THEORETICAL STATISTICS

Chairman: PROF. S. N. BOSE

6th January, 1939.

[Joint Meeting with the Section of Mathematics and Physics of the Indian Science Congress]

A SAMPLE SURVEY OF THE ACREAGE UNDER JUTE IN BENGAL

P. C. MAHALANOBIS

INTRODUCTION

The object of this paper is to give a brief account of certain recent investigations regarding the application of the method of random samples for estimating the acreage under jute in Bengal. It will be useful to mention certain basic facts in this connexion. It is estimated that during ten years from 1927-28 to 1936-37 the average world consumption of jute was a little over ten million or one crore bales[1] per year ; in the period of depression the consumption fell to as little as eight million bales and in a period of prosperity went up to twelve million bales per year. Roughly 42 per cent of the annual production is exported out of India in the form of raw fibre, and about 45 per cent in the form of manufactures. The total value of Jute exports on an average amounted to 50·3 crores of rupees or about 24 per cent of the total value of all exports during the period 1927-28 to 1936-37. Jute along with cotton thus form the two most important individual items in the export account of India.[2] As regards production it is estimated that on an average about 85 per cent is grown in Bengal.

2. The importance of preparing accurate forecasts of the jute crop will therefore be easily realised. This question has engaged the attention of administrators and business men for a long time. The present forecasts are published by the Director of Agriculture, Bengal, on the basis of information supplied to him by District Collectors who receive, through subordinate officers, returns from presidents of Union Boards. In each Union (which usually comprises from 10 to 20 *mauzas* or villages covering roughly from 10 to 20 sq. miles) a detailed list is prepared showing the name of each individual cultivator and the area sown by him with jute. These lists are supposed to be actually checked on the field but the results are extremely unreliable. The Bengal Jute Enquiry Committee of 1934 noted that:—

"It is known, however, that the checking process is too great a task for the Union Boards and that the estimates furnished are usually mere guesses which are always conservative and in most cases far from truth" (Majority Report, p. 11).

[1] One bale contains 5 maunds of jute. (1 maund = 82·284 lbs.).

[2] The price of raw jute varies widely from year to year, and also from month to month within the same year ; and naturally depends on the quality of the jute as well as on the world demand. The average money value of the crop has been estimated at thirty three crores of rupees per year during the period 1915-1938; in boom years the value may exceed eighty or ninety crores.

3. In recent years the average area under jute in Bengal has been roughly of the order of two and a quarter million acres or about 3500 sq. miles[3]. As the size of the plot in Bengal is rather small (less than half an acre) the number of plots on which jute is grown is very large. These plots do not also occur in compact blocks ; but are scattered over the whole province. Confining our attention to the 19 districts in which jute is grown in any appreciable quantity it is estimated that about six million of plots under jute are scattered among over 80 million of plots of all kinds covering an area of about 55,000 sq. miles. In a complete census the 5 or 6 million of plots under jute must be identified and this requires that each of the 80 million plots must be examined individually. After the field work is over the actual area of each plot under jute must be copied from revenue records ; and the final results will have to be obtained by direct addition. In organizing the field work it will have to be kept in mind that actual enumeration can start only after the cessation of sowing and must be completed before the beginning of harvesting. Jute being an exceptionally short crop this allows only two months to complete the work in any given area. A complete census is therefore bound to be expensive and difficult to organize.[4]

USE OF THE SAMPLE SURVEY

4. In March 1936, in reply to an enquiry from the Secretary, Imperial Council of Agricultural Research, I suggested exploring the possibilities of random sampling methods for estimating the area under different crops in Bengal.

5. The basic principles of random sample surveys may be briefly explained at this stage. In this plan the whole of the area to be surveyed (in this case roughly 55,000 sq. miles) will be divided into a number of zones of suitable size. The size of each zone need not be exactly the same, but it is desirable that each zone should be as homogeneous as possible in regard to the intensity of cultivation, that is, the proportion of land under jute. A number of points are then selected strictly at random within each zone. At each of these points a sampling unit (which may be conveniently called a "grid") of a suitable size, of the order of say 1, 4, 16 or 40 acres, is surveyed in detail. In this way the proportion under jute in each "grid" can be determined. If we have a large number of "grids" in each zone, the average proportion calculated from the "grids" within the zone can be taken to be the representative figure for the zone. Multiplying by the total area of each zone (which is known) it is then possible to estimate the area under jute in each zone. Adding the figures for the different zones, the total area under jute in each district or in the whole province can be then easily obtained. A sample survey will thus involve only a fraction of the labour required for a complete census. Besides its lower cost, one great advantage of the sample method is the possibility of calculating the order of accuracy of the final estimate.

6. In 1937 a small scale experimental crop census was carried out with the help of a grant from the Indian Central Jute Committee. The field work was done under the control of the Director of Agriculture, Bengal ; while the design of the survey was prepared and the whole of the statistical analysis of the material was carried out in the

[3] The actual acreage under jute in Bengal fluctuated from 1·61 million acres in 1931 to 3·06 million acres in 1930 during the ten years 1929—1938.

[4] I may also remark in passing that the results are likely to be unreliable as was clearly demonstrated in the course of replicated enumeration in selected areas in 1937 and 1938.

Statistical Laboratory, Calcutta ; and a detailed statistical report[5] was submitted in August 1938. A more extensive scheme was prepared and carried out on the field in 1938, and the first report[6] was submitted in December 1938. I propose giving in this paper a brief description of the method of approach and certain tentative results based on these two exploratory surveys.

Statement of the Problem

7. From the statistical point of view our aim is to evolve a sampling technique which will give, for any given total expenditure, the highest possible accuracy in the final estimates. For this it is necessary to determine three things, namely,

(*a*) what is the best size of the sampling units ;

(*b*) what is the total number of such sampling units which should be used to attain the desired degree of accuracy in the final estimates ; and

(c) what is the best way of distributing these sampling units among the different districts, regions or zones covered by the survey.

It is clear that the above questions can only be answered in reference to the sampling errors and the cost of operations of the method used. This makes it necessary to study what may be called the Variance (or Error) Function and the Cost Function.

8. Let us suppose that we are using grids of size of x-acre ; and that we have, say, n such grids located at random in any given area of say, A square miles. By direct enumeration on the field we can obtain for each of these n grids a particular value of p, the proportion of land which is under jute within the grid. If $\bar{p}$ is the average of all observed values of p, then the mean value of $(p-\bar{p})^2$ is the variance for individual grids of size x-acre. This variance will clearly depend on the size of the grid and will decrease as the size is increased. The variance will also presumably depend on p the proportion of land under jute ; and we may write the Variance Function as $\psi(p, x)$ for grids of size x-acre in a homogeneous region with proportion of land under jute equal to p. The variance of the mean value $\bar{p}$ for the whole area A is then given by $\psi(p, x)/n$. It is more convenient to use a slightly different expression. Dividing n, the number of grids, by the area A square miles, we get the density or average number of grids per square mile which we may write as 'y'. We may then write the variance for the mean value of p as $\psi(p, x)/y$ per square mile.

9. We may now consider the Cost Function which will give the expenditure incurred in time or money for collecting information for grids of size x-acre located at random with a density of 'y' per square mile. The cost will clearly depend on both the size of grids (x) and their number per square mile (y) ; and we can write the Cost Function as $\phi(x, y)$ per square mile.

The Best Size of Sampling Units

10. Let us now consider a homogeneous region in which the intensity of cultivation, *i.e.* the value of p is more or less the same over the whole area. Let T be the total cost proposed to be incurred for the sample survey ; dividing the total cost T by the total

[5] *Statistical Report on the Exploratory Crop Census of 1937* published by the Indian Central Jute Committee in 1938.

