SIZE OF UNITS IN A TWO-STAGE SAMPLING SCHEME FOR THE ESTIMATION OF CROP ACREAGE

By J. M. SEN GUPTA

Indian Statistical Institute

SUMMARY. This is an empirical study on the relative efficiencies of two-stage sampling with varying sizes of village-clusters in the first stage and varying sizes of plot-clusters in the second, for the estimation of percentage area under certain crops in West Bengal.

Variability as a function of the size of sampling units in both the stages has been studied. It has been found that variability between plot-clusters is a joint function of its own size as well as the size of the village-clusters from which they are selected.

Approximate cost functions have been constructed on the basis of the results obtained earlier by Mahalanobis in his experimental surveys on Jute acreage during the years 1937-40 and certain basic assumptions made from current experiences.

It has been found that the gain by increasing the size of first stage units is not much, while efficiency increases slightly by increasing the size of plot-clusters in case of Jute and Aus but the other way for Aman when field cost alone is considered. If the cost of statistical work is also taken into account, small sized plot clusters seem to have distinct advantages.

1. Introduction

Extensive and thorough investigations were carried out in undivided Bengal early in 1937 by Mahalanobis in his pioneering work on the question of sampling for Jute acreage. The results were issued as a series of departmental Reports by the Indian Central Jute Committee (ICJC) during 1938-41 and were later published as scientific papers in various journals (1944, 1946a, 1946b).

The sampling design adopted in these years was however a unistage one, where individual Police Stations of size 150-250 sq. miles were treated as ultimate strata or as the element of a stratum. In these investigations, a study of the reduction in variance of the proportion of area under jute crop between square-shaped 'grids', with an increase in its size ranging from 1 acre (7 plots) to 30 acres (115 plots) was made. These grids enclose a few plots wholly inside while some of them were intercepted by the grid-sides, a portion falling outside the square. The proportion of land under crop is then eye-estimated for all the plots constituting the grid. It was observed that due to an incidence of positive space correlation between adjoining plots, fall in variability was much less than what could be normally expected if the plots were mutually independent in respect of Jute cultivation. A least square fit for the variance function as $V = a(x)^g$, where V is the variance and x is the size of grid in acres, a and g are unknowns was tried out and the coefficient g was found to range between -0.40 to -0.70 i.e., less than unity.

The size of sampling unit was accordingly chosen so as to ensure a maximum of precision at a given cost.

In the National Sample Surveys currently being carried out under the technical guidance of the Indian Statistical Institute, plot-clusters instead of 'grids' selected equal or unequal probability are however being employed as the ultimate sampling units. The

adopted size of clusters with ten plots each, broadly corresponds to the standard grids of 2.25 acres, employed by the Institute in its later all-Bengal surveys. It may be noted here, that in the earlier years, a very detailed analysis of the relative efficiencies of the grid-size, taking into account the cost of field work and statistical work involved in handling them, was carried out. While the costs of statistical operations depended on grid-size and its total number, field cost was a function of grid-size as well as the density with which they were spread over the entire area to be covered. The optimum choice for the best combinations of grid-size and its density per unit of geographical area, to give the highest precision at a given level of cost, varied from stratum to stratum, which were worked out accordingly. But the later surveys extending over the whole of undivided Bengal were to cover jute, autumn paddy as also other important crops in a single survey during the entire season. So the crop-specific solutions for size-density allotments to individual strata would no longer apply, and as a compromise a standard grid of size 2.25 acres had eventually to be adopted.

The National Sample Surveys covering the whole of India had necessarily to go by a much broader choice, the task of studying the crop pattern for this immense region, the demands of a large variety of crops and finally the multipurpose character of the survey tied up with many other socio-economic enquiries hardly left any freedom what a single crop and a single purpose would have permitted. With a relatively small staff in proportion to the huge area to be covered, a two-stage sampling for the estimation of crop-area had naturally to be resorted to, individual villages serving as the first-stage units with plot-clusters selected within them in the second stage. A study of the crop-specific behaviour of variability in different stages of sampling, would nevertheless be of interest and may prove useful in other circumstances.

In this paper an attempt has been made to study the relative efficiencies with clusters of varying sizes in a two-stage sampling on the basis of data obtained through a special scheme of survey in autumn of 1961-62. Although, data regarding field costs were not collected in this scheme, approximate cost functions have also been worked out partly on old findings, and partly on empirical considerations.

