Good — lots of q

PROFESSOR P. C. MAHALANOBIS then read his paper:

RECENT EXPERIMENTS IN STATISTICAL SAMPLING IN THE INDIAN STATISTICAL INSTITUTE

Introduction

I was naturally gratified to receive in March 1946 an invitation to present a paper before the Royal Statistical Society during my visit to London. I had collected some material for this purpose and had written certain portions of the paper when I was suddenly obliged to leave India at the end of April to attend a session of the United Nations Statistical Commission in New York. Owing to unforeseen difficulties and pressure of work in connection with the Royal Society and Empire Scientific Conferences in London in June and July, I had no time to finish the paper. It was not even possible to circulate the tables which had to be presented before the meeting in a mimeographed form. I greatly appreciate the kindness shown to me in relaxing the rule about submitting the written paper in advance, and secondly in arranging the present meeting during the closed session. The present paper contains the observations made at the meeting as well as certain portions of my original notes. I acknowledge with thanks the help I have received from my young colleague, Prasad Banerjee, in putting this paper in its final shape.

Restricted scope of the paper. This paper has been deliberately named as " Recent Experiments in Statistical Sampling in the Indian Statistical Institute " to make it clear that I have no desire to speak on sample surveys in general, or to make any attempt to give an account of the many interesting and significant developments in statistical sampling which have taken place in recent years outside India. A large number of papers have been published on the subject in statistical and scientific journals, and Dr. F. Yates, in the paper recently presented before the Royal Statistical Society, has given a comprehensive review. It is not necessary for me to touch the ground covered by others. The aim of the present paper is to draw attention to certain experiments in statistical sampling conducted during the period 1937–45 by the Indian Statistical Institute, of which I am the Honorary Secretary.

The Indian Statistical Institute. It will be convenient if I give at this stage a brief account of the Indian Statistical Institute itself, as this would explain the set-up in which the work was done. The Institute was inaugurated on December 17th, 1931, and was formally registered on April 28th, 1932, as a non-profit-making scientific society. Work began on a small scale, and the total expenditure in the first year (1932–33) was about £40. Special courses in statistics were started in the second year, which later developed into regular post-graduate training classes. Research work was also begun, at first on a part-time basis, but later with the help of a whole-time nucleus staff. A scheme for conducting examinations and awarding certificates of proficiency for computers and statisticians had been prepared in 1935–36 which was brought into effect from 1938. In 1941 the Institute helped in establishing a post-graduate department of statistics in the Calcutta University offering whole-time courses for the degrees in Statistics of M.A. and M.Sc.; and the actual teaching work was done in the Institute itself by a practically joint Institute and University staff. *Ad hoc* enquiries on a small scale were being undertaken from 1935, but project work began to develop on a large scale from 1937 with the initiation of a scheme for improving the forecast of the area under jute in Bengal. The project side developed rapidly during the war, as the Institute had to undertake various statistical enquiries and surveys on behalf of Government Departments. In 1945–46 the total volume of employment was about 750 man-years, divided about equally between the statistical and field branches, and the total expenditure was of the order of £100,000. Excepting for a comparatively small research grant of about £4,000, practically the whole of the income was derived in the form of contract grants from Government Departments for specific enquiries and projects. It is interesting to observe that the machinery was practically the same which had developed independently in the U.S.A., where large schemes of statistical investigations relating to the war effort were being done in different Universities and scientific institutions with the help of contract grants from the Government.

Conditions of work. Several points deserve to be emphasized. Each project or enquiry had definite practical objectives which were laid down by the client (usually a Government Department); the permissible margin of error of the results was broadly indicated; the enquiry had to be conducted in accordance with a definite time schedule so as to enable interim or final reports to be submitted by particular dates; the whole work had to be done within the limits of an all-inclusive

NB

Instrument to overstate research

contract grant, so that any excess of expenditure would mean a financial loss to the Institute. Finally, owing to difficulties created by the war situation in North-east India, the human agency and the organizational side of the project had to be very carefully planned, and usually the field survey had to be carried out under great difficulties. The research that had to be undertaken was naturally of an applied type. Nevertheless fascinating theoretical problems were continually arising, some of which were and are being tackled on fundamental lines. In preparing the design of each sample survey the three most important considerations thus were time, cost, and the human agency—all, of course, in relation to the most important factor of the permissible margin of error which was usually stipulated in advance.

"Statistical engineering." In the present paper I have very much in mind those problems of organization which arise when a sample survey has to be carried out on a very large scale. The difference in quantity is so great that it brings about practically a change in quality. The manufacture on a commercial scale of a chemical gives rise to problems of an entirely different type from those which have to be solved in small-scale working in a laboratory, and the difference is so great that we give expression to it by calling large-scale production a matter of chemical engineering rather than of pure chemistry. In the same way large-scale sample surveys may be appropriately called statistical engineering. A good deal of the present paper is in fact concerned with statistical engineering rather than the pure theory of sampling.

