SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited By : P. C. MAHALANOBIS

VOL. 12 PARTS 1 & 2 1952

SOME ASPECTS OF THE DESIGN OF SAMPLE SURVEYS

By P. C. MAHALANOBIS

Indian Statistical Institute, Calcutta

Introductory Note: In the course of a conversation with Abraham Wald the day before he started on his fateful last flight by air from Calcutta in December 1950 he referred to the connexion between sequential analysis and the concept of pilot studies in a series of repeated sample surveys which I had used in my papers and said that he had mentioned this point in his recent book on Decision Functions. I also told him briefly some of my own ideas about problems of sample design. He became keenly interested and thought he saw certain possible developments in the theory of design of sample surveys which would have some connexion with decision functions. He said he would take up the matter for study on his return to Calcutta, but this part of the story remained unfinished owing to his tragic death. As a tribute to his memory I am giving below some general observations on a topic which was the subject of my last discussion with Wald.

1. Consider the field of a single variate, x, about which it is desired to collect information by sample surveys. For convenience of discussion it is assumed that the field is finite and consists of N elementary units which are arranged in, say, a linear or one-dimensional order so that the elementary units can be numbered in a serial order 1, 2, 3,...N. It is assumed that corresponding to each elementary unit there exists, at any given instant of time, t, a finite value of $x(t)$; and this set of N values of $x_1(t)$, $x_2(t)$, $x_3(t)...x_N(t)$ completely defines the field at time, t. The field may be static in which case the values x_1, x_2,...x_N would be constants and independent of the time t; or the field may be dynamic in which case the set of N values $x_1(t)$, $x_2(t)...x_N(t)$ would be changing with time and would be functions of the time, t.

2. The problem of statistical sampling, at any instant, is (a) to ascertain the values of x for a suitably chosen set of elementary units, and (*b*) from these values of x to estimate by appropriate methods the required statistical information about the whole field.

3. The design of a sample survey (which may be more conveniently referred to as a sample-design) thus consists of a procedure for selecting a suitable set of elemen-

tary units out of the whole set of N elementary units in such manner (usually called random) as would enable statistical inferences being drawn with validity from these sample-values by means of a set of rules for drawing such inferences or for preparing appropriate estimates of the characters under investigation for the field as a whole.

4. It is possible to distinguish between two broad types of information that may be sought. In one type the information relates to certain parameters or characteristics of the field as a whole without any reference to the dimensional arrangement or ordering of the elementary units in the field. For example, the total or average value, or the standard deviation, of the N values of x, depend only on the values x themselves and not on the manner in which these values are (dimensionally) distributed among the different elementary units in the field so that any interchange of the x-values between the elementary units would leave these parameters of the population of elementary units entirely unchanged.

5. A second type of information has explicit reference to the serial order or the relative location of each elementary-unit with respect to other units, that is, to the manner in which the x-values are dimensionally distributed over the field. For example, one may be interested to investigate whether the x-values are continually increasing or decreasing, or fluctuating in any regular manner, when one passes (in serial order) from one end to the other of an one-dimensional field; or, to ascertain the variance function for sampling units which are clusters (of various sizes) of consecutive elementary units.

6. A class of investigation of the second type, to which special attention may be drawn, is the problem of preparing what may be called a map of the field showing the dimensional distribution of the x-values over the field; and I called this type of investigation a mapping problem in an earlier paper.[1] In this case, the sample-design must include rules for estimating the x-values separately and individually for the remaining $(N-n)$ non-sample units, on the basis of the ascertained x-values for the n sample-units.

7. It is conceivable that the purpose for which a map is required is adequately met by replacing the detailed map showing the x-values at each and every elementary unit by a smoothed picture, that is, a curve or surface-like representation. This is particularly so when the field has an almost continuous structure. In this case what is practicable, and which fortunately is generally what is required, is a smoothed picture of the field.[2]

8. The exact process of smoothing is a matter of definition whose adequacy is to be judged in the light of the purpose for which it is required. One way, for example, would be the use of moving averages, another process would be to fit by specified methods a smooth curve (or surface) to the whole series of x-values. A third, which is effectively the use of enlarged elementary units, involves the

[1] P. C. Mahalanobis (1944). "On Large Scale Sample Surveys". *Roy. Soc. Phil. Trans.*, B, 231, 329-45.

[2] A special type of a "mapping survey" occurs when it is desired to study how one variate x changes with variations in another variate y or how x depends on different discrete class groups of y.

substitution of the detailed map by the average x-values within clusters formed by grouping together 'adjacent' elementary units. In such cases the sampling problem is not to estimate separately and individually the actual x-values of the elementary units but the smoothed values, *i.e.* the moving averages, the curve (or surface), or the cluster-average respectively in the three cases described above.

9. Incidentally it may be pointed out that the decision as to whether a smoothed map will meet the practical ends for which it is meant depends to a large extent upon the magnitude of the 'undulations' in the smoothed map in relation to the 'divergence' of the smoothed map from the actual map. In general the usefulness of the smoothed map decreases as the fluctuations of the actual x-values about the smoothed map increases, and if the 'divergence' is too large compared to the magnitude of the 'undulations' then it may be practically useless.