2. THE SAMPLING SCHEME AND DATA DISPOSAL

A special scheme using village-clusters of varying size with clusters of 30 plots within them in the second stage was tried out during the autumn surveys in 1961-62. The sampling design was a stratified ,two-stage one, where 324 village-clusters were allocated among the strata in the proportion $g\sqrt{pq}$, where g was the geographical area of the stratum, p the proportion of area under autumn crops and $q = 1-p$. Within each stratum, the village clusters were selected with probability proportionate to geographical area from which plot-clusters were also similarly selected. The survey was conducted in three sub-rounds. In addition to the principal crops Jute and Aus paddy cultivated in autumn, data was also collected in respect of the Winter or 'Aman' paddy so far as it was sown at the time of visit. The third sub-round which was a fully representative sub-sample of the whole, was carried out in September-October, by which time the sowing of Aman paddy was definitely completed. Apart from a number of casualties, some of the four-village clusters were not surveyed in full, leaving only 218 complete first-stage units for Jute and Aus paddy, and 55 first-stage units for Aman, from which an identical number of sub-units for each different size could be formed.

In this scheme, clusters of four villages were selected in a manner such that sub-clusters of 1, 2 and 4 villages could be formed out of them in a nested formation. The clusters of 30 plots were likewise capable of being broken up into sub-clusters of 1, 2, 3, 5, 10, 15 and 30 plots in a nested pattern. In both the stages, the entry unit, village or plot selected at random, represented the nuclear unit about which the prescribed clusters of varying sizes were constituted.

For instance, if the entry village is numbered as 87 and an entry plot selected in the second stage is numbered as 47 the village-clusters of sizes 1, 2 and 4 will be formed with villages numbered 87, 87-88 and 85-88 with the various plot-clusters constituted of plots numbered 47, 47-48, 46-48 etc.

Thus, if the cluster with plots 31-60 in any of the village-clusters was enumerated, data for plot-clusters of all the other sizes would be automatically furnished. Three clusters of 30 plot each, were selected with probability proportionate to geographical area separately and independently within each of the three village-clusters of one, two and four villages. Thus, although the Investigator on reaching a village-cluster had to survey a total of 9 clusters spread over the group of 4-villages, the clusters of one, two and four villages had each its own quota of plot-clusters for purposes of estimation.

The nested sets of sub-clusters of plots in the second stage and of villages in the first stage, thus constituted, are not mutually independent and for this reason impose a good deal of control for purposes of the comparison we have in view.

Ishaque's plot-wise complete enumeration of the whole of undivided Bengal in 1940, again gives us an opportunity for working out the population variabilities for village-cluster units of any desired size. The available reports published by the Government however give the acreages under different utilisation by Union (15-20 villages) and Police Station (10-12 Unions) breakdowns. But the Institute had been able to compile the village-wise figures for Aman paddy in the year 1950-51 from the detailed data sheets which were preserved in the government records up to that time.

3. Variance function

Unistage sampling with plot-clusters. It may be noted here that the variance functions worked out on Jute in the earlier years in Bengal refer to 'grids', i.e., equivalents of plot-clusters selected unistage over the stratum, whereas in the recent West Bengal surveys, plot-clusters were selected in the second stage from villages in the first stage. An unistage variability of the order of 95% to 120% was obtained for Jute in 1940 in the eight important Jute districts of undivided Bengal with a grid-size varying between 1-acre to 9-acres. It will be interesting to compare these coefficients with the "total" variation between plot-clusters of different sizes obtained from a two-stage sample. Coefficients of total variation have accordingly been given in Table 1 for Jute, and also for Aus and Aman paddy, where n_1 and n_2 stand for the number of first-and second-stage units respectively.

It will be seen that the coefficients of variation for Jute in West Bengal are much higher than the earlier results for undivided Bengal in 1940. The high intensity of Jute cultivation in the selected eight districts of undivided Bengal would explain the difference. For Aus and Aman paddy however, we have no material in hand for a comparison.

