STUDIES IN SYSTEMATIC SAMPLING FOR TWO DIMENSIONAL FINITE POPULATION WITH SPECIAL REFERENCE TO SURVEY FOR CROP ESTIMATION OF GUAVAS

By

A.H. Manwani* and K.B. Singh
I.A.S.R.I., New Dehli-110 002

(Received : October, 1975)

1. Introduction

In surveys for estimating production of livestock products, fisheries and fruit crops, the population under study is distributed both in space and time. Sampling techniques for estimating the total produce of such commodities necessarily, involve sampling of units in time in addition to sampling in space-though one may not always be interested in estimating the production per unit of time as such. It will be of interest to examine the problems connected with the determination of sampling units, stratification and other sampling design problems which one may come across in designing sample surveys involving sampling over time. In the present paper, an attempt has been made to study the choice of various sampling units in time and compare their efficiencies, taking into view, the convenience of the field work operation in relation to conduct of sample surveys for cr op estimation of guavas, one of the important fruit crops grown in Northern India. It has been reported that on an average for the district as a whole, harvesting in a guava orchard lasts for about 80 days [1]. The entire produce of a tree weighing on an average about 18 kg. or 150 fruits, is generally harvested in 15 pickings. Hence in order to record the entire yields of a tree, enumerator on an average will be required to visit the orchard as many as 15 times resulting in enormous cost of collection of these data. The present study has been undertaken to examine the possibility of recording yields data by visiting orchards regularly at certain fixed intervals, thereby, decreasing cost of the survey. Thus, instead of asking the enumerator to record yield data for all the pickings of the selected trees in the selected orchards, he may be required to visit the

* Present address : Ministry of Industry, Government of India, New Delhi.

orchards at regular intervals. This will increase the scope for allotting him more number of primary sampling units *i.e.* villages, the variation between which forms the major component of variance of the estimate of average yield per tree. Section 2 describes type of data under study. In Section 3, different sampling schemes which could be followed for collection of the data on yield have been proposed. In Section 4, stochastic model appropriate for the study has been given and the relevant formulae corresponding to different sampling schemes which have been studied in the paper have been derived. Section 5 gives the efficiencies of different sampling schemes followed by concluding remarks given in Section 6.

2 Type of Data Under Study

Data for this investigation have been taken from a sample survey for estimating the extent of cultivation and production of guava carried out by the Institute of Agricultural Research Statistics in Allahabad district of Uttar Pradesh. The survey was carried out in three rounds. For the study of yield, a stractified (compact agro-climatic regions formed strata) sample of 50 to 60 villages was selected in the first stage and within each village, a sample of five guava orchards was selected with equal probabilities out of bearing guava orchards in the village. Within each orchard, three clusters of four bearing trees each were selected at random to record the data on yield of guava as and when the fruit was picked throughout the winter harvesting season. For the present study, the data collected during the last round of the survey from 100 orchards selected in 20 villages out of the total of 153 villages in one of the strata have been utilized. In these orchards, it was found that harvesting of guava started as early as first week of November and lasted for 136 days. In a given village, harvesting of guava lasted for minimum of 71 days and a maximum of 130 days, the average duration of harvest in a village being 106 days. In an orchard, average harvesting period was found to be 86 days during which the fruit was picked 23 times on an average. The coefficient of variation between villages in respect of yield was of the order of 35% while that between orchards within villages was 71% and between pickings within orchards 104%. With relatively high amount of variation in the yield of guava over different pickings, the sample survey reasonably envisaged recording of data for all the pickings. However, this resulted into enormous work and inconvenience in carrying out field operations. In this paper, we will examine possibility of recording data for a sample of pickings by visiting the orchards at some suitably chosen intervals and see its effect on the precision of the resultant estimate.

3. Different Sampling Schemes Proposed

Sampling in time in relation to the problem under study, involves selection of days or clusters thereof, on which yield of the selected trees in a village could be recorded. For this purpose, systematic method of sampling could be made use of with an advantage. Generally, the population spread over time presents a well defined trend. In the data under study, the trend of yield over the harvesting period of 136 days was of the form as given in the graph. In populations having trend, intra-class correlation need not continuously decrease with increase in the size of cluster formed by clubbing the consecutive days. It will be of interest to find out intra-class correlations for varying sizes of time clusters so as to suggest optimum size of cluster and also optimum interval for systematic recording of yield data.