General observations

The sample enquiries discussed in the present paper fall into two broad types. A great deal of work has been done on sample surveys on an extensive scale covering whole provinces (50,000 to 140,000 square miles in extent in the same season), in which the object has been to obtain reliable estimates of the acreage, rate of yield per acre, and total production of important food and fibre crops like rice, wheat, jute, etc.; or of economic or demographic factors relating to indebtedness, unemployment, destitution, paddy land, plough cattle, birth rate, death rate, etc., of rural families. For such work the technique used has always been the area or grid method, in which comparatively small sampling units (of area or size ranging from about two acres to one square mile) were located at random, and information was collected for each sample-unit by direct physical inspection or investigation. Estimates for districts (about 2,500 square miles or so in area on an average) or the province as a whole were then obtained directly by multiplying observed averages by appropriate multiplying factors (equal to the ratio of the total geographical area of the region to the total area covered in the sample surveys).

In a second type of work the enquiry was more localized, and related to cost and level of living, housing, consumption of food, clothes, etc.; preferences for particular types of commodities; reactions to radio programmes; public opinion on various subjects, etc.

Terminology and classification of sample surveys

Before proceeding further it will be convenient to explain the terminology and classification of sample surveys. In the area method the smallest physical element which can be separately surveyed is called a *quad*. Typical examples are a single plant for determining the yield of crop; a single family or household (having food from the same kitchen) for ascertaining the cost of living; or a single individual in anthropometric or blood-group surveys. The whole region or field to be surveyed may be then considered to consist of a very large number of such ultimate physical elements or *quads*.

Quad and configurational sampling. Sampling may then proceed in two different ways. The ultimate physical element or quad may be adopted as the sampling unit, or a group of such elements taken together may be used as the unit for sampling; these two types are called "quad" and "configurational" (or "grid" or "cluster") sampling. The grid, in the area method, consists of a suitable number of adjoining quads or ultimate physical units. For example, in crop surveys in Bengal it was found that a suitable sampling unit was a grid of square shape and size 2·25 acre. Each grid thus consists of a compact group of adjoining quads.

It must be noted, however, that the sample-unit may also consist of quads separated from one another, but in a fixed spatial configuration—for example, 4 quads at the corner of a square of a given size. It is convenient therefore to distinguish between the "grid" and the "cluster," and restrict the former to a compact group of adjoining individual units, and the latter to other types

of configurational sampling where the quads are not adjoining but are arranged in some particular pattern in one or two dimensions. Every *n*th individual unit in a linear series would also in this sense be an example of cluster or configurational sampling. From the present point of view all sampling methods may then be characterized as being of either (1) of *quad*, (2) of *grid*, or (3) of *cluster* type.

Zonal and non-zonal. The important point to be noted is that all the sample-units have to be located at random. But this can be done in two ways—namely, distributing the sample units over the region under survey as a whole, or by first dividing the whole region into a suitable number of sub-divisions or zones, allocating a suitable number of sample units to each zone, and then locating such sample units, grids, or clusters at random within each zone separately. This gives rise to two types—namely, (1) without zoning or non-zonal, and (2) zonal sampling.

Multi-stage. So far methods have been considered in which the sample units are located at random in one single stage over each zone or over the whole area under survey. In many enquiries this method cannot be used because of its high cost. For example, in estimating the rate of yield per acre of crops it is usually necessary to proceed by stages. A number of zones are first selected at random; within each selected zone a number of villages are next selected at random; within each village so selected a number of fields are then selected at random; and within each field so selected one or more sample cuts are finally located at random. In this method the act of randomization is performed, not in one single stage, but in successive stages. The coverage of the process of randomization takes place repeatedly over smaller and smaller regions. This type of sampling is called multi-stage.

Replicated networks of sample units. Information for each zone (or for the entire field under survey) may be collected in one single network of sample units; or the information may be collected separately for two or more networks of sample units. Each such network would give an independent estimate, and differences between such estimates supply immediately measures of the over-all or effective margin of error. A characteristic feature of the Statistical Institute's work has been the use of such independent replicated networks of samples.

It would be noticed that quad or configurational, zonal or non-zonal, and replicated sampling can be used with either the uni-stage or the multi-stage methods. Thus quad sampling may be used at one or more stages, and grid or cluster at another stage. In the same way, the sampling method may be non-zonal at one or more stages and zonal at others. There may be one single network of samples at one or more stages or more than one network at other stages. The present system of classification of sampling methods is thus flexible and comprehensive, and makes it possible to describe any particular design of sample survey in a precise and non-ambiguous manner. It is suggested that the present terminology (with such modifications that may be considered desirable) should be adopted as a standard system for the classification of sample surveys.