[6] *First Report on the Crop Census of 1938* published by the Indian Central Jute Committee in 1939.

area say A square miles proposed to be covered by the survey, we get $t=T/A$ as the assigned cost per square mile. If we now put $\phi(x,y)=t$, we can get the value of y (for any given value of x) which we can afford to use at the given level of expenditure. If we substitute this value of y in the formula for variance, that is, in $\psi(p,x)/y$, we immediately obtain the corresponding value of the variance per square mile. Our problem then is to choose that value of x, i.e., that size of grids, for which (with any assigned value of t or the cost per square mile) the variance of the final estimate will be a minimum.

11. If the functional form $\psi(p,x)$ and $\phi(x,y)$ can be determined then we may use analytic methods to obtain the appropriate value of x (and necessarily also of y) for any assigned value of t. The value of x will of course, in general, depend on the value of t, the rate of expenditure. When the functional forms are not known but only discrete values are available we can use numerical methods for comparing the efficiency of different sizes of grids.

GENERAL SOLUTION IN ABSTRACT FORM

12. We may now state the problem more generally. Let us assume that the whole area to be surveyed has been sub-divided into k zones each of which is more or less homogeneous in regard to the intensity of cultivation, that is, the proportion of land under jute. Let us write:—

A_i = area of the i-th zone in sq. miles ... ... (1·1)

p_i = proportion of land under jute in the i-th zone ... ... (1·2)

$q_i = (1-p_i)$

x_i = size of grids used in the i-th zone ... ... (1·3)

y_i = density or number of grids per sq. mile in the i-th zone ... (1·4)

We also assume that the Variance Function in the i-th zone is given by

$$v_i = \psi(p_i, x_i) \quad \ldots (2)$$

and the Cost Function in the i-th zone is given by

$$t_i = \phi(x_i, y_i) \quad \ldots (3)$$

13. The area under jute in the i-th zone is then given by $(A_i p_i)$ and since $A_i y_i$ is the total number of grids used in the i-th zone, the variance of the estimate of area under jute for the i-th zone is given by

$$V_i = \frac{A_i v_i}{y_i} = \frac{A_i \cdot \psi(p_i, x_i)}{y_i} \quad \ldots (4)$$

The final estimate of the area under jute for the whole area is given by $\Sigma(A_i p_i)$, and, assuming that the form of the Variance Function is identical in all zones, the variance of this final estimate may be written

$$V = \sum_{i=1}^{i=k} (V_i) = \sum_{i=1}^{i=k} \frac{A_i \cdot \psi(p_i, x_i)}{y_i} \quad \ldots (5)$$

14. Again the cost for the i-th zone is clearly given by $A_i . \phi(x_i, y_i)$; and, if we assume that the form of the cost function is the same in every zone, the total cost may be written as

$$T = \sum_{i=1}^{i=k} A_i . \phi (x_i, y_i) \qquad \ldots (6)$$

Our problem then is to choose (x_i, y_i) for each zone in such a way that V as defined in equation (5) is a minimum subject to the total cost T having any assigned value in equation (6). Differentiating with respect to x_i's and y_i's we have for minimum V and fixed T

$$\delta V = \sum_{i=1}^{i=k} \left[\frac{A_i}{y_i} . \frac{\partial \psi}{\partial x_i} . \delta x_i - \frac{A_i . \psi}{(y_i)^2} . \delta y_i \right] = 0 \qquad \ldots (7)$$

$$\delta T = \sum_{i=1}^{i=k} \left[A_i . \frac{\partial \phi}{\partial x_i} \delta x_i + A_i . \frac{\partial \phi}{\partial y_i} \delta y_i \right] = 0 \qquad \ldots (8)$$

Using Lagrange's method of undetermined multipliers we get

$$\left[\frac{1}{y_i} . \frac{\partial \psi}{\partial x_i} + \lambda \frac{\partial \phi}{\partial x_i} \right] = 0 , \qquad (i=1, 2,\ldots k) \quad \ldots (9)$$

$$\left[\frac{1}{(y_i)^2} . \psi - \lambda \frac{\partial \phi}{\partial y_i} \right] = 0 , \qquad (i=1, 2,\ldots k) \quad \ldots (10)$$

15. Solving these equations we can theoretically obtain the optimum set of values of (x_i, y_i) for each zone for any assigned value of the total cost. The solution in any concrete case will naturally depend on the form of the Variance and Cost Functions ; and in actual practice it may not be possible to obtain exact algebraic expressions. Numerical or semi-graphical methods may, however, be always used to obtain approximate solutions. It is thus possible to form a Cost-Error Table showing size and density of grids and error of the final estimate for different values of the total expenditure. From this Table we can then pick up the best size (and best density) of grids for any assigned level of expenditure.

16. One other point is worth noting in the present connexion. If we assume the cost to be proportional to the total number of sampling units, we may ignore the question of cost and keep constant the total number of sampling units. If Y is the total number of samples we may write :—

$$Y = \sum_{i=1}^{i=k} (y_i) = \text{constant} ; \text{ and } \delta Y = \sum_{i=1}^{i=k} (\delta y_i) = 0 \qquad \ldots (8{\cdot}1)$$

Minimizing V subject to this equation (8·1), and writing ψ_i for $\psi (p_i, x_i)$ we easily get in terms of the Lagrangian multiplier λ :—

$$\frac{A_i \psi_i}{(y_i)^2} - \lambda = 0 \qquad (i=1, 2,\ldots k) \quad \ldots (9{\cdot}1)$$

Using $Y=\Sigma(y_i)$ we immediately obtain the usual formula for stratified sampling :—

$$y_i = \frac{Y\sqrt{(A_i\psi_i)}}{\Sigma\sqrt{(A_i\psi_i)}} \qquad \ldots \quad (10{\cdot}1)$$

GENERAL CONSIDERATIONS

17. The mathematical theory in abstract form is simple. The basic ideas underlying the theory are also easy to understand from general physical considerations. Let us first consider the Cost Function. This will depend mainly on the amount of time taken for the actual field survey, as supervision and other expenses in the final analysis may be expressed in terms of the number of hours of work done by primary investigators. Now the total time spent by the primary field investigators may be broadly divided under four heads :

(*a*) *Enumeration,* which includes the time spent in identifying and inspecting each of the plots included within the sample grids by reference to large scale village maps ($16''=1$ mile) ; and noting which of them is under jute.

(*b*) *Journey,* which includes the time spent in going from camp to camp (camp being defined as a place where the night is spent) ; from camp to a sample unit ; and from sample unit to camp. This therefore covers all journeys undertaken for the purpose of carrying out the field enumeration.

(*c*) *Miscellaneous,* which includes the time spent for making all preliminary arrangements, copying the field record and despatching the same to Headquarters ; receiving instructions etc.

(*d*) *Indirect,* which consists of the time spent in sleep, rest, food etc.; and includes wastage on account of sickness, holidays, cessation of work owing to drought, excessive rain or unforeseen reasons. The indirect time in the case of primary investigators thus represents the total time spent for indirect or non-productive purposes.

18. The time required for actual enumeration may be expected to increase with the size of the grid ; and for the same size of the grid may be expected to increase more or less proportionately to y, the density or number of grids per sq. mile. The time spent in miscellaneous productive work will also probably increase with y, the density or number of grids per sq. mile.

19. It is also easy to see in a general way how the journey time is likely to behave. Let us suppose that n sampling units are scattered at random within any given area ; and let us assume that we may treat each such sample unit as a geometrical point. We may also assume that arrangements will usually be made to move from one sample point to another in such a way as to keep the total distance travelled as small as possible ; that is, we may assume that the path traversed in going from one sample point to another will follow a straight line. In this case it is easy to see that the mathematical expectation of the total length of the path travelled in moving from one sample point to another will be $(\sqrt{n}-1/\sqrt{n})$. The cost of the journey from sample to sample will therefore be roughly proportional to $(\sqrt{n}-1/\sqrt{n})$. When n is large, that is, when we consider a sufficiently large area, we may expect that the time required for moving from sample to sample will be roughly proportional to $\sqrt{n}$, where n is the total number of samples in the given area. If we consider the journey time per sq. mile, it will be roughly proportional to $\sqrt{y}$, where y is the density of number of sample units per sq. mile.

