ON SAMPLING WITH VARYING PROBABILITIES AND WITH REPLACEMENT IN SUB-SAMPLING DESIGNS

By J. N. K. Rao

Iowa State University

(Received on 26 April, 1961)

1. Introduction

In sub-sampling designs, it is well known that selection of primary units with varying probabilities (*e.g.*, probability proportional to size) often leads to more efficient estimates than selection with equal probabilities. Due to difficulties in the theory of sampling with varying probabilities and without replacement, it is usual practice to select the primaries with replacement and with varying probabilities. This leads to three different methods of selecting the secondaries. In method 1 (Sukhatme, 1954), if the i-th primary is selected λ_i times, $m_i\lambda_i$ secondaries are selected without replacement and with equal probabilities from the i-th primary. In method 2 (Cochran, 1953), if the primary is selected λ_i times, λ_i sub-samples of size m_i are independently drawn without replacement and equal probabilities from the i-th primary, each sub-sample being replaced after it is drawn. In method 3 (Hartley, 1954; Des Raj, 1954) when i-th primary is selected λ_i times, a fixed size of m_i secondaries is drawn from the i-th primary with equal probabilities and without replacement and the estimate from the i-th primary is weighted by λ_i. It was shown (*e.g.*, Des Raj, 1954) that method 1 has smaller variance than method 2 and method 2 has smaller variance than method 3. But, if we assume that the expected cost in a primary is proportional to the expected sub-sample size from the primary, then the three methods have different expected costs. Therefore it would appear more reasonable to compare the efficiency of the three methods for the same expected cost or expected sample size. Here a comparison of the variances for the three methods has been made for the same expected sample size but interestingly the conclusions remain the same regarding efficiency.

2. Variance Formulæ

Let y_{ij} denote the value of j-th secondary in the i-th primary and let M_i be the number of secondaries in the i-th primary. Let P_i be the

probability of selecting the i-th primary. Let Y_i denote the total for the i-th primary and Y be the population total. From the N primaries in the population, a sample of n primaries is selected with probabilities P_i and with replacement.

Method 1.—The estimate of the total is

$$\hat{Y}_1 = \frac{1}{n}\sum_{i=1}^{N} \lambda_i \frac{M_i}{P_i} \bar{y}_{i, m_i\lambda_i} \tag{1}$$

where $\bar{y}_{i, m_i\lambda_i}$ is the mean of the $m_i\lambda_i$ units from the i-th primary. It is assumed for the largest value of λ_i (namely n) that $m_i\lambda_i \leq M_i$.

$$\text{Var}\,(\hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{M_i^2}{P_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$- \frac{(n-1)}{n}\sum_{i=1}^{N} M_i S_i^2 \tag{2}$$

where S_i^2 is the mean square for the i-th primary.

Method 2.—The estimate of the total is

$$\hat{Y}_2 = \frac{1}{n}\sum^{(n)} \frac{M_i}{P_i} \bar{y}_{i, m_i} \tag{3}$$

where the summation is taken over all the n primaries in the sample.

$$\text{Var}\,(\hat{Y}_2) = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{M_i^2}{P_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2. \tag{4}$$

Method 3.—The estimate of the total is

$$\hat{Y}_3 = \frac{1}{n}\sum_{i=1}^{N} \lambda_i \frac{M_i}{P_i} \bar{y}_{i, m_i} \tag{5}$$

$$\text{Var}\,(\hat{Y}_3) = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \sum_{i=1}^{N}\left(\frac{1}{nP_i} + \frac{(n-1)}{n}\right)$$

$$\times M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2. \tag{6}$$

It is easily seen that

$$\text{Var}(\hat{Y}_1) < \text{Var}(\hat{Y}_2) < \text{Var}(\hat{Y}_3) \tag{7}$$

3. Efficiency Comparisons for the Same Expected Cost

Method 1 vs. *Method* 2.—In method 1, the expected sample size from the i-th primary is

$$E(m_i\lambda_i) = m_i n P_i \tag{8}$$

since λ_i is a binomial variable. To find the expected sample size from the i-th primary for method 2, we have to evaluate the expected number of distinct units in λ_i sub-samples of size m_i, each sub-sample drawn with equal probability and without replacement and each sub-sample is replaced after it is drawn.