TABLE 1. COEFFICIENT OF 'TOTAL' VARIATION BETWEEN PLOT-CLUSTER OF VARIOUS SIZES FOR THE PROPORTION OF LAND UNDER DIFFERENT CROPS

size of clusters in terms of plots	total variability between plot cluster in two-stage sampling			total variability between plot cluster in two-stage sampling			total variability between plot cluster in two-stage sampling		
	1-village unit in 1st stage	2-village unit in 1st stage	4-village unit in 1st stage	1-village unit in 1st stage	2-village unit in 1st stage	4-village unit in 1st stage	1-village unit in 1st stage	2-village unit in 1st stage	4-village unit in 1st stage
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	($\bar{p}$ = .0638 n_1 = 218 n_2 = 654)			($\bar{p}$ = .0735, n_1 = 218, n_2 = 654)			($\bar{p}$ = .5596, n_1 = 55, n_2 = 165)		
	jute			aus			aman		
1	348	331	347	305	307	410	93	92	83
2	313	303	310	284	286	335	88	85	79
3	280	288	278	263	269	317	83	87	78
5	268	262	252	248	248	268	82	80	77
10	227	232	220	222	226	237	76	79	73
15	208	205	210	215	212	215	76	79	76
30	187	191	180	191	192	192	72	75	68

Two-stage sampling with village-clusters in the first stage.

Variability (true) of village-clusters as first stage units. Table 2 below gives the observed coefficients of 'true' variation between first-stage units in cols. (2), (4), (6) and (8). The coefficients based on Ishaque's complete census of 1946 in col. (8) gives the variability of the proportion of area under winter paddy between individual villages which were obtained by a model sampling on the village-wise material. A sample of villages was chosen with probability proportionate to geographical area and with a stratification and allocation of villages exactly parallel to the one employed in the autumn survey of 1961-62. The clusters of 2-villages, 4-villages, 8-villages, etc. were built up in a nested formation with the originally selected village as entry. This makes the two sets of results readily comparable. Except in case of Ishaque's where the villages were completely enumerated, directly giving the 'population' coefficients of variation between village units, the 'true' coefficients had to be estimated in all other cases by an analysis of the total variance into their stage components, as :

C_1 = variability between village-clusters as $100\sqrt{\left|\frac{V_1 - V_2}{3}\right|}$ mean

C_2 = variability within village-clusters $100\sqrt{V_2}$ mean,

where V_1 and V_2 are the observed variances in a two-stage analysis, there being 3 plot-clusters per village-cluster. The largest, i.e., the full size plot-clusters of 30 plots (in 1961-62) has been accepted giving the best estimates of variance between villages, which of course should be independent of the size of second-stage units.

It will be seen that the variability between village-clusters falls generally with an increase in its size in all cases, Jute, Aus and Aman paddy. The coefficients given in Tables 1 and 2 all refer to the pooled measure of variation within all the strata.

Graduation of the coefficients of first-stage variation. A least square fit in the form $C_1 = A(v)^{g_1}$ was made on the coefficients of first-stage variation on the size offirst-stage units in the proportion of area under Jute, Aus and Aman paddy as observed in 1961-62 and under Aman paddy according to the Ishaque's in 1946. The graduated values of first-stage variability are given in cols. (3), (5), (7) and (9) of Table 2 below. The ratios of the components of variation accounted for by the fit to the residual i.e., the deviation

TABLE 2. GRADUATED VALUES OF FIRST STAGE VARIABILITY (C_1) OBTAINED BY A LEAST SQUARE FIT ON SIZE, i.e., NUMBER OF VILLAGES v CONSTITUTING THE FIRST STAGE UNIT IN THE FORM $C_1 = A(v)^{g_1}$

Size of village cluster in the first stage (v)	jute, 1961–62 $\bar{p} = .0638$ $n_1 = 218$		aus, 1961–62 $\bar{p} = .0736$ $n_1 = 218$		aman, 1961–62 $\bar{p} = .5596$ $n_1 = 55$		Ishaque's aman, 1946 $\bar{p} = .4762$ $n_1 = 60$	
	observed	expected	observed	expected	observed	expected	observed	expected
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
one	134	138	111	110	36	36	60	58
two	146	135	109	104	36	36	54	53
four	128	132	99	98	33	33	49	47
eight	—	129	—	93	—	31	45	43
sixteen	—	126	—	88	—	29	40	38
constants	$A = 137.80$		$A = 109.80$		$A = 38.39$		$A = 58.32$	
	$g_1 = -.0320$		$g_1 = -.0809$		$g_1 = -.1029$		$g_1 = -.1499$	

therefrom, were found to be highly significant, indicating that the fits were good. The coefficients g are found to be negative in all cases. The coefficients of variation for Jute, Aus and Aman could thus be predicted even beyond the four-village size with some degree of assurance, assuming that there is no abrupt break in the continuity of its behaviour pattern. For Aman in Ishaque's time we can safely predict for a size of 16-villages without any such presumptions.