20. As regards the time spent for indirect or non-productive purposes, this is likely to be more or less constant. This was what we actually found to be the case during the field work in 1938; about 70 per cent of the time was required every day for indirect purposes. This means that only about 30 per cent or a little over seven hours per day are actually available for net productive work.

21. It is now easy to see in a general way the nature of the Cost Function. For example, if the sample units are very widely scattered, that is, if we work with a comparatively small total number of grids, a great deal of time will be consumed in travelling from one grid to another. In this case it will be probably economical to use grids of rather large size as with only a little additional expenditure of time it will be then possible to collect information for a much greater number of plots. On the other hand, when the number of grids is very large, the fraction of the time spent in travelling will be less; but in this case even a small increase in the size of each grid is likely to increase the cost per sq. mile very appreciably. By using a small number of sampling units, that is, by scattering them very wide, the cost may be kept low, but a large fraction of the time will be wasted in moving from one sample to another. On the other hand, by using a high density of sampling units, the wastage in travelling from one sample to another will be small, but the total cost will be large.

22. We must now consider errors of sampling. It is true that by increasing the size of the grid the sampling error will be reduced; but increasing the size of the grid means a good deal of additional expenditure in case we keep the total number or the density of grids the same. If we reduce the number of grids, that is, scatter them more widely, we shall be wasting proportionally a greater amount of time in travelling from one grid to another. In this situation the real question is whether the reduction in the variance or the sampling error brought about by an increase in the size of the grids will or will not be offset by the increase in the expenditure. It will be seen therefore that the best size of grids can only be settled by taking into consideration the Variance (or Error) Function in conjunction with the Cost Function. This is just what the mathematical theory given above is intended to achieve.

Nature of the Variance Function

23. With a view to studying the Variance Function, that is, how the sampling error changes with the size of the sampling unit, we had arranged to obtain during the experimental crop census of 1937 detailed information regarding each individual plot in an area of about 124 sq. miles. The primary investigators went to the field with large scale (16 inches=1 mile) official village maps used for settlement and revenue purposes (on which the position of each plot or revenue holding is marked) and recorded the name of the crop grown on each individual plot by direct inspection on the field. The area (in acre) of each plot was subsequently obtained from the *Khatian* or official revenue records. Later on in the Laboratory we made arrangements for marking grids of various sizes directly on the village maps. From the knowledge of the crop grown in each individual plot (or fraction of a plot) included within each grid, it was easy to obtain the value of p or the proportion under jute (or of other crops) included within the grid. The observed variance for each size of grid could be then directly worked out.

24. A large number of model sampling experiments were carried out in this way in May and June 1938 in the case of thana Deganga for 108 villages comprising 1,71,250 plots covering 49,920 acres or 78 sq. miles in district 24-Parganas. The observed values of

variance[7] of p for different sizes of sampling units are given in Table (1). This material suggested the following form for the Variance Function :—

$$v_i = \psi(p_i, x_i) = \frac{p_i q_i}{(b_i x_i)^{g_i}} \quad \ldots (2{\cdot}1)$$

$$\text{or :} \qquad g_i \log (b_i x_i) = \log \left(\frac{p_i q_i}{v_i}\right) \quad \ldots (2{\cdot}2)$$

where p is the proportion of land under jute, $q=(1-p)$; and "b" and "g" are statistical parameters, values of which were obtained by least square methods in the logarithmic form (2·2), and are given in Table (1).

TABLE 1. VARIANCE FUNCTION : OBSERVED AND GRADUATED VALUES

Size of Grids (in acres)	24 PARGANAS : THANA DEGANGA (1937)				MYMENSINGH : THANA ISWARGANJ (1938)			
	Number of Grids	$\log_{10}$ (variance)		Chi-square	Number of Grids	$\log_{10}$ (variance)		Chi-square
		Observed	Graduated			Observed	Graduated	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1·00	559	1·0655	1·0409	0·8965	1,756	0·9509	0·9645	0·8520
2·25	...	...	...	...	1,377	1·0900	1·0834	0·1596
4·00	316	1·0716	1·1745	8·8647	1,161	1·1812	1·1678	0·5503
5·00	809	1·2055	1·1960	0·1951	...	...	...	...
6·25	...	...	...	...	951	1·2389	1·2332	0·0795
9·00	...	...	...	...	832	1·2968	1·2867	0·2248
12·25	...	...	...	...	752	1·3424	1·3319	0·2142
16·00	215	1·3478	1·3080	0·8987	618	1·3778	1·3711	0·0726
25·00	...	...	...	...	476	1·4001	1·4366	1·6749
36·00	213	1·3979	1·3862	0·0779	365	1·4657	1·4900	0·5747
	$p=0{\cdot}0193$		Chi-square = 10·9329		$p=0{\cdot}3712$		Chi-square = 4·4026	
	$b=0{\cdot}00085$		D.F. = 3		$b=9{\cdot}73$		D.F. = 7	
	$g=0{\cdot}2219$		P = 0·0122		$g=0{\cdot}3377$		P = 0·7315	

In this Table (1) col. (1) shows the size of grids ; col. (2) the number of grids used in the model sampling experiments ; col. (3) the logarithms of observed variances, and col. (4) corresponding values of the logarithms of variances calculated from equation (2·2). The next column (5) gives the value of chi-square for the comparison of the observed and

[7] The standard deviations were given in the *Statistical Report on the Crop Census of 1937* printed by the Indian Central Jute Committee in 1938.

calculated values of the logarithms of variances.[8] As judged by the chi-square test the graduation is encouraging ; the probability of occurrence of a system of deviations equal to or greater than that actually observed was greater than 12 per cent.

25. During the field survey of 1938 further material was collected for 824 villages in 8 thanas comprising 4,78,639 plots covering a total area of 2,64,855 acres or about 414 sq. miles. Extensive model sampling experiments were carried out in the case of thana Iswarganj (in district Mymensingh) which included 340 villages comprising 145,557 individual plots covering a total area of 77,675 acres or 121·4 sq. miles. Altogether 8,268 grids of 9 different sizes (1, 2·5, 4, 6·25, 9, 12, 16, 25 and 36 acre) were used in this investigation ; and in each case the samples were collected in a number of instalments so as to supply material for independent comparisons for each size of grid. The results for Iswarganj thana are given in cols. (6)-(9) of Table (1) from which it will be seen that the exponential formula gives excellent graduation.

26. One or two comments on the Variance Function will not be out of place. The parameter 'b' is a dimensional constant which probably represents something like the average value of the reciprocal of the smallest unit of the area on which jute is sown. Thus when the size of the sampling unit is equal to this constant, that is, $x=1/b$, we have $bx=1$, and the Variance Function automatically reduces to the binomial distribution as it should.

27. Secondly, the parameter 'g' represents the association between neighbouring plots in regard to the occurrence of jute. In fact if we write ρ_x as the average value of the intra-class correlation of p (the proportion of land under jute) between all the plots included within grids of size x-acre, then it is easy to see that

$$\rho_x = \frac{(bx)^{1-g}-1}{(bx)-1} \qquad \ldots (11)$$

This gives the relation between 'g' and ρ_x . The form of the Variance Function adopted by us thus apparently requires a change of the intra-class correlation ρ_x with x in accordance with equation (11).

