S_y^2 + k^2 S_z^2 - 2k\rho_{yz} S_z S_y\right) \qquad \text{...(8)}$$

For the second term on the right-hand-side of equation (5), we obtain

$$V_2(\hat{Y}_2 \mid n_1) = \frac{1}{n_2}\sum_{i=1}^{n_1} p_i\left[\frac{N}{n_1}\left(\frac{y_i - kZ_i}{p_i} + \frac{n_1}{N}kZ_0\right) - \frac{N}{n_1}\sum_{i=1}^{n_1} p_i\left(\frac{y_i - kZ_i}{p_i} + \frac{n_1}{N}kZ_0\right)\right]^2,$$

$$= \frac{N^2}{n_1^2}\frac{1}{n_2}\sum_{i=1}^{n_1} p_i\left(\frac{d_i}{p_i} - d_{n_1}\right)^2; d_{n_1} = \sum_{i=1}^{n_1} d_i .$$

From Raj (1968) this can be expressed as

$$V_2(\hat{Y}_2 \mid n_1) = \frac{N^2}{n_1^2}\frac{1}{n_2}\sum_{i=1}^{n_1}\sum_{j>i}^{n_1} x_i x_j\left(\frac{d_i}{x_i} - \frac{d_j}{x_j}\right)^2 .$$

Thus

$$E_1 V_2(\hat{Y}_2 \mid n_1) = \frac{N^2}{n_1^2}\frac{1}{n_2}\frac{n_1(n_1-1)}{N(N-1)}\sum_{i=1}^{n_1}\sum_{j>i}^{n_1} x_i x_j\left(\frac{d_i}{x_i} - \frac{d_j}{x_j}\right)^2 ,$$

$$= \frac{N}{N-1}\frac{n_1 - 1}{n_1 n_2}V_p(d),$$

$$= \frac{N}{N-1}\frac{n_1 - 1}{n_1 n_2}\left[V_p(y) + k^2 V_p(z) - 2k\delta_{yz}\sigma_p(z)\sigma_p(y)\right]. \qquad \text{....(9)}$$

Substitution of (8) and (9) in (5) gives the result

$$V(\hat{Y}_2) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(S_y^2 + k^2 S_z^2 - 2k\rho_{yz} S_z S_y\right) + \frac{N}{N-1}\frac{n_1 - 1}{n_1 n_2}\left[V_p(y) + k^2 V_p(z) - 2k\delta_{yz}\sigma_p(z)\sigma_p(y)\right] \qquad \text{... (10)}$$

which is the sampling variance of $\hat{Y}_2$. To obtain an unbiased estimator of $V(\hat{Y}_2)$ in Equation (10), we only need to replace V₁ E₂ ($\hat{Y}_2 \mid n_1$) and E₁ V₂ ($\hat{Y}_2$ |n₁) in equation (5) by their unbiased estimators.

Now an unbiased estimator of $V_1 E_2(\hat{Y}_2 \mid n_1)$ is given by

$$\hat{V}_1 E_2(\hat{Y}_2 \mid n_1) = N_2\left(\frac{1}{n_1} - \frac{1}{N}\right)s_d^2$$

where $s_d^2 = \dfrac{1}{n_1 - 1}\left[\sum_{i=1}^{n_1} d_i^2 - \dfrac{1}{n_1}\left(\sum_{i=1}^{n_1} d_i\right)^2\right]$ is calculated from the initial sample information.

Under the sample design under discussion, the information needed to calculate d$_i$ was not obtained from S(n₁) and hence an unbiased estimator of S_d^2 is obtained from the second sample [Raj (1968)] as

$$s_d^2 = \frac{1}{n_1 - 1}\left\{X_1\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{x_i} - \frac{X_1^2}{n_2(n_2-1)}\left[\left(\sum_{i=1}^{n_2}\frac{d_i}{x_i}\right)^2 - \sum_{i=1}^{n_2}\frac{d_i^2}{x_i^2}\right]\right\}$$

Thus an unbiased estimator of $V_1 E_2(\hat{Y}_2 \mid n_1)$ from the second sample information is given by

$$\frac{N(N-n_1)}{n_1 n_2(n_1 - 1)}\left\{X_1\sum_{i=1}^{n_2}\frac{d_i^2}{x_i} - \frac{X_1^2}{n_2 - 1}\left[\left(\sum_{i=1}^{n_2}\frac{d_i}{x_i}\right)^2 - \sum_{i=1}^{n_2}\frac{d_i^2}{x_i^2}\right]\right\} \qquad \text{....(11)}$$

An unbiased estimator of $E_1 V_2(\hat{Y}_2 \mid n_1)$ is also obtained from the second sample information as

$$\frac{N^2 X_1^2}{n_1^2 n_2(n_2 - 1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{x_i} - \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{x_i}\right)^2 \qquad \text{....(12)}$$