It will be seen that the first-stage variability in 1946 (Ishaque) is much higher than that observed in 1961-62. This is perhaps natural, considering that the cultivation of Aman has since then appreciably increased, most of the villages being now under some degree of cultivation.

Variability of plot clusters as second-stage units. We are thus in a position to work out the number of village-clusters of specified sizes to be allotted to an individual stratum, in order to attain any given level of precision, for an unistage sampling scheme. In a two-stage sampling, variability of plot-clusters of a given size within village-clusters in the first-stage is also involved.

Table 3 in cols. (2), (4) and (6) below gives the observed coefficients of variation C_2 between plot-clusters of different sizes (in terms of the number of plots p) within village-clusters of varying sizes (in terms of the number of village v) for Jute, Aus and Aman in 1961-62. It will be seen that the second-stage variability decreases for Jute, but increases for Aus, as the size of village-clusters increases from one to four, although their behaviour is rather irregular in the intermediate stage, i.e., with two-village clusters. For Aman, it increases up to 2 village-cluster and then falls with a further increase of the size of the first-stage units.

TABLE 3. GRADUATED VALUES OF SECOND-STAGE VARIABILITY (C_2)OBTAINED BY A LEAST SQUARE FIT ON THE SIZE OF FIRST-STAGE UNITS (v)

AND SECOND-STAGE UNITS (p) IN THE FORM $C_2 = a(v)^{\beta_1}.(p)^{\beta_2}$

size of clusters in terms of plots	1-village unit in 1st stage		2-village unit in 1st stage		4-village unit in 1st stage	
	observed	expected	observed	expected	observed	expected
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	crop : jute					
	($p = .0638$ $n_1 = 218$, $n_2 = 654$; $a = 341.30$ $\beta_1 = -.0610$ $\beta_2 = -.3336$)					
1	304	341	294	327	300	314
2	276	271	265	260	267	249
3	226	237	237	227	221	217
5	224	200	216	191	193	183
10	158	158	173	152	139	146
15	145	138	133	133	132	127
30	106	110	91	105	91	101
	crop : aus paddy					
	($p = .0735$ $n_1 = 218$ $n_2 = 654$; $a = 296.45$ $\beta_1 = .1361$ $\beta_2 = -.2377$)					
1	284	296	282	326	401	358
2	253	251	260	276	329	304
3	231	228	233	251	309	276
5	211	202	209	222	246	244
10	178	171	183	188	208	207
15	182	166	161	171	186	188
30	137	132	135	145	151	160
	crop : aman paddy					
	($p = .5596$ $n_1 = 55$, $n_2 = 165$; $a = 82.02$ $\beta_1 = -.0078$ $\beta_2 = -.0677$)					
1	82	82	87	82	80	81
2	77	77	81	77	72	76
3	70	74	79	74	71	74
5	69	71	73	71	69	70
10	61	67	69	67	68	66
15	67	65	69	64	61	64
30	59	61	65	61	58	60

For a given size of the first-stage units, the second-stage variability however falls steadily with an increase in the size of second-stage units, i.e., plot-clusters for all the three crops. From its behaviour as exhibited above, it seems reasonable to presume that the second stage variability is a joint function of the size of sampling-unit in both stages.

Graduation of the coefficients of second-stage variations. A multiple fit in the form $C_2 = a(v)^{g_1} . (p)^{g_2}$, where C_1 is the coefficient of variation between first-stage units of size v villages, C_2 is the same for second-stage units of size p plots, within first-stage unit, has therefore been tried out by the method of least squares.

In Table 3, the coefficients a, g_1 and g_2 have been given at the top of the table, the graduated values of C_2 being shown for each size of stage unit in cols. (3), (5) and (7) along with the corresponding observed values. For testing the goodness of fit, analysis of the total variance C_2 into its two components, namely, that due to the fit with its 2 degrees of freedom and the rest, has been done and the results are shown in Table 4.