For the sake of simplicity, we shall assume that each village consists of a fixed number of orchards and each orchard consists of a fixed number of trees. Also, we assume that the number of trees in any village are not too many as could not be observed for yield in a single day. Let, N and M be number of villages and length of harvesting period (in days) respectively, so that the population spread in space and time consists of NM units with values Y_{ij}'s, the yield obtained from i-th village on j-th day, some Y_{ij}'s may be zero.

Let us suppose that a sample of t units is selected at random out of NM units in the population. This sample may be selected by any one of the following alternative sampling schemes :

(*A*) Without any restriction with simple random sampling without replacement (SRS) out of the totality of NM units in the population;

(*B*) By selecting n_o villages with SRS out of N and recording yield data on all the days so that $n_o = t/M$.

(*C*) By selecting m_o days with SRS out of M and recording yield data from all the N villages in the population on only those selected days, so that $m_o = t/N$.

(*D*) By selecting independently with SRS, n villages out of N and m days out of M with $nm = t$ and recording yield data from the n selected villages on each of the fixed m selected days. The sample of m days out of M should be selected either : (*i*) as a stratified sample with weeks or fortnights taken as strata, or (*ii*) as a systematic sample with an interval of k days between different recordings.

(*E*) By selecting a sample of n villages with SRS out of N and recording yield data on a sample of m randomly selected days within each selected village, the sample of days within each village being selected independently. The random selection of m days should be done either in the form of (*i*) clusters of two or more consecutive days; or (*ii*) a systematic sample with an interval of varying number of k days.

Each of the schemes except (*A*) is feasible from the point of view of organizing field work of the survey. Thus, scheme (*B*) is feasible when the data are to be collected by an *ad-hoc* field staff who may be allotted specified number of villages to observe yield data over the entire period of harvesting of the crop. Scheme (*C*) could be operated by assigning the data collection work as a part-time job to the existing local staff posted in the villages. Scheme (*D*) and (*E*) envisage optimum utilisation of part-time or fulltime *ad-hoc* field staff. The next section describes stochastic model for studying the efficiency of each of the schemes listed above.

4. Stochastic Model for Estimating Mean of a Two Dimensional Population—Efficiencies of Different Sampling Schemes

Let Y_{ij}, the yield from i-th village on j-th day in a given region and a given crop season, be expressed as :

$$Y_{ij}=u+v_i+d_j+e_{ij},\ i=1,2,\ldots\ldots N;\qquad j=1,2,\ldots\ldots M; \tag{4·1}$$

Where,

v_i = fixed average effect of i-th village due to particular soil and agro-climatic factors obtaining in i-th village.

d_j = fixed average effect of j-th day due to market demand and maturity of fruit attained on j-th day.

e_{ij} = random effect due to several causes obtaining in i-th village on j-th day.

Without loss of generality, we may put the restrictions

$\Sigma v_i=0$ and $\Sigma d_j=0$

With these restrictions, we obtain,

$$\begin{aligned} E(\bar{Y}_{..})&=u\\ E(Y_{ij})&=u+v_i+d_j\\ E(\bar{Y}_{i.})&=u+v_i\\ E(\bar{Y}_{.j})&=u+d_j \end{aligned} \quad \ldots(4.2)$$

Where,

$$\bar{Y}_i = \sum_j \frac{Y_{ij}}{M} \quad \text{mean yield per day in } i\text{-th village.}$$

$$\bar{Y}_{\cdot j} = \sum_i \frac{Y_{ij}}{N} = \text{mean yield per village on } j\text{-th day.}$$

$$\bar{Y}_{\cdot\cdot} = \sum_{i,j} \frac{Y_{ij}}{NM} = \text{mean yield per day per village in the region under survey.}$$

We are interested in estimating $\bar{Y}_{\cdot\cdot}$ from a sample of recordings made on t units selected at random out of NM units in the population.