Let Z_{ij} denote the 'indicator variable' for the j-th unit in the i-th primary, defined by

$Z_{ij} = 1$ if j-th unit in the i-th primary is included in the sample

$= 0$ otherwise.

Then, the expected number of distinct units taken from the i-th primary is

$$\sum_{j=1}^{M_i} E(Z_{ij}) = \sum_{j=1}^{M_i} E \cdot E\left(\frac{Z_{ij}}{\lambda_i}\right) \tag{10}$$

where

$E(Z_{ij}/\lambda_i)$ = probability that j-th unit is included at least once in the sample of λ_i sub-samples from the i-th primary

= 1 — (probability that it is not included in any one of the λ_i sub-samples)

$$= 1 - \left(\frac{M_i - m_i}{M_i}\right)^{\lambda_i}. \tag{11}$$

Expanding binomially we get

$$E\left(\frac{Z_{ij}}{\lambda_i}\right) = \frac{m_i\lambda_i}{M_i} - \frac{\lambda_i(\lambda_i - 1)}{2}\frac{m_i^2}{M_i^2} \tag{12}$$

neglecting higher terms

$$\therefore \sum_{j=1}^{M_i} E(Z_{ij}) = \sum_{j=1}^{M_i} \left\{ \frac{m_i n P_i}{M_i} - \frac{m_i^2}{M_i^2} \frac{n(n-1)P_i^2}{2} \right\}$$

$$= m_i n P_i \left[1 - \frac{(n-1)P_i}{2} \cdot \frac{m_i}{M_i} \right] \tag{13}$$

since

$$E(\lambda_i) = nP_i \quad \text{and} \quad E\{\lambda_i(\lambda_i - 1)\} = n(n-1)P_i^2. \tag{14}$$

In order to make the expected sample sizes equal, we still select m_i units in method 2, but in method 1 we select $m_i^* \lambda_i$ units from the i-th primary if it is selected λ_i times where m_i^* is given by

$$m_i^* n P_i = m_i n P_i \left[1 - \frac{(n-1)P_i}{2} \frac{m_i}{M_i} \right] \tag{15}$$

or

$$\frac{1}{m_i^*} \doteq \frac{1}{m_i} \left[1 + \frac{(n-1)P_i}{2} \frac{m_i}{M_i} \right] \tag{16}$$

neglecting terms involving higher powers of P_i. Now,

$$V(\hat{Y}_i) = \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y \right)^2 + \frac{1}{n} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i^*} - \frac{1}{M_i} \right) S_i^2$$

$$- \frac{(n-1)}{n} \sum_{i=1}^{N} M_i S_i^2 \tag{17}$$

substituting for $1/m_i^*$ from (16) in (17), we get

$$V(\hat{Y}_i) = \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y \right)^2 + \frac{1}{n} \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2$$

$$- \frac{(n-1)}{2n} \sum_{i=1}^{N} M_i S_i^2 \tag{18}$$

$$= V(\hat{Y}_2) - \frac{(n-1)}{2n} \sum_{i=1}^{N} M_i S_i^2 \tag{19}$$

i.e.,

$$V(\hat{Y}_1) < V(\hat{Y}_2).$$

Method 1 vs. *Method* 3.—To obtain the expected sample size from the i-th primary for method (3), we have to find the probability that the i-th primary is included at least once in the sample. It is easily seen that this probability is equal to $1 - (1 - P_i)^n$.

$\therefore$ Expected sample size from the i-th primary for method 3

$$= m_i [1 - (1 - P_i)^n]$$

$$= m_i n P_i \left[1 - \frac{(n-1)}{2} P_i\right]. \tag{20}$$

Therefore, for method 1, $m_i^* \lambda_i$ units are selected from the i-th primary if the i-th primary is selected λ_i times, where m_i^* is given by

$$m_i^* n P_i = m_i n P_i \left[1 - \frac{(n-1)}{2} P_i\right] \tag{21}$$

or

$$\frac{1}{m_i^*} = \frac{1}{m_i}\left(1 + \frac{(n-1)}{2} P_i\right) \tag{22}$$

whilst for method 3, m_i units are still selected. Substituting for $1/m_i^*$ from (22) in (17), we obtain

$$V(\hat{Y}_1) = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \frac{1}{n}\sum_{i=1}^{N} \frac{M_i^2}{P_i}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$+ \frac{(n-1)}{2n}\sum_{i=1}^{N} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$- \frac{(n-1)}{2n}\sum_{i=1}^{N} M_i S_i^2 \tag{23}$$