28. We may also consider briefly the question of the best size of the grids in the light of the empirical knowledge we have gained regarding the Variance Function. We may, therefore, now write the variance of the i-th zone in the form :—

$$V_i = \frac{A_i\, p_i q_i}{(b_i x_i)^{g_i}} \cdot \frac{1}{y_i} \qquad \ldots (4{\cdot}1)$$

This gives us immediately the variaunce of the final estimate as

$$V = \sum_{i=1}^{i=k} \frac{A_i\, p_i\, q_i}{(b_i\, x_i)^{g_i}} \cdot \frac{1}{y_i} \qquad \ldots (5{\cdot}1)$$

with the $2k$ equations in x_i's and y_i's, in the form :—

$$\left[\frac{p_i\, q_i\, g_i}{y_i (b_i)^{g_i}\ (x_i)^{g_i+1}} + \lambda \frac{\partial \phi}{\partial x_i}\right] = 0 \quad (i=1,\ 2,\ 3\ .\ .\ k) \qquad \ldots (9{\cdot}1)$$

[8] The standard error in this case is simply inversely proportional to the square root of the size of the sample.

$$\left[\frac{p_i q_i}{(y_i)^2 (b_i)^{g_i} (x_i)^{g_i}} + \lambda \frac{\partial \phi}{\partial y_i}\right] = 0 \quad (i = 1, 2, \ldots k) \quad \ldots (10{\cdot}1)$$

Eliminating λ between equations (9·1) and (10·1), we get

$$x_i \frac{\partial \phi}{\partial x_i} = g_i y_i \frac{\partial \phi}{\partial y_i} \quad \ldots (12{\cdot}0)$$

where x_i and y_i have values corresponding to minimum V at the given cost.

29. Two special cases are of some interest in the present connexion. Let us consider a homogeneous region in which p is sensibly constant ; in this case we can drop the suffix 'i' in all the terms in equations (9·1) and (10·1). When $g = 1$, or the Variance Function reduces to the normal (or binomial) distribution, we get the simple relation :—

$$x \frac{\partial \phi}{\partial x} = y \frac{\partial \phi}{\partial y} \quad \ldots (12)$$

where x and y have the values corresponding to minimum V at the given cost. If in addition the Cost Function is simply proportional to the total area covered in the sample survey, that is, $\phi(x,y)$ is of the form constant $\times (xy)$, we find that all sizes of grids have the same efficiency.

30. On the other hand, if we assume that the Cost Function is of this special form, *i.e.* $T = \phi(x,y) = C.(xy)$, but the Variance Function is not normal (or binomial) then the variance of the final estimate is given by

$$V = \frac{A^2 C}{(b)^g} \cdot \frac{pq}{T} (x)^{1-g} \quad \ldots (5{\cdot}2)$$

where T is the total cost of the sample survey, and C is the constant of proportionality in the Cost Function. Since $0 \leqq g \leqq 1$, it will be seen that V will decrease as x is decreased. That is, the smaller the size of the sampling unit the greater will be the accuracy of the final estimate for any assigned value of the total expenditure. Finally, if in addition $g = 1$, or the Variance Function is of the normal (or binomial) form, then the variance of the final estimate reduces to :—

$$V = \frac{A^2 C p q}{T} \quad \ldots (5{\cdot}3)$$

The accuracy of the estimate is independent of the size of the sampling unit ; and the variance is inversely proportional to the total cost ; or the available 'information' (in the Fisherian sense) is proportional to the total cost.

ACCURACY OF SAMPLE SURVEYS

31. It will be convenient at this stage to consider briefly the question of accuracy of sample surveys. I am giving in Table (2) the results of a comparison of estimates of the area under jute based on model sampling experiments with different sizes of sampling units and "true" values obtained by complete enumeration in thana Deganga in 1937. This Table gives for each Union (administrative unit comprising roughly from 10 to 20

villages and covering an area from 10 to 20 sq. miles) the values of the 't'-statistic together with the corresponding degrees of freedom.

TABLE 2. VALUES OF t-STATISTIC FOR COMPARISON OF ESTIMATES OF AREA UNDER JUTE BY USING SAMPLING UNITS OF DIFFERENT SIZES IN DISTRICT 24 PARGANAS, THANA DEGANGA IN 1937

Union	Random Plot		1-acre grid		4-acre grid		5-acre grid		16-acre grid		36-acre grid	
No.	D. F.	t	D. F.	t	D. F.	t	D. F.	t	D. F.	t	D. F.	t
(1)	(2·1)	(2·2)	(3·1)	(3·2)	(4·1)	(4·2)	(5·1)	(5·2)	(6·1)	(6·2)	(7·1)	(7·2)
1	570	0·86	38	1·00	21	0·75	56	1·18	14	0·13	14	0·04
2	980	1·62	61	1·72	33	0·26	88	1·07	23	0·85	22	1·97
3	552	0·21	40	0·17	23	0·19	51	0·71	15	2·30*	14	0·62
4	703	0·12	44	0·26	25	0·23	58	0·57	16	1·57	18	0·28
5	804	1·47	51	0·73	31	0·42	78	0·18	20	0·81	18	0·33
6	1047	0·41	59	0·85	31	0·85	98	3·28**	22	0·76	22	0·02
7	1123	2·61**	63	1·23	39	0·80	111	3·51**	23	1·20	21	0·11
8	845	1·56	67	2·18*	35	...	89	0·29	23	0·39	25	0·30
9	1010	0·36	68	0·47	35	1·32	93	0·93	25	3·40	25	0·76
10	910	0·44	58	...	33	...	78	2·76**	24	0·56	24	0·71
Whole Thana	8544	1·22	549	0·003	306	1·47	800	0·53	205	1·18	203	0·04

In Table (2) observed values of 't' which exceed the five per cent. level of significance, have been marked by a single star, and those which exceed the one per cent. critical limit by two stars following the usual convention. It will be noticed that out of 63 values of the 't'-statistic, six exceed the critical value at the 5 per cent level of significance. The agreement between the results obtained by the sampling method and by complete enumeration is not therefore unsatisfactory.

32. We may also compare the results of actual sample surveys carried out in the field during the 1938 season. Table (3) gives the values of the 't'-statistic for testing

TABLE 3. VALUES OF t-STATISTIC FOR COMPARISON OF ESTIMATES OF AREA UNDER JUTE BY USING SAMPLING UNITS OF DIFFERENT SIZES IN 1938

Sampling Methods	Murshidabad		Mymensingh		Pabna		Rajshahi		Rangpur		Tipperah	
	D. F.	t	D. F.	t	D. F.	t	D. F.	t	D. F.	t	D. F.	t
(1)	(3·1)	(3·2)	(4·1)	(4·2)	(5·1)	(5·2)	(6·1)	(6·2)	(7·1)	(7·2)	(8·1)	(8·2)
1-acre and 4-acre	278	0·065	2148	1·62	570	2·27*	...	...	939	1·86	1194	0·403
1-acre and 16-acre	211	0·077	1812	1·74	403	0·138	323	0·925	807	1·32	1047	0·543
4-acre and 16-acre	83	0·054	484	1·13	183	0·788	...	...	254	0·199	233	0·353

the significance of differences between independent estimates of the acreage under jute obtained from sample surveys carried out in the field in 1938 with grids of different sizes. It will be noticed that out of 16 values of the 't'-statistic given in this table, in only one

case the observed value of 't' is greater than two; in fact, in most cases the observed value is less than one.

33. We may also compare the "true" value of the area under jute based on complete enumeration with the estimates obtained from sample surveys. The values of the 't'-statistic for such comparisons are shown in Table (4).

34. We have in Tables (2)-(4) the results of 192 comparisons; on the 5 per cent level we therefore expect that in 5 per cent of cases, or say in 10 instances, the observed differences should exceed the five per cent critical limit; we also expect that only two of the differences should exceed the one per cent limit. In actual fact, however, 16 or 17 of the observed values of 't' exceed the 5 per cent limit; and 6 the one per cent limit. This is higher than the theoretical values, but is of the same order.[9] The agreement is therefore not unsatisfactory; and properly organized sample surveys may be expected to give results with limits of accuracy of the order expected on theoretical grounds.

