A NOTE ON SAMPLE DESIGN OF THE COMBINED SURVEY OF COCONUT AND ARECANUT IN INDIA

BY K. V. R. SASTRY

F.A.O. Regional Coconut Statistician for Asia and Far East

SEVERAL States in India have initiated dual purpose sample surveys with the object of simultaneously estimating the area and production of coconut and arecanut crops. These two perennial crops are often grown under similar conditions and therefore a certain amount of economy is expected from a dual purpose survey where common field personnel are used for enumerating both crops. The primary units of sampling are villages which are geographically stratified under administrative divisions of the State such as districts, taluks or taluk groups. Villages in each stratum are further categorised under the three following groups according to the reported presence of either one of both crops.

Category 1.—Consisting all villages reportedly growing both crops.

Category 2.—Consisting of villages reported growing coconut alone, and

Category 3.—Consisting of villages reported growing arecanut alone.

This categorisation thus leads to a two-way stratification. The total number of sample villages for both crops is fixcd with reference to the available strength of field staff and a reasonably satisfactory level of precision for the results. The allocation between different strata is proportional to the respective area under the two crops together. According to this procedure, strata belonging to category 1 seem to be receiving an excessive representation with respect to either of the crops compared to strata belonging to the other category. Even the relative allocation of sample villages among strata belonging to the first category will be satisfactory only if there is perfect correlation between the stratum areas under coconut and arecanut. The present note describes a more reasonable approach to the problem of allocation.

2. For simplicity, let us consider the problem of estimating correct area under each crop based on a single stage design with villages as sampling units. Let N_{hi} be the total number of villages in the h-th geographical stratum belonging to the i-th category. ($h = 1, 2, \ldots, t$; $i = 1, 2, 3$) so that

$$\left.\begin{aligned} N_{h.} &= \sum_i N_{hi} \\ N_{.i} &= \sum_h N_{hi} \\ \text{and} \quad N_{..} &= \sum_h N_{h.} = \sum_i N_{.i} \end{aligned}\right\} \qquad (1)$$

respectively denote the total number of villages in the h-th geographical stratum, i-th category (crop-stratum) and the entire population.

3. Let us first ignore geographical stratification and decide the sample allocation between the three main crop-strata. Let c_1 denote the cost of enumerating a village of the first category for both crops, c_2 denote the cost of enumerating a village of the second category for coconut area alone and c_3 denote the cost of enumerating a village of the third category for arecanut area alone. Then the varying component C of the total cost can be expressed as:

$$C = c_1 n_{.1} + c_2 n_{.2} + c_3 n_{.3} \qquad (2)$$

where $n_{.i}$ denotes the number of sample villages selected from the i-th crop-stratum. The variances of the estimated coconut area and arecanut area are respectively given by

$$V_c = \frac{N_{.1}^2 \sigma_{.1}^2}{n_{.1}} + \frac{N_{.2}^2 \sigma_{.2}^2}{n_{.2}} \qquad (3)$$

and

$$V_a = \frac{N_{.1}^2 \delta_{.1}^2}{n_{.1}} + \frac{N_{.3}^2 \delta_{.3}^2}{n_{.3}} \qquad (4)$$

where $\sigma_{.1}$ and $\delta_{.1}$ are the standard deviations of the coconut and arecanut areas between villages of the first category, $\sigma_{.2}$ is the standard deviation of the coconut area between villages of the second category and $\delta_{.3}$ is the standard deviation of the arecanut area between villages of the third category. The general problem of allocation is one of choosing $n_{.1}$, $n_{.2}$ and $n_{.3}$ for a fixed value of $C = C_0$ in such a way that V_c and V_a assume simultaneously possible minimum values. This is mathematically equivalent to minimizing with reference to $n_{.1}$, $n_{.2}$ and $n_{.3}$

$$\phi_1 = \frac{N_{.1}^2\sigma_{.1}^2}{n_{.1}} + \frac{N_{.2}\sigma_{.2}^2}{n_{.2}} + \lambda(c_1 n_{.1} + c_2 n_{.2} + c_3 n_{.3} - C_0) \tag{5}$$

and

$$\phi_2 = \frac{N_{.1}^2\delta_{.1}^2}{n_{.1}} + \frac{N_{.3}^2\delta_{.3}^2}{n_{.3}} + \mu(c_1 n_{.1} + c_2 n_{.2} + c_3 n_{.3} - C_0) \tag{6}$$

subject to

$$c_1 n_{.1} + c_2 n_{.2} + c_3 n_{.3} = C_0 \tag{7}$$

where λ and μ are the usual Lagrangean multipliers.