Now we define variations in v_i's, d_j's and e_{ij}'s as

$$\sigma_v^2 = \sum_{i=1}^{N} v_i^2 / N ; \sigma_d^2 = \sum_{j=1}^{M} d_j^2 / M; \quad \sigma_{ij}^2 = E(e_{ij})^2 \quad \text{and}$$

$$\sigma_e^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \sigma_{ij}^2 / NM \qquad \ldots(4\cdot3)$$

Also, for any positive integer $k \geqslant 2$, we define,

$\rho_k \ \sigma_{ij}^2 = E\ (e_{ij} .\ e_{ij+k})$, so that for large M

$$\rho_k \ \sigma_e^2 \cong \sum_{i=1}^{N} \sum_{j=1}^{M-K} E\ (e_{ij\cdot,ij+k}) / N(M-k) \qquad \ldots(4\ 4)$$

Let M days be grouped into M_o clusters each of size $l \geqslant 2$ we define,

$$\rho_l^* \ \sigma_e^2 \cong \sum_{i=1}^{N} \sum_{|j-j'| \leqslant l} E\ (e_{ij}.e_{ij}') / NM\ (l-1) \qquad \ldots(4\cdot5)$$

Similarly,

$$r_k \ \sigma_d^2 = \sum_{j=1}^{M-k} d_j . d_{j+}k / (M-k) \qquad \ldots(4\cdot6)$$

and

$$r_l^* \ \sigma_d^2 \cong \sum_{|j-j'| \leqslant l} d_j . d_j' / M(l-1) \qquad \ldots(4\cdot7)$$

Also,

$$\bar{\rho}_k = \left(\rho_k \ \sigma_e^2 + r_k \ \sigma_d^2 \right) \Big/ \left(\ \sigma_e^2 + \ \sigma_d^2 \right) \text{ and} \qquad \ldots(4\cdot8)$$

$$\bar{\rho}_l^* = \left(\rho_l^* \sigma_e^2 + r_l^* \sigma_d^2 \right) \Big/ \left(\sigma_e^2 + \sigma_d^2 \right) \qquad \ldots(4.9)$$

Now assuming, $E\,(e_{ij}.e_{ij})=0$,

$$\sum_{i \neq i'} v_i\, v_i' = 0 \text{ for all } i \neq i'$$

We will work out variance of $\bar{y}..$, the simple mean estimator of $\bar{Y}..$ based on t units (Y_{ij}'s) selected according to schemes $A, B...E$ listed in Section 2 as follows.

For a given sampling scheme $ S, we will denote variance of $\bar{y}..$ by $V(\bar{y}..)_S$ and efficiency of S compared to an alternative Scheme T as

Eff. $(S \sim T) = V(\bar{y}..)_T / V(\bar{y})_S$

(*i*) For Scheme A, ignoring finite population corrections,

$$V\,(\bar{y}..)_A = \left(\sigma_v^2 + \sigma_d^2 + \sigma_e^2 \right) / t$$

$$= \sigma^2\,(1+\lambda_1+\lambda_2)/t \qquad \ldots(4.10)$$

Where $\lambda_1 = \sigma_d^2 \;/\; \sigma_v^2$

and $\lambda_2 = \sigma_e^2 \;/\; \sigma_v^2$

(*ii*) For Schemes B and C,

$$V\,(\bar{y}..)_B = \sigma_v^2 \;/ n_0\; \sigma_e^2 \;/\; t.$$

$$V(\bar{y}..)_C = \sigma_d^2 \;/\; m_0 + \sigma_e^2 \;/\; t$$

Since, $n_0 = t/M$ and

$m_0 = t/N$

$$V\,(\bar{y}..)_B = \sigma_v^2 \;\; (M+\lambda_2)/t \qquad \ldots(4.11)$$

$$V\,(\bar{y}..)_C = \sigma_v^2 \;\; (N\,\lambda_1+\lambda_2)/t \qquad \ldots(4.12)$$

Hence

Eff. $(B \sim A) = (1+\lambda_1+\lambda_2)/(M+\lambda_2)$ and

Eff. $(\mathrm{C} \sim B) = (M+\lambda_2)/(N\lambda_1+\lambda_2) \qquad \ldots(4.13)$

$\cong M/(N\lambda_1)$

for sufficiently large M and N.