TABLE 4 : ANALYSIS OF VARIANCE OF THE SECOND-STAGE VARIABILITY (C_2) WITHIN FIRST-STAGE UNITS FOR TESTING THE GOODNESS OF FIT IN THE FORM $C_2 = a(v)^{g_1} . (p)^{g_2}$

source	jute			aus			aman		
	d.f.	variance	F	d.f.	variance	F	d.f.	variance	F
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. due to regression	2	.271400	209.4	2	.148500	130.3	2	.018639	37.1
2. deviation	18	.001296		18	.001140		18	.000502	
3. total	20			20			20		

4. Cost function

As already stated, data relating to the cost of field work or statistical work had not been collected during this investigation. The parameters of field cost worked out by Mahalanobis in earlier years no longer hold good because of radically changed circumstances, partly due to changes in the performance rates but mainly due to the changed pattern of movements in adopting a multistage sampling scheme.

An empirical structure of operational costs. An approximate cost structure can however be built up on the basis of our past experiences with certain *ad hoc* assumptions. Accordingly, the enumeration costs per plot-cluster of different sizes were worked out using Mahalanobis's equations. Assuming that three standard clusters of 10 plots each spread over a single village can be surveyed in one working day (on *a priori* consideration), the number that could be surveyed in bigger clusters of 2 and 4 villages were calculated on the concept that the inter-cluster journey component varied inversely with density, while enumeration cost depends on the size of plot-clusters alone. The cost of statistical operations have however been based on recent experiences.

Field costs. Table 5 below gives an empirically built-up structure of field cost covering various size combinations of sampling units in the two stage. Against each combination of the size of village-cluster in the first stage with size of plot-clusters within them, the expected number of plot-clusters that can be surveyed per day of gross haltage, have been given in cols (2)—(4). The procedure adopted for working out these performance rates have been given in paras 1—3 of the Appendix.

TABLE 5. NUMBER OF PLOT CLUSTERS OF p PLOTS PER VILLAGE-CLUSTER OF SIZE v-VILLAGES ($n_{p,v}$) WHICH CAN BE COVERED PER GROSS DAY OF FIELDWORK.

size of plot cluster (p)	number of plot clusters per cluster of v villages in the first stage		
	one-village cluster	two-village cluster	four-village cluster
(1)	(2)	(3)	(4)
1	5.58	5.10	3.68
2	4.70	3.72	2.85
3	3.92	3.22	2.50
5	3.32	2.80	2.22
10	3.00	2.58	2.08
15	2.72	2.38	1.95
30	2.28	1.92	1.60

Cost of statistical operations. Total cost of statistical operation has been worked out for each of the size-combinations corresponding to the sample-size that can be surveyed at a given field cost and the same expressed as indices to the same for a standard sized cluster of 10 plots in one-village units, have been given in Table 6 below. The detailed breakdowns of the components of statistical operations have been given in para 4 of the Appendix. It will be seen that for a given size of village-cluster the cost goes down as the size of plot-cluster falls, reaching a minimum for clusters of one plot, where the heaviest component of cost, namely area extraction and calculation of proportions at the cluster level is entirely eliminated. For a given size of plot-cluster the overall statistical cost also goes down somewhat with an increase in the size of village-clusters. This is due to a reduction in the number of plot-clusters that can be surveyed at a given total of field cost, journey time between plot-clusters within larger village-clusters being greater there.

TABLE 6. COST OF STATISTICAL OPERATIONS FOR DIFFERENT SIZE-COMBINATIONS CORRESPONDING TO THE DATA COLLECTED AT A GIVEN TOTAL OF FIELD COST EXPRESSED AS INDICES TO THE SAME IN RESPECT OF TEN PLOT-CLUSTERS IN ONE-VILLAGE UNITS IN THE FIRST STAGE

size of plot-cluster (plots)	size of village cluster		
	1-village	2-village	4-village
(1)	(2)	(3)	(4)
1	15.3	13.1	11.6
2	39.7	32.6	27.0
3	45.6	38.6	31.9
5	59.5	51.1	42.6
10	100.0	86.8	71.9
15	132.9	116.8	97.8
30	216.2	184.0	154.7

5. Relative efficiencies of clusters of varying sizes in the first and second stages

Percentage error of the proportion of area under a crop, sampled in two stages can be calculated from the relation :

$$C_1^2/N + C_2^2/N \cdot n_{p,v} \cdot h = C \qquad \ldots \quad (1)$$

while total field cost T in gross days is obtained as

$$N(J+h) = T \qquad \ldots \quad (2)$$

where,

C_1 and C_2 are the coefficients of true variation in the first and second stage, and C is the percentage error of the estimated mean,

N is the number of first-stage units of v villages,

$n_{p,v}$ is the number of second-stage units of p plots per first-stage unit that can be enumerated in one day of gross haltage, as given in Table 5,

h = gross days of haltage per first-stage unit

T = total field days spent and

J = gross days of journey per first-stage unit, which may be taken as 1.5 irrespective of their size.