Substitution of (11) and (12) for their respective estimands in (5) gives the result

$$\hat{V}(\hat{Y}_2) = \frac{N(N-n_1)}{n_1 n_2 (n_1 - 1)}\left\{X_1 \sum_{i=1}^{n_2} \frac{d_i^2}{x_i} - \frac{X_1^2}{n_2 - 1}\left[\left(\sum_{i=1}^{n_2} \frac{d_i}{x_i}\right)^2 - \sum_{i=1}^{n_2} \frac{d_i^2}{x_i^2}\right]\right\} + \frac{N^2 X_1^2}{n_1^2 n_2 (n_2 - 1)} \sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{d_i}{x_i}\right)^2 .$$

VARIANCE OF $\hat{Y}_2$ FOR INDEPENDENT SAMPLES

If the two samples, S(n$_1$) and (Sn$_2$), are drawn independently of each other, the estimator, $\hat{Y}_2$, is still unbiased for Y. To verify the unbiasedness of $\hat{Y}_2$ under this condition, we rewrite equation (4) as

$$\hat{Y}_2 = \left(\frac{N}{n_1} X_1\right)\left(\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{y_i}{x_i} - \frac{k}{n_2}\sum_{i=1}^{n_2} \frac{Z_i}{x_i}\right) + kZ_0$$

Let

$$\hat{X} = \frac{N}{n_1} X_1 = N\bar{x}_1, \hat{Z} = N\bar{z}_1$$

$$\hat{R} = \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{y_i}{x_i} - \frac{k}{n_2}\sum_{i=1}^{n_2} \frac{Z_i}{x_i} = \frac{1}{X}\left(\frac{X}{n_2}\sum_{i=1}^{n_2} \frac{y_i}{x_i} - \frac{k}{n_2} X \sum_{i=1}^{n_2} \frac{Z_i}{x_i}\right),$$

$$= \frac{1}{X}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{y_i}{p_i} - \frac{k}{n_2}\sum_{i=1}^{n_2} \frac{Z_i}{p_i}\right).$$

Now

$$E(\hat{X}) = \sum_{i=1}^{N} X_i = X . \quad ..... (13)$$

$$E(\hat{R}) = \frac{1}{X}\sum_{i=1}^{N} p_i \left(\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{y_i}{p_i} - \frac{k}{n_2}\sum_{i=1}^{n_2} \frac{Z_i}{p_i}\right) = \frac{1}{X}(Y - kZ),$$

$$= R = \frac{D}{X}; D = Y - kZ, \quad ......(14)$$

where $Y = \sum_{i=1}^{N} Y_i$, $Z = \sum_{i=1}^{N} Z_i$.

Then

$$\hat{Y}_2 = \hat{X}\hat{R} + kZ_0,$$

and

$$E(\hat{Y}_2) = E(\hat{X})E(\hat{R}) + kZ_0 = X\left[\frac{1}{X}(Y - kZ_0)\right] + kZ_0 = Y$$

This proves that $\hat{Y}_2$ is an unbiased estimator of Y when $S(n_2)$ and S(n$_1$) are independent.

$V(\hat{Y}_2) = V(\hat{X}\hat{R}) + V(kZ_0) + 2Cov(\hat{X}\hat{R}, kZ_0) = V(\hat{X}\hat{R})$

since kZ$_0$ is a constant. Raj (1968) has given the formula for the variance of a product, $\hat{X}\hat{R}$, when $\hat{X}$ and $\hat{R}$ are independent as

$$V(\hat{X}\hat{R}) = [E(\hat{R})]^2 V(\hat{X}) + [E(\hat{X})]^2 V(\hat{R}) + V(\hat{X})V(\hat{R}) \quad ..(15)$$

Now

$$V(\hat{X}) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right) S_x^2 \quad ....(16)$$

$$\left[E(\hat{R})\right]^2 = R^2 .$$

$$V(\hat{R}) = \frac{1}{X^2} V\left(\frac{1}{n}\sum_{i=1}^{n_2} \frac{d_i}{p_i}\right) = \frac{1}{X^2}\frac{1}{n_2}\sum_{i=1}^{N} p_i\left(\frac{d_i}{p_i} - D\right)^2,$$

where $D = \sum_{i=1}^{N} d_i$. From our earlier discussions, we express this result as

$$V(\hat{R}) = \frac{1}{x^2}\frac{1}{n_2}\sum_{i=1}^{N}\sum_{j>1}^{N} x_i x_j \left(\frac{d_i}{x_i} - \frac{d_j}{x_j}\right)^2 = \frac{1}{X^2}\frac{1}{n_2}V_p(d) \qquad \text{...(17)}$$

Substitution of (13), (14), (16) and (17) in (15) gives the result

$$V(\hat{Y}_2) = R^2N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)S_x^2 + \frac{1}{n_2}V_p(d)\left[1 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\frac{S_x^2}{\overline{X}^2}\right]$$

$$= N^2\left(\frac{1}{n_1} - \frac{1}{n}\right)\left(R_1^2 - 2kR_1R_2 + k^2R_2^2\right)S_x^2 + \frac{1}{n_2}\left[1 + \left(\frac{1}{n_1} - \frac{1}{N}\right)C_x^2\right]\left[V_p(y) + k^2V_p(z) - 2k\delta_{yz}\delta_p(z)\sigma_p(y)\right] \qquad \text{...(18)}$$

where R₁ = Y/X, R₂ = Z/X and $C_x = S_x / \overline{X}$. This result is different from the corresponding result for $\hat{Y}_1$ which appears as equation (28) in Udofia (1995).