When days are selected in the form of stratified sample with one unit per stratum out of $l=M/m_0$ units in each stratum,

$$\text{Eff. } (C_{st}\sim B)=M/[N\lambda_1+\lambda_2]\,\alpha^*_l \qquad \ldots(4.14)$$

where

$$\alpha^*_l=\left(1-\bar{\rho}^*_l\right)$$

When days are selected in the form of systematic sample with an interval of k days,

$$\text{Eff. } (C_{sys}\sim B)\cong M/\,[N\lambda_1+\lambda_2]\,\alpha_k \qquad \ldots(4.15)$$

where,

$$\alpha_k=1+(m_0-1)\bar{\rho}_k$$

(*iii*) For Scheme D,

$$V(\bar{y}..)_D=\sigma^2_v\;/n+\sigma^2_d\;/m+\sigma^2_e\;/t$$

The optimum choice of m and n subject to restriction $t=mn$, is given by

$$m\,\sigma^2_v\;=n\,\sigma^2_d$$

or

$$m/n=\lambda_1$$

$$V(\bar{y}..)_D=(C\sigma^2_v\;/t)\,(2m+\lambda_2) \qquad \ldots(4.16)$$

Hence

$$\text{Eff. } (D\sim B)=(M+\lambda_2)/(2m+\lambda_2) \qquad \ldots(4.17)$$

$$\cong M/2m>1,$$

if

$$m<M/2$$

(*iv*) For Scheme E,

$$V(\bar{y}..)_E=\sigma^2_v\;/n+\sigma^2_d\;/mn+\sigma^2_e\;/mn$$

$$=\left(\sigma^2_v\;/t\right)[m+\lambda_1+\lambda_2]$$

$$\text{Eff. } (E\sim B)=(M+\lambda_2)/(m+\lambda_1+\lambda_2)$$

$$\text{Eff. } (E\sim A)=(1+\lambda_1+\lambda_2)/(m+\lambda_1+\lambda_2)$$

When m days are selected in the form of stratified sample with one unit per stratum out of $l=M/m$ units in each stratum,

$$\text{Eff. } (E_{st}\sim A)=(1+\lambda_1+\lambda_2)/\left[m+(\lambda_1+\lambda_2)\,\alpha^*_l\right] \qquad \ldots(4.18)$$

Also, when m days are selected systematically with an interval of k days,

$$\text{Eff. } (E_{sys}\sim A)=(1+\lambda_1+\lambda_2)/[m+(\lambda_1+\lambda_2)\,\alpha_k] \qquad \ldots(4.19)$$

From the above formula we conclude that as compared to Scheme A which is infeasible from the point of view of field work, every other Scheme as such, is less efficient. However, E_{st} and E_{sys} could be equally or even more efficient depending upon the nature of intraclass correlations. Comparison of the last three Schemes to B indicates that whereas Schemes D and E will always be more efficient so long as $M>2m$ and $M>(m+\lambda_1)$ respectively, the Scheme C will be preferable only when the number of time units M is greater than λ_1 times the number of special units N.

5. Efficiencies of Alternative Schemes of Sampling Proposed in Section 3

Consistent with the formula developed in Section 4 we will work out efficiencies of alternative schemes proposed in the earlier sections. Table 3 gives analysis of variance of the data described in Section 2, assuming that each village uniformally consists of 5 orchards with 12 bearing trees of guava planted in each orchard and the duration of harvesting as 136 days.

TABLE 3

Analysis of variance for estimation of different components of variance (Figures in Kg.^{2})

Source of variation	*d.f.*	*M.S.*	*E (M S.)*	*Est. (Components of variance)*
Between villages	19	1428	$\sigma_e^2+136\ \sigma_v^2$	9·0
Between days	135	630	$\sigma_e^2+\ 20\cdot\sigma_d^2$	21·4
Residual	2565	202	σ_e^2	202·0
Total	2719	232	σ^2	232·4

It will be seen from the table that variation due to random effects in yield is nearly 20 times of that between days. Also, the variation in village effects is considerably small being nearly half of that between days. Estimates of λ_1 and λ_2 as defined in 4.10 are obtained as 2.38 and 22.44 respectively.