The values of N and T for 10-plot clusters in one-village units, has been computed from (1) and (2) above, substituting $n_{p,v} = 3.00$ from Table 5, when $C = 5\%$, and $h = 4$.

Once we know the value of total cost = T-days, we can work out the value of N for any other haltage h from equation (2). Substituting $n_{p,v}$, based on Table 5, we can determine the resultant percentages of error for any other size-combination p ,v for any given haltage h, with the help of equation (1). It may be noted here that with a given haltage per first stage units, the number of second-stage units per first-stage unit decreases as the size of unit in either stage increases. The results as given in Table 7 separately worked out for Jute, Aus and Aman paddy, using haltages of 4 days and 8 days per first-stage unit.

It will be seen that for Jute, 857 one-village units with standard clusters of 10 plots at a cost of 4714 field days are needed for precision of 5%. The same for Aman works out to be 74 one-village units at a field cost of 407 days only.

For the State as a whole which will be broken up into a number of strata, the over-all size of the sample will remain more or less of the same order.

We have here considered the relative merits of the different size-combinations against the performances of standard clusters of 10 plots in one-village units at a given cost in respect of field work alone, ignoring costs of statistical operations involved in each.

For both Jute and Aman paddy, there is only a slight gain by increasing the size of first-stage units. On the other hand, while Jute and Aus gain a little by increasing the size of plot-clusters, it is the other way for Aman, there being some gain by reducing the size of plot clusters. For Aus paddy, with a haltage of 4 days, an increase in the size of first unit results in a loss, while there is a slight gain when the haltage is increased. An increase in the size of plot-clusters also tends to reduce the percentage error. The overall merits of the small-sized plot-cluster can however be judged when cost of statistical operations is also taken into

TABLE 7. P.C. VARIABILITIES OF THE PROPORTION OF AREA UNDER JUTE, AUS AND AMAN PADDY, ATTAINED BY THE VARIOUS SIZE COMBINATIONS OF CLUSTERS IN THE TWO STAGES AT A GIVEN FIELD COST (THAT WOULD ENSURE 5%ERROR WITH STANDARD 10 PLOT CLUSTERS FROM ONE-VILLAGE UNITS AT 4 DAYS OF HALTAGE PER VILLAGE), SEPARATELY FOR HALTAGE OF 4 DAYS AND 8 DAYS PER VILLAGE CLUSTERS

size of plot clusters (in plots) (1)	scheduled haltage of 4 days per village cluster of			scheduled haltage of 8 days per village cluster of		
	1-village (2)	2-village (3)	4-village (4)	1-village (5)	2-village (6)	4-village (7)
			jute (p = .0038)			
1	5.3	5.3	5.3	6.5	6.4	6.4
2	5.2	5.2	5.2	6.4	6.4	6.3
3	5.2	5.1	5.1	6.4	6.3	6.3
5	5.1	5.0	5.0	6.4	6.3	6.2
10	5.0	4.9	4.9	6.3	6.2	6.1
15	5.0	4.9	4.8	6.3	6.2	6.0
30	4.9	4.8	4.7	6.2	6.1	6.0
			aus (p = .0735)			
1	5.1	5.2	5.6	6.4	6.3	6.4
2	5.1	5.2	5.5	6.4	6.3	6.4
3	5.1	5.2	5.4	6.4	6.2	6.3
5	5.1	5.1	5.3	6.4	6.2	6.2
10	5.0	4.9	5.0	6.3	6.1	6.0
15	5.0	4.9	4.9	6.3	6.1	6.0
30	4.9	4.8	4.8	6.3	6.0	5.9
			aman (p = .5596)			
1	4.8	4.7	4.6	6.1	5.8	5.6
2	4.9	4.8	4.7	6.2	5.9	5.6
3	5.0	4.8	4.7	6.2	5.9	5.7
5	5.0	4.8	4.7	6.2	5.9	5.7
10	5.0	4.8	4.7	6.2	5.9	5.7
15	5.0	4.8	4.7	6.2	5.9	5.6
30	5.0	4.9	4.8	6.2	5.9	5.7

account. The trend of this cost on the size of plot-cluster, as given in Table 6, seems to indicate, that small-sized plot-clusters have distinct advantages because of a low cost, a saving of 85%, when one-plot cluster is used, instead of ten-plot clusters. It may be noted here that the primary cost of statistical operations up to the stage of prediction (without evaluating the standard errors) represent only 3%-4% of the field cost, but a saving at this stage may help in obtaining quick results.