OBSERVED VALUES OF THE COST FUNCTION

35. We may now consider the Cost Function again. In preparing the plan for the field work in 1938 it was our intention to collect information regarding the cost of operations for different combinations of size and density of grids. Owing to various administrative difficulties the field work was however delayed, and actual enumeration was started at the end of July when the jute season was well advanced. Unfortunately early and exceptionally heavy floods occurred in most of the districts selected for the sample survey; some of the villages which were dry at the time of starting the work gradually went under water so that the field survey had to be suspended or abandoned in many cases. This caused a great deal of delay and retarded progress; and it was not possible to complete our original programme. The actual amount of field work done is indicated below:

(*i*) Complete enumeration of 824 villages in 8 thanas comprising 4,78,639 plots covering about 2,64,855 acres or 414 sq. miles;

(*ii*) individual plots selected at random with densities of 3, 6, 9, 12 and 15 plots per sq. mile in 243 villages comprising 2,540 plots altogether;

(*iii*) 1-acre grids with densities of 2, 4, 6, 8 and 10 grids per sq. mile in 923 villages comprising 6,339 grids (39,669 plots);

(*iv*) 4-acre grids with densities of 2, 3, 4, 5 per sq. mile in 368 villages comprising 1221 grids (20,086 plots); and

(*v*) 16-acre grids with densities of 2 and 3 per sq. mile in 106 villages comprising 328 grids (18,956 plots).

[9] Observed values in excess of the critical limits appear to be roughly double the expected number. The expected values of the 't'-statistic have been, of course, obtained on the assumption of a normal distribution of the variate in the parent population. It is not, however, clear how far this assumption is true in the present case; and the observed excess of large values of 't' may be partly due to a real difference in the relevant distribution function; it may also be partly due to errors of recording.

TABLE 4. VALUES OF *t*-STATISTIC FOR COMPARISON OF ESTIMATES OF AREA UNDER JUTE BY USING SAMPLING UNITS OF DIFFERENT SIZES AND COMPLETE ENUMERATION IN 1938.

Size of Grids	Union Number	Degrees of Freedom	't'
(1·1)	(1·2)	(1·3)	(1·4)
MYMENSINGH : Iswarganj			
1-acre	2	156	0·05
	6	109	0·16
	7	147	1·09
	8	136	2·24*
	9	137	2·36*
	10	128	1·31
	12	147	1·29
	15	124	0·17
	16	127	1·08
	17	161	1·10
	18	107	0·75
	19	112	0·70
	20	119	1·12
	Thana	1710	0·59
4-acre	2	42	2·06*
	6	25	0·27
	7	32	1·34
	8	65	1·53
	9	43	0·08
	10	43	1·33
	15	21	0·27
	16	18	0·82
	17	59	1·85
	18	24	0·72
	20	28	0·17
	Thana	400	0·18
16-acre	2	12	0·01
	8	10	0·54
	9	18	0·97
	10	18	0·64
	18	4	0·49
	20	5	0·51
	Thana	67	0·77
RAJSHAHI : Badalgachhi			
Random Plots	5	17	0·47
	6	3	1·17
	Thana	20	0·82
1-acre	1	29	0·92
	5	11	0·52
	Thana	40	1·01
16-acre	1	8	0·43
	5	2	0·75
	Thana	10	0·62

Size of Grids	Union Number	Degrees of Freedom	't'
(2·1)	(2·2)	(2·3)	(2·4)
TIPPERAH : Laksham			
Random Plots	1	56	1·23
	3	51	0·97
	8	37	0·37
	9	76	0·19
	Thana	2·20	1·02
1-acre	1	89	1·39
	3	61	1·83
	6	50	2·36*
	7	52	1·05
	8	33	1·27
	9	62	0·48
	Thana	347	3·35*
4-acre	1	23	1·70
	3	9	0·89
	7	12	1·62
	8	5	3·00*
	9	10	0·38
	Thana	59	2·15*
16-acre	1	6	1·57
	7	9	0·65
	Thana	15	1·67
PABNA : Sara			
Random Plots	3	54	1·73
	5	85	0·86
	Thana	139	1·75
1-acre	2	33	1·69
	3	29	22·60*
	5	33	1·52
	6	70	0·15
	Thana	165	0·09
4-acre	2	12	1·18
	3	10	2·14
	5	28	1·28
	6	24	0·06
	Thana	74	2·52*
16-acre	3	6	0·11

Size of Grids	Union Number	Degrees of Freedom	't'
(3·1)	(3·2)	(3·3)	(3·4)
RANGPUR : Domar			
Random Plots	2	11	0·23
	4	24	0·41
	8	20	0·67
	Thana	55	0·30
1-acre	2	27	0·45
	4	56	1·34
	6	12	0·48
	8	17	0·00
	9	46	1·22
	Thana	158	0·91
4-acre	2	6	0·24
	4	22	0·98
	8	2	0·75
	9	12	0·36
	Thana	42	0·91
16-acre	2	3	0·18
	4	8	0·01
	9	1	7·18
	Thana	12	0·34
DACCA : Kaliganj			
1 acre	4	66	0·18
	7	60	1·16
	Thana	126	0·90
MURSHIDABAD : Beldanga			
1-acre	1	56	0·56
	3	20	1·17
	4	23	0·71
	5	17	0·40
	8	1	0·08
	9	11	0·10
	Thana	128	1·20
4-acre	1	6	0·43
	3	12	2·85*
	4	5	0·55
	5	8	0·26
	9	2	0·62
	Thana	33	1·33
16-acre	3	5	0·03

36. The available information was not sufficient to enable the question of the form of the cost function being investigated with any chance of success. In this situation I decided to pool together the information collected partly on the field in 1937, partly in the Statistical Laboratory, and partly during the field work of 1938. In this way by combining all available information and using free-hand graduation to smooth the data it was possible to construct a table showing the cost for different sizes and densities of sampling units. The relevant material is given in Table (5) in which col. (1·1) gives the

TABLE 5. ESTIMATED COST IN RUPEES PER SAMPLE AND PER SQUARE MILE

Size of grid	Density per Sq. mile	Number of units on the Field	Cost per Sample			Cost per Square Mile		
			Laboratory	Field	Total	Laboratory	Field	Total
(1·1)	(1·2)	(1·3)	(2·1)	(2·2)	(2·3)	(3·1)	(3·2)	(3·3)
Random Plot	3	289	0·143	0·324	0·467	0·429	0·972	1·401
	4	.....	0·139	0·320	0·469	0·556	1·280	1·836
	6	242	0·137	0·310	0·447	0·822	1·860	2·682
	9	1242	0·136	0·212	0·348	1·224	1·908	3·132
	12	666	0·135	0·191	0·326	1·620	2·292	3·912
	15	267	0·132	0·164	0·296	1·980	2·460	4·440
1-acre	1	.....	0·598	0·996	1·594	0·598	0·996	1·594
	2	183	0·563	0·879	1·442	1·126	1·758	2·884
	3	.....	0·545	0·765	1·310	1·635	2·295	3·930
	4	385	0·527	0·689	1·216	2·108	2·756	4·864
	5	.....	0·516	0·557	1·073	2·580	2·785	5·365
	6	847	0·502	0·464	0·966	3·012	2·784	5·796
	8	754	0·470	0·400	0·870	3·760	3·200	6·960
4-acre	1	83	0·658	1·188	1·846	0·658	1·188	1·846
	2	228	0·641	1·080	1·721	1·282	2·160	3·442
	3	296	0·612	0·915	1·527	1·836	2·745	4·581
	4	295	0·601	0·851	1·452	2·404	3·404	5·808
	5	39	0·591	0·809	1·400	2·955	4·045	7·000
5-acre (1937)	12	890	1·005	0·421	1·386	11·580	5·052	16·632
16-acre	1	47	1·407	1·561	2·968	1·407	1·561	2·968
	2	132	1·351	1·341	2·692	2·702	2·682	5·384
	3	64	1·332	1·096	2·428	3·996	3·288	7·284
36-acre (extrapolated)	3	.....	2·657	2·250	4·907	7·971	6·750	14·721