$$= V(\hat{Y}_3) - \frac{(n-1)}{2n}\sum_{i=1}^{N} \frac{M_i^2}{m_i} S_i^2 \tag{24}$$

i.e.,

$$V(\hat{Y}_1) < V(\hat{Y}_3).$$

Method 2 vs. *Method* 3.—We have already found that expected sample size from the i-th primary for method 2 is given by (13) and for method 3 by (20). Therefore, to make the expected sample sizes equal, we select m_i^* units from the i-th primary for method 3 where m_i^* is given by

$$m_i^* nP_i\left(1-\frac{(n-1)}{2}P_i\right)=m_i nP_i\left(1-\frac{(n-1)P_i}{2}\frac{m_i}{M_i}\right) \quad (25)$$

whilst for method 2, m_i units are still selected. So,

$$m_i^*=m_i\left(1-\frac{(n-1)P_i}{2}\cdot\frac{m_i}{M_i}\right)\left(1+\frac{(n-1)}{2}P_i\right)$$

or

$$\frac{1}{m_i^*}=\frac{1}{m_i}\left(1+\frac{(n-1)P_i}{2}\frac{m_i}{M_i}\right)\left(1-\frac{(n-1)}{2}P_i\right)$$

$$=\frac{1}{m_i}\left(1-\frac{(n-1)}{2}P_i+\frac{(n-1)P_i}{2}\cdot\frac{m_i}{M_i}\right). \quad (26)$$

Now,

$$V(\hat{Y}_3)=\frac{1}{n}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-y\right)^2+\sum_{i=1}^{N}\left(\frac{1}{nP_i}+\frac{(n-1)}{n}\right)$$

$$\times M_i^2\left(\frac{1}{m_i^*}-\frac{1}{M_i}\right)S_i^2. \quad (27)$$

Substituting for $1/m_i^*$ from (26) in (27), we get

$$V(\hat{Y}_3)=V(\hat{Y}_2)+\frac{(n-1)}{2n}\sum_{i=1}^{N}M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)$$

$$\times S_i^2\,[1-(n-1)P_i]. \quad (28)$$

Usually $nP_i\leqslant 1$, unless some of the P_i are large. So, if we restrict to such sizes for which $nP_i\leqslant 1$, then $V(\hat{Y}_3)>V(\hat{Y}_2)$. Even if some of the P_i are such that $nP_i>1$, the second term in the r.h.s. of (28) can still be positive so that we expect $V(\hat{Y}_3)>V(\hat{Y}_2)$ in most of the situations.

4. Concluding Remarks

We note that the between-primary variation component for all three methods is the same. So, if the between-primary component is

large compared to the within-primary component, then all the three methods will have approximately the same variance. But, cases can arise where the within-primary component is of the same order as the between-primary component in which case the decision among the three methods may be important. From the efficiency point of view, we have seen that method 1 is best of all the three methods and that method 2 is better than method 3. But, method 2 has an advantage with regard to the simplicity of the estimate of the variance which is simply the mean square between sample totals of the primaries included in the sample, whereas for methods 1 and 3, the estimate of variance involves within-primary components also. Method 3 has an advantage over methods 1 and 2 regarding the fixed sample size from each primary. In methods 1 and 2, the size of the sample from each primary is a random variable, so that method 3 has an advantage with regard to the optimum sample size from a primary. For with method 3 the actual (fixed) sub-sample size can be equated to the optimum value, whilst with methods 1 and 2, only expected sub-sample size can be equated to the optimum value.

5. Acknowledgment

The author expresses his sincere thanks to Dr. H. O. Hartley for helpful discussions on this topic.

6. References

Cochran, W. G. .. *Sampling Techniques*, John Wiley & Sons, Inc., 1953, p. 252.

Des Raj .. "On sampling with varying probabilities in multi-stage designs," *Ganita*, 1954, **5**, 45-51.

Hartley, H. O. .. Unpublished lecture notes on Theory of Advanced Design of Surveys, Statistical Laboratory, Iowa State University, Ames, 1954.

Sukhatme, P. V. .. *Sampling Theory of Surveys with Applications*, Ind. Soc. Agric. Stats., New Dehli and Iowa State College Press, Ames, Iowa, 1954, p. 379.