A serious reduction in the total number of plots to be enumerated however creates some difficulty in the selection of fields for crop-cutting experiments, which have to be kept confined within the enumerated plot-clusters only and thus may prove to be inadequate for meeting the requirement. Furthermore, identification of a single plot is always fraught with some danger. It is perhaps advisable to employ a cluster of at least 2 plots, so that the identification of one on the ground may be mutually corroborated with an identification of the other.

ACKNOWLEDGEMENT

The author gratefully acknowledges the help received from Sri Paul Jacob in writing up this note and specially from Sri Maloy Chanda in all stages of analysis.

Appendix

(1) *Direct field cost in hours of primary field work.* The number of plot-clusters that could be enumerated per village-cluster within a specified haltage has been approximately determined, on the following considerations:

According to our present experiences with one-village-cluster as the first-stage unit, 3 standard clusters of 10 plots can be surveyed per day gross haltage with net 3.75 hours available for work, allowing for leave etc. at 25%. The net working hours would consist of two components, actual enumeration and inter-cluster journeys, other over-head or miscellaneous activities being ignored. Or,

$$3.75 = n_{10,1}(c_{10}+j_{10,1}). \qquad \ldots \quad (1.A)$$

These would be our basic assumptions.

(2) *Enumeration.* Expected enumeration hours per plot-cluster of size x may be worked out from the least square fit $c_x = A + B(x)$ as obtained in our earlier investigations, where x is the size of grid units in terms of acres and c_x represents the enumeration hours per grid. Enumeration hours can be refitted on the size of sample units in terms of the number of plots constituting it, i.e., in terms of average plots per grid. Table A.1 below gives the observed and expected values of enumeration hours per grid in terms of acres (from Table 26, ICJC Report, 1940) and average number of plots per grid of varying sizes.

TABLE A.1. ENUMERATION HOURS PER GRID AS A FUNCTION OF GRID SIZE IN ACRES BY A LEAST SQUARE FIT IN THE FORM

$$c_x = .7559 + .0600(x)$$

grid size in acre (x)	average no. of plots per grid	enumeration hours per grid	
		observed	expected
(1)	(2)	(3)	(4)
1	7.3	0.87	0.82
4	17.8	1.00	1.00
9	34.7	1.35	1.30
16	55.8	—	1.72
36	115.3	—	2.92

From a least square fit of col. (4) on col. (2), we may obtain a new fit $c_p = 0.650 + .0190(p)$.... where 'p' is the number of plots in the cluster.

The graduated values of c_p have been shown below in Table A.2, against various sizes of plot-clusters i.e., p, where p takes up the values 1, 2, 3, 5, 10, 15 and 30.

It may be noted here that the smallest clusters lying below 7 plots are outside our observed range. A general fit in the form of a straight line is not therefore adequate for sizes much below 7. On account of this, the graduation was made graphically for this portion, i.e., for clusters of 1, 2 and 3 plots.

(3) *Journey.* The journey hours per cluster of p plots in a cluster of v village may be taken as

$$j_{p,v} = j_{10,1}\sqrt{\frac{n_{10,1}}{n_{p,v}} \cdot \frac{v}{1}} \qquad \ldots \quad (2.A)$$

assuming that journey time per plot-cluster would vary inversely as the square root of their relative density.

TABLE A.2. ENUMERATION HOURS PER CLUSTER AS A FUNCTION OF CLUSTER SIZE IN TERMS OF THE NUMBER OF PLOTS CONSTITUTING IT, BY A LEAST SQUARE FIT IN THE FORM $e_p = 0.650 + .0196(p)$.

size of plot cluster = p	graduated value of e_p
1	(.30)
2	(.50)
3	(.62)
5	.75
10	.85
15	.94
30	1.24