Equating

$$\frac{\delta\phi_1}{\delta n_1}, \quad \frac{\delta\phi_1}{\delta n_2}, \quad \frac{\delta\phi_2}{\delta n_1} \quad \text{and} \quad \frac{\delta\phi_2}{\delta n_3}$$

to zero we get

$$n_{.1} = \frac{N_{.1}\sigma_{.1}}{\sqrt{\lambda c_1}} = \frac{N_{.1}\delta_{.1}}{\sqrt{\mu c_1}} \qquad \therefore \quad \frac{1}{\sqrt{\mu}} = \frac{1}{\sqrt{\lambda}}\frac{\sigma_{.1}}{\delta_{.1}}$$

$$n_{.2} = \frac{N_{.2}\sigma_{.2}}{\sqrt{\lambda c_2}}$$

and

$$n_{.3} = \frac{N_{.3}\delta_{.3}}{\sqrt{\mu c_3}} = \frac{N_{.3}\delta_{.3}}{\sqrt{\lambda c_3}}\frac{\sigma_{.1}}{\delta_{.1}}.$$

Using these values in (7) we have

$$\frac{1}{\sqrt{\lambda}}\left(N_{.1}\sigma_{.1}\sqrt{c_1} + N_{.2}\sigma_{.2}\sqrt{c_2} + N_{.3}\delta_{.3}\frac{\sigma_{.1}}{\delta_{1}}\sqrt{c_3}\right) = C_0$$

or

$$\frac{1}{\sqrt{\lambda}} = \frac{C_0\delta_{.1}}{N_{.1}\sigma_{.1}\delta_{.1}\sqrt{c_1} + N_{.2}\sigma_{.2}\delta_{.1}\sqrt{c_2} + N_{.3}\delta_{.3}\sigma_{.1}\sqrt{c_3}}.$$

Thus the allocation is given by

$$\left.\begin{aligned} n_{.1}c_1 &= C_0 k N_{.1}\sigma_{.1}\delta_{.1} \\ n_{.2}c_2 &= C_0 k N_{.2}\sigma_{.2}\delta_{.2} \\ \text{and} \quad n_{3}c_3 &= C_0 k N_{.3}\sigma_{.3}\delta_{.3} \end{aligned}\right\} \tag{8}$$

where

$$k = \frac{1}{(N_{.1}\sigma_{.1}\delta_{.1}\sqrt{c_1} + N_{.2}\sigma_{.2}\delta_{.1}\sqrt{c_2} + N_{.3}\delta_{.3}\sigma_{.1}\sqrt{c_3})}.$$

In the absence of knowledge about $\sigma_{.1}$ and $\delta_{.1}$, we may assume $\sigma_{.1}$ as proportional to the corresponding average "reported" area under coconut per village and $\delta_{.1}$ as proportional to the corresponding average "reported" area under arecanut per village so that (8) can be rewiitten as

$$\left.\begin{aligned} n_{.1}\sqrt{c_1} &= C_0 k^1 C_{.1} A_{.1} \\ n_{.2}\sqrt{c_2} &= C_0 k^1 C_{.2} A_{.1} \\ n_{.3}\sqrt{c_3} &= C_0 k^1 C_{.1} A_{.3} \end{aligned}\right\} \qquad (9)$$

where

$$k^1 = \frac{1}{(C_{.1}A_{.1}\sqrt{c_1} + C_{.2}A_{.1}\sqrt{c_2} + C_{.1}A_{.3}\sqrt{c_3})}$$

with $C_{.i}$ and $A_{.i}$ denoting the "reported" coconut and arecanut areas in the i-th crop-stratum.

4. Sometimes consideration of relative costs are not relevant and in such cases we often have a fixed total sample size n_0. This is equivalent to assuming $c_1 = c_2 = c_3 = c$ in the condition at (7) so that it can be written as

$$n_{.1} + n_{.2} + n_{.3} = \frac{C_0}{c} = n_0. \qquad (10)$$

The allocation for this case can be derived from (9) as below:

$$\left.\begin{aligned} n_{.1} &= n_0 L C_{.1} A_{.1} \\ n_{.2} &= n_0 L C_{.2} A_{.1} \\ \text{and} \quad n_{.3} &= n_0 L C_{.1} A_{.3} \end{aligned}\right\} \qquad (11)$$

where

$$L = \frac{1}{(C_{.1}A_{.1} + C_{.2}A_{.1} + C_{.1}A_{.3})}.$$

5. Having thus determined the sample sizes from each main crop-stratum, we can very often gain further improvements in precision or operational simplification by adopting a proper allocation of the samples fixed for each crop-stratum, between different geographical

strata. Let n_{hi} denote the number of sample villages belonging to the h-th geographical stratum and i-th crop-stratum so that $\sum_h n_{hi} = n_{.i}$.