size, col. (1·2) the desity per sq. mile, and col. (1·3) the number of sampling units for which information was collected on the field for each size of sampling units. The next three columns give the estimated cost in rupees per sampling unit ; thus col. (2·1) shows the cost of the laboratory portion of the work, namely, the cost for locating and marking of the sample units on the maps, listing of plots and other preliminary arrangements, and for tabulation and analysis of the material collected on the field ; col. (2·2) shows the cost of the field work ; and col. (2·3) the total cost in rupees per sampling unit. The next three columns (3·1), (3·2) and (3·3) give the corresponding cost per sq. mile which is more useful for our present purposes.

37. Information is generally meagre for random plots and for some of the densities for 1-acre and 4-acre ; the material for 16-acre is particularly deficient while no information for the field portion of the work is available for 36-acre for which the cost figure had to be obtained by extrapolation. The material is also definitely heterogeneous and of varying reliability in different parts. It must be remembered therefore that I am using it only for

purposes of illustration, and I should like to emphasise here that much significance should not be attached to particular results.

38. It will be noticed that so far as the laboratory portion of the work is concerned there is practically no increase in the cost per sampling unit with the increase in density of grids. There is however a marked increase in the cost with an increase in the size of grids. This is of course just what is to be expected ; since the locating of sample points at random, marking of sample units on the maps, listing of revenue plots, preparation of crop schedules, tabulation, addition, checking etc., vary almost directly as the total number of plots covered. As the cost for the laboratory portion of the work on a sample basis is practially constant, the cost per sq. mile naturally increases linearly with increasing density.

39. In the field portion of the work, on the other hand, it will be seen that the cost per sample decreases quite rapidly as the density is increased, owing to the saving in the time of identifying the sample plots and in moving from one sample to another. Consequently the cost per sq. mile increases more slowly than the increase in density.

40. As already noted no attempt was made to obtain any mathematical expression for the form of the cost function owing to the meagreness of the data. We may however proceed by numerical methods to compare the relative efficiency of sampling units of different sizes. Suppose, for example, that it is desired to spend three rupees per sq. mile (which will amount to rupees one lakh and sixtyfive thousand altogether for a full scale provincial survey covering 55,000 sq. miles). By numerical interpolation in Table (5), we find that with this expenditure we can use random plots with a density 7·9 per sq. mile, or 1-acre grids with a density 2·1 per sq. mile, 4-acre grids with a density of 1·7 per sq. mile, or 16-acre grids with density of 0·64 per sq. mile. In the same way, we may find the density of grids which we can afford to have for different sizes of grids for other rates of expenditure. Such interpolated values of the density or number of sampling units per sq. mile for different sizes are shown in Table (6).

TABLE 6. INTERPOLATED VALUES OF THE DENSITY OF SAMPLING UNITS PER SQUARE MILE (y) FOR DIFFERENT RATES OF EXPENDITURE

Expenditure per square mile		Random Plots	1-Acre	4-Acre	16-Acre	36-Acre
(1)		(2)	(3)	(4)	(5)	(6)
Rs.	As.					
1	0	2·0	0·6	0·5	0·3	0·21
1	8	3·2	0·9	0·8	0·5	0·32
2	0	4·5	1·3	1·1	0·7	0·43
2	8	6·1	1·7	1·4	0·9	0·53
3	0	7·9	2·1	1·7	1·0	0·64
3	8	10·0	2·6	2·1	1·2	0·76
4	0	12·4	3·1	2·4	1·4	0·85
5	0	17·6	4·4	3·2	1·8	1·06
6	0	...	6·1	4·1	2·3	1·28

THE BEST SIZE OF SAMPLING UNITS

41. If we know the standard deviation for any particular size of grids we can now easily obtain the standard error per sq. mile for any given rate of expenditure. I am giving below in Table (7) illustrative examples for two thanas. The observed value of the standard deviation is given at the top in each case together with the number of units on which they are based. Necessary calculations are easy and straightforward. At any given level of expenditure the number of sampling units per sq. mile which we can afford to have is given in Table (6) ; dividing the observed value of the standard deviation by the root of the corresponding number we get directly the standard error shown in Table (7).

TABLE 7. STANDARD ERRORS PER SQUARE MILE FOR DIFFERENT RATES OF EXPENDITURE AND DIFFERENT SIZES OF SAMPLING UNITS (1938)

Size of Units	MYMENSINGH : THANA ISWARGANJ				RANGPUR : THANA DOMAR			
	Random plots	1-Acre	4-Acre	16-Acre	Random plots	1-Acre	4-Acre	16-Acre
Number of Units=	791	1,739	411	75	328	747	194	62
Standard Devn. =	0·445	0·327	0·271	0·252	0·416	0·329	0·298	0·228
(1)	(2·1)	(2·2)	(2·3)	(2·4)	(3·1)	(3·2)	(3·3)	(3·4)
Rate Rs. As.								
1 0	0·315	0·422	0·384	0·459	0·294	0·425	0·421	0·416
1 8	0·249	0·345	0·303	0·356	0·233	0·347	0·333	0·322
2 0	0·210	0·287	0·259	0·301	0·196	0·289	0·284	0·272
2 8	0·180	0·251	0·229	0·265	0·169	0·253	0·252	0·240
3 0	0·158	0·226	0·208	0·252	0·148	0·227	0·228	0·228
3 8	0·141	0·203	0·187	0·230	0·132	0·204	0·205	0·208
4 0	0·130	0·186	0·175	0·213	0·122	0·187	0·192	0·193
5 0	0·106	0·156	0·152	0·188	0·099	0·157	0·166	0·170
6 0	...	0·132	0·134	0·166	...	0·133	0·147	0·150

It will be noticed from this table that in Mymensingh (thana Iswarganj) random plots appear to be more economical in the sense of giving the lowest standard error for any given expenditure ; but as regards grids, the best size appears to be 4-acre rather than 1-acre. In Rangpur also we find that random plots and 4-acre grids are more economical at comparatively low rates of expenditure ; at higher rates 1-acre grids appear to be more efficient than 4-acre but random plots retain their superiority.

42. If we divide the standard errors by the corresponding average values of p, the proportion of land under jute, we can get the percentage error per sq. mile for the final estimate for different sizes of grids at different levels of expenditure. This will enable us to find out, for example, what will be the expenditure required to attain the same level of accuracy (as measured by the percentage error) in working with grids of different sizes. I am showing this in Table (8) in which the cost per sq. mile is given for different sizes of grids for different levels of percentage error per sq. mile.

In the case of Mymensingh it will be noticed, for example, that at the level of percentage error per sq. mile of 50 the cost will be Rs. 2·90 per sq. mile in working with random plots or less than half of Rs. 6·10, the cost for working with grids of size 16-acre. Table (8) shows many instances of similar striking contrasts. We find then that there is a wide variation in the relative efficiency of grids of different sizes ; and thus Table (8) brings out most clearly how important it is to use grids of the proper size.