Substituting $e_{10} = 0.85$ (Table A.2) and $n_{10,1} = 3$ in equation (1.A), the value of $j_{10,1}$ is obtained as 0.10. It may be noted here that enumeration time of per-cluster depends only on the size of plot-cluster and independent of the size of village-cluster. Once $j_{10,1}$ is known, value of the $j_{p,v}$ per cluster of p plots within a cluster of v villages may be evaluated as $j_{p,v} = .095\sqrt{\frac{v}{n_{p,v}}}$ given by equation (2.A). Thus, for any of the combinations of p, and v, where 'p' assumes the values 1, 2, 3, 5, 15 and 30 and v the values 1, 2 and 4 we can work out enumeration hours as well as journey hours per ultimate unit, i.e., per plot-cluster. Then since, $n_{p,v}(e_p + j_{p,v}) = 3.75$ hours, we can determine the value of $n_{p,v}$ i.e., the number of plot clusters surveyed within a gross haltage of one-day in village-clusters of any size. The corresponding values for larger haltages will be increased in the proportion of net working hours. Net hours spent on enumeration and journeys per plot-cluster unit as obtained, has been given in Table A.3 below.

Table A.3. COST IN HOURS OF FIELD WORK (ENUMERATION PLUS JOURNEYS) PER PLOT-CLUSTER OF VARYING SIZES SELECTED FROM VILLAGE-CLUSTERS OF VARYING SIZES IN THE SECOND STAGE

size of plot cluster	cost of field work per second stage unit in hours		
	1-village cluster	2-village cluster	4-village cluster
(1)	(2)	(3)	(4)
1	0.57	0.73	1.02
2	0.82	1.01	1.32
3	0.97	1.17	1.49
5	1.13	1.34	1.68
10	1.25	1.46	1.81
15	1.36	1.58	1.92
30	1.70	1.95	2.31

SIZE OF UNITS IN A TWO-STAGE SAMPLING SCHEME

The longer-journey time per first-stage unit irrespective of their size has been taken as 1.5 days (gross) on general considerations, a flat allowance permitted for changes of camp, so long as the area to be covered is of the order of a district or below. Total cost in gross field days T will come out as

$T = N(1.5+h)$, where h = gross haltage in days per first-stage unit and N is the total number of firststage units.

(4) *Direct cost of statistical operations in hours of manual computation.* The cost of statistical operations, all the stages being operated manually, will be constituted of two major components, one at the level of plot-clusters and the other at the level of village-clusters or first-stage units. Contribution of these components per unit of the respective stages, based on our current experiences has been given below by their detailed breakdown in Tables A.4 and A.5 below.

TABLE A.4. COMPONENTS OF THE COST OF STATISTICAL OPERATIONS DEPENDENT ON THE NUMBER OF SECOND STAGE UNITS

size of plot clusters (p)	cost per plot-cluster in hours					
	selection	listing of sample plots	area extraction	calculation of crop acreages	tabulation of crop proportion	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	.0333	.0100	—	—	.0530	.0963
2	.0333	.0200	.2500	.0400	.0530	.3963
3	.0333	.0300	.3750	.0600	.0530	.5513
5	.0333	.0500	.6250	.1000	.0530	.8613
10	.0333	.1000	1.2500	.2000	.0530	1.6363
15	.0333	.1500	1.8750	.3000	.0530	2.4113
30	.0333	.3000	3.7500	.6000	.0530	4.7363

TABLE A.5. COMPONENT OF THE COST OF STATISTICAL OPERATIONS DEPENDENT ON THE NUMBER OF FIRST STAGE UNITS

size of village clusters (v)	cost per village cluster in hours				
	selection of sample units	listing of sample villages	allocation of 2nd stage units and handling of maps	estimation	total
(1)	(2)	(3)	(4)	(5)	(7)
1	.2500	.0833	.0500	.1736	.5569
2	.2500	.1667	.0875	.1736	.6778
4	.2500	.3332	.1750	.1736	.9318

REFERENCES

MAHALANOBIS, P. C. (1938) : Statistical Report on the Experimental Crop Census of 1937, Indian Central Jute Committee.

——— (1939) : Progress Report of the Jute Census Scheme for 1939, Indian Central Jute Committee.

——— (1941) : Report on the Sample Census of the Area under Jute in Bengal in 1940, Indian Central Jute Committee.

——— (1944) : On large scale sample survey. *Phil. Trans. Roy. Soc.*, 231, Series B (584), 329—451.

——— (1946) : Sample survey of crop yields in India. *Sankhyā*, 7(3), 269-280.

——— (1946) : Recent experiments in statistical sampling in the Indian Statistical Institute. *J.R.S.S.* 109(4), 325-378.

Paper received : April, 1967.

Revised : January, 1968.