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.

PLANNING OF EXPERIMENTS

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.