TABLE 8. EXPENDITURE IN RUPEES PER SQUARE MILE FOR DIFFERENT VALUES OF PERCENTAGE ERROR AND DIFFERENT SIZES OF SAMPLING UNITS (1938)

Percentage Error	MYMENSINGH : THANA ISWARGANJ				RANGPUR : THANA DOMAR			
	R. P.	1-Acre	4-Acre	16-Acre	R. P.	1-Acre	4-Acre	16-Acre
(1)	(2·1)	(2·2)	(2·3)	(2·4)	(3·1)	(3·2)	(3·3)	(3·4)
25	...	...	...	...	6·20	...	..	...
50	2·90	4·70	4·40	6·10	2·60	4·80	5·10	5·40
75	1·50	2·50	2·20	3·00	1·40	2·70	2·70	2·50
100	0·90	1·60	1·30	1·80	0·90	1·70	1·60	1·50
150	0·10	0·60	0·40	0·90	...	...	...	...

43. From the same Table (8) it is possible, by multiplying the percentage error per sq. mile by the reciprocal of the square root of the area (in sq. miles) under survey, to obtain the standard error of the final estimate. For example, provided the best size of grids is used, with an expenditure of three rupees per sq. mile or a total expenditure of less than two lakhs of rupees (inclusive of overhead charges for supervision, maps etc.) it should be possible to obtain an estimate of the area under jute in Bengal with a margin of error of the order of say three per cent for the whole province. This shows how a sample survey can furnish, at a moderate cost, a forecast of the jute area with an accuracy which will be probably sufficient for all practical purposes.

DISTRIBUTION OF SAMPLES IN DIFFERENT ZONES

44. I may, at this stage, consider very briefly the question of the best distribution of the sampling units in different zones. It will be noticed that for homogeneous zones, in which the intensity of cultivation (*i.e.*, the proportion of land under jute) is more or less the same, the density of the sampling units will be automatically decided as soon as the rate of expenditure is settled. This because, once the best size of grids is settled, the number of such grids which we can afford to use is also determinate. The question, however, is more complicated when we have to cover areas in which the intensity of cultivation varies appreciably from one region to another. In this case, for a complete solution of the problem, it is necessary to consider the variation of both the size and the density at the same time. I have already indicated in the abstract outline of the mathematical theory how this can be done when the form of the Variance and the Cost Functions are known. When the form is not known we must have recourse to a method of successive approximations for obtaining numerical solutions. We may, for example, find out what would be the best size of the sampling units in case only one single size was to be used over the whole area proposed to be surveyed. Once this is settled the total number of grids which would be available with the money at our disposal would also be

automatically determined. We can then proceed to distribute this total number of grids among the various zones in such a way as to reduce the error of the final estimate to a minimum. This will give us an intermediate solution of the problem which however is not exact. In the next stage we can find out separately for each zone what would be the best size of the grids in the neighbourhood of the appropriate density as obtained in the first approximation. We can then use these values to obtain a second approximation. Even now the solution will not be exact but may be sufficiently accurate for all practical purposes. The important point to be noted, however, is that in a complete solution we may find it necessary to use grids of different sizes (as well as of different densities) in different regions. For example, we may find that it is desirable to use grids of comparatively small size or random plots in regions in which the intensity of cultivation and hence the density of sampling units is large, while in other regions in which the proportion of land under jute is small it may be advisable to use grids of comparatively large size. The observational material is not however adequate to enable me to discuss this question in greater detail at the present stage.

SUMMARY OF RESULTS

45. I may now give a summary of the results reached on the basis of the experimental crop census carried out in two thanas in two districts in 1937 and in nine thanas in eight districts in 1938.

(*i*) The estimated area under jute obtained by different methods of random sampling, that is, by using grids of different sizes, are in satisfactory agreement. The method of a sample census may therefore be used with success for improving the jute forecast.

(*ii*) The variance or standard error decreases as the size of the grid is increased but at a much slower rate than the normal or binomial distribution; and the Variance Function is given by a simple exponential formula.

(*iii*) The cost per sampling unit for the laboratory portion of the work is practically constant; so that the cost per sq. mile increases linearly with the density of grids.

(*iv*) The cost per sample for the field work decreases appreciably with increasing density of grids; and hence the cost per sq. mile increases at a slower rate than the density or the number of grids per sq. mile.

(*v*) With any assigned rate of expenditure per sq. mile the error of the final estimate depends on both the size as well as the density or number of grids per sq. mile.

(*vi*) Available material suggests that on the whole sample units or grids of small size are more efficient or more economical than grids of large size in the sense that the error of the final estimate is smaller with any given expenditure. To put the matter in a slightly different way, we find that, to attain the same accuracy in the final estimate, the cost is likely to be distinctly lower in working with sampling units of smaller size. It is most important therefore that a proper selection should be made of the size and density of grids; the cost may easily be doubled or trebled by a wrong choice.

(*vii*) The intensity of cultivation (that is, the proportion of land under jute) is known to vary widely in different areas. It is therefore necessary to use different densities of grids in different regions in order to attain the highest accuracy in the estimate. It is possible therefore that for maximum efficiency it may be necessary to use different sizes as well as different densities) of grids in different regions.

(*viii*) It is also possible that the Cost Function, that is, the relation between the cost of operations and the size and density of grids, will be different in different regions. The question of stratification or sub-division of the whole area for sampling purposes is thus more complicated in the present case; but is automatically included in the solution of the problem of finding out the best size and density of grids for each local region.

(*ix*) As regards actual cost of operations, available evidence suggests that with an expenditure of three rupees per sq. mile or less than two lakhs of rupees for the whole province it should be possible to obtain an estimate of the area under jute in Bengal with an error of the order of three per cent or less.

(*x*) The possibility of carrying out such a Sample Census with success will, however, depend on (*a*) the statistical technique being adequately worked out, and (*b*) the necessary human agency being built up for this purpose on efficient lines.

Advantages of a Sample Census

46. We have seen that the sample census is capable of providing estimates of the area under jute in Bengal with reasonable accuracy at a moderate cost. We may, in conclusion, compare it with a complete census. It must be admitted that theoretically the complete census has two distinct advantages. As there is no error of sampling, the accuracy of the final result is determined simply by the accuracy of the primary enumeration on the field. In fact, if errors of recording can be eliminated, the final result should be hundred per cent. correct. Secondly, a complete census will give detailed information about every individual plot.

47. In the sample census, on the other hand, in addition to errors of recording we shall also have errors of sampling. This, however, is not a serious difficulty. For, by increasing the number of sampling units, we can reduce the sampling error to as low a figure as may be desired. The really serious theoretical weakness in the sampling method is a lack of detailed information for individual plots. In fact, in the sample method the relative error of the estimates for smaller regions is bound to be comparatively high.

48. In theory there can be no doubt therefore that the complete census is superior to the sample census. The real objections to a complete census are of a purely practical nature. First of all, the cost is high. For jute alone the total expenditure may easily come to ten or twelve lakhs of rupees in Bengal. But even the question of cost is in one sense secondary. In view of the importance of the problem it is conceivable that Government may be willing to provide necessary funds for this purpose. But even if the money is available the difficulties of organization will be great. To serve any useful purpose the census must be completed in one season. This because of the large fluctuations in the area under jute which are known to occur from year to year ; so that a divided census spread over 5 years, or even 3 years, will be practically useless, as the figures collected in different regions in different seasons will not be comparable. At a rough estimate a complete census will require an army of say five thousand field workers who will be scattered over an area of say 55,000 sq. miles ; the difficulties of organization can be easily imagined.