Table 4 gives estimates of intra-class correlation coefficients $\bar{\rho}_k$ and $\bar{\rho}^*_l$ as defined in (4.8) and (4.9) respectively for varying values of k and l. Since the harvesting period (M_i) varied from village to village, the values of $\bar{\rho}_k$ have also been worked out taking into account varying lengths of the harvesting period ineach village.

TABLE 4

Estimates of values of intra-class correlation coefficients

Values of intra-class correlation coefficients	*Interval (k) of systematic sample or size (l) of stratum in days*							
	Two	*Three*	*Four*	*Five*	*Six*	*Seven*	*12 days*	*14 days*
1. $\bar{\rho}_k$ with constant period of harvesting in each village.	·0016	—·0167	—·0100	—·0260	—·0219	·1477	—	—
2. $\bar{\rho}_k$ with varying period of harvesting.	·0283	—·0114	—·0025	—·0230	—·0192	·1833	—	—
3. $\overline{\rho_l^*}$ with constant size (l) of each stratum.	·2876	·3318	·3471	·3634	·3858	·3911	·3845	·3666

It is seen from the table that intra-class correlations between yields harvested on alternate and every seventh day is positive, while those between yields harvested either every third, fourth, fifth or sixth day aie negative irrespective of the fact whether harvesting period remains constant or varies from village to village. Also, the harvests made on consecutive days are positively related, the intra-class correlation coefficient increases consistently up to a period of seven days and thereafter starts decreasing slowly. Using the figures given in tables 3 and 4, efficiencies of Schemes B and C as compared to Scheme A are found to be only 16.0% and 6.7% respectively.

The efficiency of Scheme E under various alternative selection procedures is given in table 6.

TABLE 6

Percentage efficiency of Scheme E as compared to Scheme A for varying sizes of m, the secondary stage sample of days

Type of Scheme for second stage selection	*Size of secondary stage sample selected as one day out of every*					
	Two	*Three*	*Four*	*Five*	*Six*	*Seven days*
1. Selecting *independently* from *constant* No. of *M* days of harvesting in each village	28	37	44	50	54	57
2. Selecting in the form of a *stratified* Sample assuming *constant* number of time strata in each village.	30	42	51	60	68	74
3. Selec:ing in the form of *systematic* sample with constant period of harvesting for each village.	24	43	45	64	65	22
4. Selecting in the form of systematic sample with varying periods of harvest for each village.	28	51	52	75	74	24

The figures given in table 6 show that selection of units in two stages results into loss in efficiency. This loss could be substantially compensated by selecting the units at the second stage in the form of either stratified or systematic sample. However, stratified sampling is rather inconvenient from the point of view of field work. The systematic sampling, which is equally efficient, can be adopted. The five and six days' intervals of sampling are equally optimum. The loss due to second stage selection is then considerably reduced. For organizing the field work of the survey, the sampled villages should be randomly divided into five or six groups. The numbers 1 to 5 or 1 to 6 should then be allotted at random to each of the such groups. The yield data in the villages in each group should be collected on every fifth or sixth day starting from the randomly allotted number. Thus the party of fieldmen would collect data from all the 1/5th or 1/6th of