6. Let C_{hi} and A_{hi} denote the "reported" areas under coconut and arecanut for all villages coming under the h-th geographical stratum and i-th crop-stratum. Clearly $C_{h3} = A_{h2} = 0$

$$\sum_h C_{hi} = C_{.i} \quad \text{and} \quad \sum_h A_{hi} = A_{.i}.$$

In so far as the second crop-stratum which reportedly contains villages with coconut alone, it can be easily shown that an efficient allocation between the geographical strata is approximately given by

$$n_{h2} = n_{.2} \frac{C_{h2}}{C_{.2}}. \tag{12}$$

Similarly for the third crop-stratum

$$n_{h3} = n_{.3} \frac{A_{h3}}{A_{.3}}. \tag{13}$$

In the first crop-stratum, where the presence of both crops is reported, the choice of n_{h1}, which leads to the minimum possible sampling variance in the estimation of coconut area, is given by

$$n_{h1} = n_{.1} \frac{N_{h1}\sigma_{h1}}{\sum_h (N_{h1}\sigma_{h1})} \tag{14}$$

the corresponding variance being given by

$$v_{c0} = \frac{1}{n_{.1}} (\Sigma N_{h1}\sigma_{h1})^2 \tag{15}$$

where σ_{h1} is the standard deviation of the coconut area between villages in the h-th geographical stratum of the first crop-stratum. Similarly the optimum choice of n_{h1} and corresponding minimum variance in the estimation of arecanut area are given by

$$n_{h1} = n_{.1} \frac{N_{h1}\delta_{h1}}{\sum_h (N_{h1}\delta_{h1})} \tag{16}$$

and

$$v_{a0} = \frac{1}{n_{.1}} (\Sigma N_{h1}\delta_{h1})^2 \tag{17}$$

where δ_{h1} has the same meaning with arecanut as σ_{h1} has for coconut.

7. The allocations given by (14) and (16) will be identical when

$$\frac{\sigma_{h1}}{\delta_{h1}} = \text{constant} \tag{18}$$

under certain assumptions, this condition can be shown as equivalent to :

$$\frac{C_{h1}}{A_{h1}} = \text{constant.} \tag{19}$$

This condition will thus mean perfect correlation between the reported coconut areas and arecanut areas of the geographical strata where the presence of both crops is reported. In practice, we cannot however expect such a perfect correlation and therefore if we choose one of the two allocations (14) or (16), it may sometimes result in a serious loss of precision in estimating area under the other crop. For any given allocation, the sampling variance of the estimated coconut area is given by

$$v_c = \sum_h \frac{N_{h1}{}^2 \sigma_{h1}{}^2}{n_{h1}}.$$

The relative efficiency of the corresponding optimum allocation is given by

$$\frac{v_c}{v_{c0}} = \frac{1}{v_{c0}} \sum \frac{N_{h1}{}^2 \sigma_{h1}{}^2}{n_{h1}}.$$

Similarly the relative efficiency of the corresponding optimum allocation for arecanut is given by

$$\frac{v_a}{v_{a0}} = \frac{1}{v_{a0}} \sum \frac{N_{h1}{}^2 \delta_{h1}{}^2}{n_{h1}}.$$

As a compromise between (14) and (16), we shall choose an allocation which gives a minimum average efficiency of the respective optimum allocation procedures. We shall therefore minimise

$$\phi = \frac{1}{v_{c0}} \sum \frac{N_{h1}{}^2 \sigma_{h1}{}^2}{n_{h1}} + \frac{1}{v_{a0}} \sum \frac{N_{h1}{}^2 \delta_{h1}{}^2}{n_{h1}} \tag{20}$$

subject to

$$n_{.1} = \sum_h n_{h1} \tag{21}$$

Adopting the usual method of Lagrangean multipliers, it can be easily shown that

$$n_{h1} \propto \sqrt{\frac{N_{h1}{}^2 \sigma_{h1}{}^2}{v_{c0}} + \frac{N_{h1}{}^2 \delta_{h1}{}^2}{v_{a0}}}.$$

If we once again assume σ_{h1} and δ_{h1} as proportional to the respective average "reported" coconut area per village and arecanut area per village, the above solution can be simplified as

$$n_{h1} = n_{.1} \frac{\sqrt{\frac{C_{h1}^2}{C_{.1}^2} + \frac{A_{h1}^2}{A_{.1}^2}}}{\sum \sqrt{\frac{C_{h1}^2}{C_{.1}^2} + \frac{A_{h1}^2}{A_{.1}^2}}}. \qquad (22)$$

8. After fixing $n_{.1}$, $n_{.2}$ and $n_{.3}$ with reference to a fixed total cost or total sample size, another alternative allocation may be obtained as follows: the total number of villages to be enumerated for coconut area is $n_{.1} + n_{.2}$ while that for the arecanut area will be $n_{.1} + n_{.3}$. These numbers can then be split up between the various substrata in proportion to the "reported" coconut and arecanut areas respectively. With reference to substrata of the second and third categories, the procedure will lead to a single set of numbers which represent the numbers of villages to be selected and enumerated for either of the two crops. However, for the substrata of the first category, we shall get one set of numbers representing the allocation for coconut and another set representing the allocation for arecanut. The smaller of these two numbers may be taken as the common sample for both crops, while the difference between this number and the larger number will be the number of villages to be further selected for enumerating area of the crop to which the larger number corresponds. In practice, the total number of distinct villages that will be chosen in this manner will be slightly different from $n_{.1} + n_{.2} + n_{.3}$ as initially determined. From empirical considerations, this approach seems to result in greater precision per unit cost or per sample village for either crop. However, the analytical aspect of this approach has still to be investigated.