49. The strongest argument against the complete census is, however, in my opinion, the unrealiability of the results. I discussed this question in considerable detail in my reports on the Experimental Crop Census of 1937 and 1938. In both these years arrangements were made to carry out repeated enumeration by independent sets of workers by selecting plots and villages at random and also by using grids of different sizes in areas in which complete enumeration also had been carried out. A detailed comparison of the duplicate, or sometimes triplicate or even quadruplicate, records for the same plots showed that the absolute percentage discrepancy varied from 21 per cent to over 200 per cent in different villages and was so high as 58 per cent for Deganga thana as a whole in 1937. The detailed information is thus thoroughly unreliable. The percentage algebraic error for Deganga thana as a whole was, however, small and only $-4{\cdot}8$ per cent. This shows that positive and negative errors of recording had balanced out to a great extent in the final estimate, but records for individual plots or for individual villages were completely

untrustworthy. The quality of work had improved to some extent in 1938; but the absolute percentage discrepancy still remained high and of the order of 25 or 30 per cent in the case of thana Iswarganj for which a complete census had been carried out. In view of such large discrepancies between results of enumeration carried out by independent sets of field workers we find that the chief theoretical advantage of the complete census, namely, detailed information about individual plots will be actually lacking in practice.

50. In this situation the sample survey, carried on from year to year, offers the most promising line of advance in the immediate future. This plan has many advantages. It will require a comparatively small staff of a few hundred men against so many thousands for a complete census; and will reduce appreciably the difficulties of organization and inspection. In a programme extended over several years it will be possible to give proper training to the workers and also to eliminate unreliable enumerators. This will enable the accuracy of the results being improved, and the cost of the sample survey reduced progressively from year to year. Finally the expenditure involved will be comparatively small and of the order of say two lakhs of rupees against ten or twelve lakhs in a complete census. The plan is also flexible and will allow the work being expanded or intensified in accordance with current needs.[10]

DISCUSSION ON PLANNING OF EXPERIMENTS

P. C. MAHALANOBIS

The preceding paper was followed by a discussion in which Professors S. N. Bose, N. R. Sen (Calcutta), K. Krishnan (Calcutta), K. B. Madhava (Mysore), Dr. K. R. Ramanathan (Poona), Mr. S. N. Roy (Statistical Laboratory, Calcutta) and others participated. In course of the discussion P. C. Mahalanobis pointed out the need of proper planning of experiments in physics and other natural sciences, a written summary of his observations is given below:

As a simple case we may consider a problem in which only two variables are involved, and in which the specification or the form of the equation connecting the two variables is known. For example in the model sampling experiments described above it is assumed that the variance (v) changes with the size of the sampling unit (x) in accordance with formula : $g_i \log (b_i x_i) = \log (p_i q_i / v_i)$. To test this formula it is necessary to carry out model sampling experiments with a number of different sizes of sampling units. In physics an exactly similar problem arises in the case of Boyle's Law which gives the relation between the pressure (p) and the volume (v) of a gas in the simple form pv=constant when the temperature of the gas is kept constant.

In such cases it is usual to select approximately a number of suitable values of one of the variables, which may be called the independent variable (x) for convenience of reference; and to make a number of observations on either the dependent variable (y) or on both the variables (x and y) at each such selected value of the independent variable (x). To fix our ideas let us assume that $x_1, x_2, \ldots . x_k$ are the selected values of the independent variable, and $n_1, n_2, \ldots . n_k$ the corresponding number of observations of say both x and y at each of these selected values. The total number of observations is thus $n_1+n_2+ \ldots . +n_k=N$.

Now in actual practice, especially in the case of experiments of a routine or survey type, we cannot afford or do not desire to increase indefinitely the total number of observations. That is, there is usually an upper limit N for the total number of observations; or more generally, there is an upper limit to the time (and/or cost) which we can afford to spend for completing the whole set of observations; so that the total number of observations N itself may vary but the total time (and/or cost) is kept constant.

[10] *Notes added in November, 1939.* An exploratory sample census was carried out in 1939 in five districts at a total cost of about eighty thousand rupees. A detailed statistical analysis of the results has fully confirmed the exponential form of the Variance Function; and has enabled a more accurate estimate being made of the Cost Function. In fact it was found that, at the level of expenditure of about Rs. 2/- or Rs. 3/- per square mile, grids of size 4-acre are likely to be most efficient on the whole. The material has been fully discussed in the *Report on the Sample Census of Jute in 1939* submitted to the Indian Central Jute Committee on the 8th November, 1939.

An interesting problem in the planning of experiments arises in this situation. Let us consider the case in which the total number of observations N is to be kept constant; also let us assume that the range of the independent variable (over which the observations would be made) is settled beforehand.

For example, in the Boyle's Law experiment on air we may decide to make, say, $N=100$ observations altogether; and also that the experiments would extend roughly from a pressure of say 6 cms. of mercury to say 106 cms. of mercury, i.e. over a range of about 100 cms. of mercury. Now we may distribute our 100 observations over this range in different ways; we may, for example, take one single reading at 100 different values of the pressure; or two observations at each of fifty different values; or five observations at each of twenty points, or fifty observations at each of only two points of the range. Instead of taking an equal number of observations at each of the selected value of the independent variable we may also take different number of observations at different points. This we may call the pattern of the observations.

Our first task then is to settle the pattern of the observations, namely, 100 sets of one observation each, or 50 sets each of 2 observations, or 20 sets each of 5 observations, or 2 sets each of 50 observations etc. In case we decide to take an unequal number of observations in each set, we shall have to specify the pattern by stating the number of observations in each set, namely, $n_1, n_2, \ldots . n_k$ respectively in k sets of observations subject to the condition $n_1+n_2+ \ldots . +n_k=N$

In the second place we have to settle approximately the particular values of the independent variable, namely, $(x_1, x_2 \ldots . x_k)$ where these k successive sets of observations would be made. This we may call fixing the location of the observations.

We may now define our problem. We have to find out that particular pattern and location of the observations which would enable the parameters of the equation of specification being estimated with the lowest sampling variance. It may turn out that all random patterns and all random locations are equally efficient (in the sense in which R. A. Fisher has used the term efficiency); if so, it would be of great theoretical interest as well as of practical value to prove this result. On the other hand, if different patterns and/or different locations differ in efficiency it is certainly of importance to be able to select the more efficient designs.

This is the problem which arises in model sampling experiments for estimating the parameters of the Variance Function. We know the form of the function; we know, or can know by a preliminary series of experiments, the approximate time or cost of finding the variance for a given number of sampling units of a given size; we also know what is the total time or cost which we desire to spend on the work; what we have to find out is how many and what different sizes of sampling units (*i.e.* values of x) we should choose and how may units of each size we should use for our investigations.

It is scarcely necessary to remark that the problem of planning (in the sense of selecting the most efficient pattern and location of the observations) arises only because the equation of specification connecting the different variables being more or less a good fit to the observations is essentially of an empirical or statistical (and not of a rigorous mathematical) nature. The solution of the problem is, therefore, necessarily of an iterative character. If we have no knowledge regarding the relative cost (in time or money) of making observations or regarding their relative accuracy at different points of the range there is nothing to guide us in deciding our programme, and we may have to perform the experiments in a haphazard manner or use general common sense principles such as distributing the observations over the whole range and including replications at each point of observation. But as soon as relevant information regarding accuracy and cost becomes available, either in the course of actual experimentation or from preliminary exploratory surveys, we are in a position to use this information for planning the next phase of the work. It should be possible in this way to improve the efficiency of the experiments in each subsequent stage by utilizing the information collected in previous stages of the work (unless, of course, all designs have the same efficiency).

It is worth noting the contrast between the statistical analysis of observations which have been already made, and the statistical planning of the programme for obtaining observations which are to be made in future. In the method of fitting curves by least squares the observations are given and a certain mathematical procedure is used for estimating the parameters in the equation of specification in order that the residual errors of the estimates are reduced to a minimum. In the present case we have the converse problem; we want to decide where and how many observations we should take, subject to the total time spent on the work being kept constant, in order to reduce the sampling error of the estimate to a minimum. In large scale surveys or in routine measurements such a situation frequently occurs in practice. The question is thus of both practical and theoretical importance and deserves the serious notice of mathematicians.