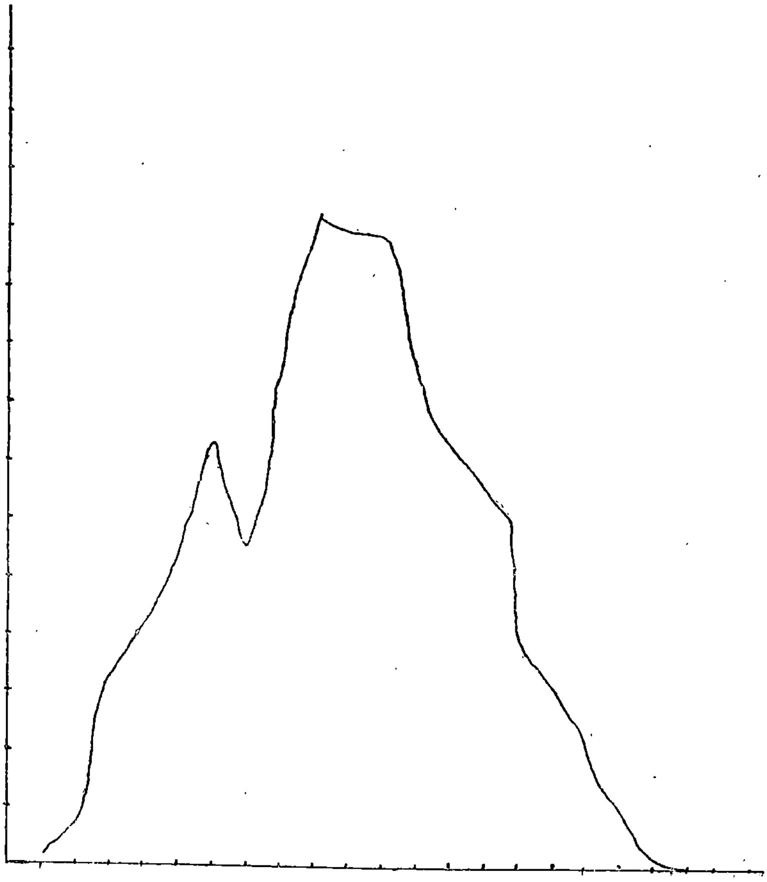

Fig. 1,

villages in the group and move to the next group of villages next day returning back to the same group after a lapse of either four or five days. By this way, data collection work would be very convenient.

6. Concluding Remarks

The formulae for comparing efficiencies have been derived under two major assumptions *viz.* (*i*) the linearity of effects, and (*ii*) absence of finite population corrections. The first assumption is not justified specially in view of the fact that, the population indicates a trend (cf. Appendix I) over time which appears to be approximately of parabolic form. Consistent with this trend, one may attempt the comparison of various sample schemes under a quadratic model of the form

$$Y_{ij}=u+v_i+\alpha d_j+\beta d_j^2+e_{ij} \qquad \ldots(6.1)$$

Under this model, the efficiency formula will be rather complicated and will depend upon the values of α and β in addition to the variability parameters and intra-class correlations. The estimation of α and β will also present a problem. Utility of the present study is to give an indication about the likely optimum sampling scheme prior to planning of the survey, without bringing in a complicated analysis. So far as the second assumption regarding ignoring of finite population corrections in various sampling schemes is concerned, the optimality conditions of the design may not undergo any change.

The Scheme D which involves two way independent selection and is operationally very convenient—perhaps the most convenient one in case the data collection work is done on part-time job basis, deserves one remark.

Under systematic sampling,

$$\text{Eff. }(D_{Sys}\sim E_{Sys})=\frac{M+(\lambda_1+\lambda_2)\alpha_k}{2m+\lambda_2\alpha'_k} \qquad \ldots(6.2)$$

$$\text{where } \alpha_k=1+(m-1)\,\overline{\rho_k}$$

$$\text{and } \alpha'_k=1+(m-1)\,\rho_k$$

Clearly, *D* will be less efficient than *E* for moderately large values of *m* approximately greater than $\lambda_1\alpha_k$. Under model 6.1, we may perhaps find that loss in efficiency is not substantial. Hence, from the point of view of convenience of field work, the Scheme D may be preferable over Scheme E.

Summary

The paper deals with comparing efficiencies of different sampling schemes for estimating mean of a finite population spread in space and time. For the type of data studied in this paper and the pattern of intra-class correlations revealed by the data over time, it has been found that two-stage sampling design with special units (villages) as primary and time units (days) as secondary sampling units, selected in the form of a systematic sample is the optimum feasible design. The optimum intervals of systematic sampling are found to be 5 or 6 days, both being equally efficient. Systematic sampling with alternative or weekly intervals have been found to be the worst sampling scheme.

Reference

[1] Seth, G.R., Sukhatme, B.V. and Manwani, A.H. : "Survey on Mango and Guava in Uttar Pradesh", *Technical Bulletin*, Indian Council of Agricultural Research, New Delhi.