Large-scale surveys

When the work is done on an extensive scale covering a large geographical area it is almost inevitable that the sample units must be located at a considerable distance from one another. The investigators have therefore to spend a good deal of time in travelling from one sample unit to another. The total time (or the total cost) of field operations thus consists of two broad portions—namely, the actual time or cost required for enumeration or investigation of the sample units, and the time required for the journeys between different sample units. In such surveys it would be obviously uneconomical to collect information for a single quad in each locality. It is therefore usually arranged to use a grid or a cluster—that is, a group of quads for investigation in each locality visited by the investigator.

Cost function. Large-scale sample surveys are necessarily expensive, and considerations of cost are therefore of great importance.

As already noted, in surveys covering large geographical areas the cost of field operations depends on the time required for actual enumeration or investigation and the time required for journeys from one sample unit to another. Collecting information for grids of a large size would naturally take more time. If grids of a large size are used, then the total number which can be surveyed must be necessarily small. Grids of a large size would be thus more widely scattered, so that the time required for journeys would be larger for each journey on an average, but there would be fewer journeys to perform. Working with grids of a small size it

would be possible to use a large number so that the average distance apart would be less. The time for journeys on an average would be less, but a large number of journeys would have to be undertaken. The total cost of operations thus depends on the size of the grids and their total number. Working within a fixed budget this means that once the size of the grids is settled, their total number must also be determined by the budget limit.

In enquiries which are geographically more localized, but which are conducted on a sufficiently large scale, the cost of operation depends on the way the field survey is organized—that is, on the nature of the human agency employed for the work, the methods adopted for collecting the primary data, arrangements for checking and supervision, etc. In every large-scale enquiry the cost is, therefore, determined by the particular plan or design of survey proposed to be adopted.

Variance function. Cost, however, is not the only consideration. The precision of the final results is also important. The variance of grids or sample units decreases as the size (*i.e.*, area) of the grid or sample unit is increased. In crop-survey work it became clear quite early that the decrease in variance with an increase in the size of the grid was appreciably slower than the normal rate of decrease (namely, inversely as the size of the sample unit). What particular size of grid would be most economical would therefore depend on how the variance changes with the size of the grid. This necessitates a study of the variance function which gives the relation between the variance and the size of the grid. More generally, each way of distributing the sample units—that is, each particular design of sampling—involves its own appropriate sampling error, so that corresponding to each design there exists its specific variance function.

Optimum solutions. It is now possible in terms of the cost and variance functions to state the conditions for an optimum or most economical design for sampling. The aim may be stated in either of the following two alternative forms: (*a*) to determine the size and distribution or density of grids or sample units in such a way that the variance (or margin of error) of the final estimate is a minimum when the total cost of the sample survey is fixed; or, alternatively, (*b*) to determine the size and distribution of sample units so as to reduce the variance (or margin of error) of the final estimate to any desired level at a minimum cost. It has been shown elsewhere how concrete solutions (which are the same for either of the alternative forms) can be obtained with the help of an empirical knowledge of the variance and cost functions.

Two points are worth noting. The above approach is indispensable when the scale of operations is sufficiently large, and individual grids or sampling units are so widely separated that an appreciable amount of time and cost is incurred in journeys between grids of sample units, or when a large number of investigators are employed for the field survey so that the cost depends materially on the way in which the field staff is organized. Secondly, such large-scale sample surveys usually give rise to many problems involving the human agency which have engaged the special attention of the Indian Statistical Institute for a long time, and to which more detailed references have been made in later sections.

Exploratory (or sequential) development

The planning of a sample survey thus has several aspects—namely, (1) zoning and/or stratification; (2) size of the grid or sample unit; (3) arrangements for replication; (4) density or distribution of grids in different zones or strata; (5) preparation of forms and schedules; (6) structure and organization of the field staff; (7) arrangements for the statistical processing of the material, etc. It has been explained above how the efficiency of the design of the sample survey can be maximized from joint considerations of variance and cost functions. To use such methods it is necessary to have previous information about those two functions. Other relevant information relating to the region or universe to be surveyed is also of great value in preparing the design. In fact, the greater the amount and accuracy of such information the greater is the possibility of reducing the cost of operations.

When absolutely no information is available, and yet the survey has to be completed at one single operation, there is no other alternative but to use unrestricted random sampling without zoning, or with such zoning or stratification as may be considered convenient from a purely organizational point of view.

Fortunately, some information is usually available. Also, especially in large-scale work, the survey has often to be continued from year to year or at suitable intervals. It is therefore usually possible to adopt the exploratory method, in which a sample survey is first carried out on a very small scale with the primary object of collecting basic information required for preparing an efficient