From Raj (1968), we obtain the unbiased estimator of $V(\hat{X}\hat{R})$ from the formula

$$\hat{V}(\hat{X}\hat{R}) = \hat{X}^2\hat{V}(\hat{R}) + \hat{R}^2\hat{V}(\hat{X}) - \hat{V}(\hat{X})\hat{V}(\hat{R}) \qquad \text{...(19)}$$

Now

$$\hat{X}^2 = \left(\frac{N}{n_1}X_1\right)^2 = \frac{N^2}{n_1^2}X_1^2;\quad X_1 = \sum_{i=1}^{n} x_i \qquad \text{(20)}$$

$$\hat{V}(\hat{R}) = \frac{1}{X^2 n_2}\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i}\right)^2 \qquad \text{...(21)}$$

$$\text{and } \hat{R}^2 = \left(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i}\right)^2,\ R = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i} \qquad \text{...(22)}$$

$$\hat{V}\left(\hat{X}\right) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)s_{1x}^2;\quad s_{1x}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 \qquad \text{...(23)}$$

Substitution of (20), (21), (22) and (23) in (19) gives the result

$$\hat{V}(\hat{Y}_2) = \frac{\bar{x}_1^2}{\overline{X}^2 n_2(n_2 - 1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \hat{R}\right)^2 + N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)\left\{\left[\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i}\right)^2 - \frac{1}{X^2 n_2(n_2 - 1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \hat{R}\right)^2\right]s_{1x}^2\right\} \qquad \text{...(24)}$$

The corresponding estimator for the variance of $\hat{Y}_1$ under the same condition is given by Udofia (1995) as

$$\hat{V}\left(\hat{Y}_1\right) = \frac{\bar{x}_1^2}{\overline{X}^2 n_2(n_2 - 1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \hat{R}\right) + N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)\left\{\left[\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{d_i}{p_i}\right)^2 - \frac{1}{X^2 n_2(n_2 - 1)}\sum_{i=1}^{n_2}\left(\frac{d_i}{p_i} - \hat{R}\right)^2\right]s_{1x}^2 + k^2 s_{1z}^2\right\} \qquad \text{...(25)}$$

A COMPARISON BETWEEN V($\hat{Y}_1$) AND $V(\hat{Y}_2)$

As stated earlier, the estimator $\hat{Y}_1$ discussed by Udofia (1995) requires information on X and Z from the initial sample while the estimator $\hat{Y}_2$ proposed and discussed in this paper requires information on Z from the subsample S(n₂) of S(n₁) and obtains the total of Z for the initial sample from the known population total.

By using (2) and (10) we obtain the following difference between V ($\hat{Y}_1$) and $V(\hat{Y}_2)$:

$$V(\hat{Y}_1) - V(\hat{Y}_2) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)kS_Z(2\rho_{yz}S_x - kS_Z)$$

This shows that $V(\hat{Y}_1) > V(\hat{Y}_2)$ if $2\rho_{yz}S_x > kS_Z$ or

$$\rho_{yz} > \frac{1}{2}k\frac{S_Z}{S_y},\ k > 0$$

Also V($\hat{Y}_1$) = $V(\hat{Y}_2)$ if $\rho_{yz} = \frac{1}{2}k\frac{S_z}{S_x}$. This shows that $\hat{Y}_2$ should be preferred to $\hat{Y}_1$ where there is a high correlation between the auxiliary variable and the study variable and the population total of the auxiliary variable can be obtained. Where these conditions can be satisfied, the collection of information on the auxiliary variable, Z, needed to increase the precision of the estimator can be limited to the second sample. A comparison of $V(\overline{Y}_2)$ in equation (18) to the corresponding expression for $V(\overline{Y}_1)$ in Udofia (1995) under the assumption of independence of S(n₁) and S(n₂) shows that the above inequality also holds.

REFERENCES

Raj, D., 1956. Some estimators in sampling with varying probabilities without replacement. Journ. Amer. Stat. Assoc., 51: 269 – 284.

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

Udofia, G. A., 1995. Methods of differences for global estimation under double sampling for inclusion probabilities Tropical Journal of Applied Sciences, Vo1 5, 56 – 64