10. To every estimate provided by a sample survey (for purposes of mapping or otherwise) is associated a sampling error (or in case sources of error in the process of ascertainment of sample-values and other errors are included, a margin of uncertainty); and it is of course necessary to set up suitable measures of this margin of error (or of uncertainity). In the mapping problem in principle it is possible, for example, to use the root mean square discrepancy between the actual (or 'true') values or the smoothed values as the case may be, and the estimated values of x for all sampling units as the 'deviation' (on a per unit basis) of the estimated map. The expected value (under the adopted sample-design) of the square of the deviation then supplies a measure of the sampling variance. The full significance of this measure remains to be seen.

11. In a sample-design there should be some provision for obtaining a valid estimate of the sampling error (or of the margin of uncertainty), so that an idea of the reliability of the estimate may be obtained. The problem of estimating the margin of error presents some difficulty in the case of the mapping problem. It appears, however, that if the sample-survey be conducted in the form of a number of interpenetrating net-work of sub-samples it may be then possible to estimate the sampling variance on the basis of independent estimates provided by the sub-samples, by making use of their mutual 'deviations'. However, this is a technical point, and it may be assumed for the purposes of the present general discussion that it would be possible to develop suitable measures of the sampling error and appropriate methods of its estimation in the case of mapping surveys.

12. Coming back to the basic problem of sample-design, namely, to settle a procedure for selecting the sample-units to be investigated it is essential to use a 'frame' which in its simplest from may consist of a list of the elementary units. It is also essential to have these units arranged in proper dimensional order if the second type of information (para 5) is to be collected. The frame may also contain some further information about the field.

13. When the frame consists of only a list of units and nothing else whatsoever is known about the field, the problem of sample-design reduces to the simple

case of selecting for investigation a suitable number, n, of elementary units in a random manner so that valid inferences may a be drawn from the sample by appropriate methods.

14. It is only when some previous information (which may be only approximate in nature) is available about the field that the problem of the sample-design becomes important. The object then is to use the available information in the most effective way to prepare an improved sample-design (in the sense that it would be an improvement over the best design that would otherwise be possible) so that it can be reasonably expected to reduce the cost of the survey as much as possible without sacrificing the accuracy, or alternatively, reducing the margin of error (or uncertainty) to the greatest possible extent for the same expected cost.

15. It may be pointed out here that an approximate (smoothed) map of the field in respect of the characteristic under investigation or of any other character highly correlated with it (apart from its other uses) serves usually as an excellent prior information for the improvement of sample-design. For example, if the map exhibits a non-random (in the sense explained in my 1944 memoir) or patterned field it is known that the use of stratification[3] (or other devices) on the basis of such maps often makes it possible to improve the efficiency of the sample-design.

16. On the other hand, in the case of a random-like field which really is a case in which the map is devoid of any significant features other than its randomness and therefore is not otherwise of much interest (that is, the field is not map-worthy) the sample-design may be considerably improved (from the point of view of costs) by using a very small number of very large (geographically compact) clusters of elementary units.

17. The possibility of the utilisation of the prior information available in the frame for preparing an improved sample-design shows that it is sometimes worth while carrying out certain preliminary investigations to collect relevant additional information which would pertain to or be associated with the frame thus making the sample-design essentially sequential in procedure.

18. The sequential procedure is praticularly evident in repeated surveys which are being increasingly used in some countries. At this stage it is important to define what I mean by historical or sequential and non-historical designs of sample surveys.

[3] It is worth noting that the question of stratification itself has been very little studied so far. The basic problem is to devide the field (on the basis of the knowledge of the 'frame') into a suitable number of strata. In the more familar case of "aggregate" (or non-mapping) surveys it is possible that an optimum or nearly optimum solution would be obtained when the expected contribution of each stratum to the total aggregate value of x is made equal for all strata. In fact, in general, stratification (in the sense of choosing a scheme of dividing the field into a number of strata) itself is an integral part of preparing the sample-design. Again, in mopping surveys of one-dimensional fields, it would probably be advantageous to distribute the sample-units proportionately to the rate of change of the direction of the tangent line to the smooth curve which represents the map of the field.

19. Consider, for example, a scheme for repeated surveys which consists of a simple rule that the same number, n, of sample-units would be selected at random from over the whole field at each successive repetition of the survey. In this example the information relating to n, $2n$, $3n$,...sample-units and other relevant information (cost etc.) which become progressively available in the course of the repeated surveys are not utilised to improve tha sample design of subsequent surveys. Such schemes of sampling which remain the same for each repeated survey and would not change with time may be called non-historical schemes or non-historical designs of sampling.

20. On the other hand, it is also possible to try to use the information which gradually accumulates, to improve the sample-design for later surveys. Such schemes may be called historical or sequential designs of sample surveys.

21. The use of sequential or historical sample designs is particularly attractive in repeated sample surveys. This arises from the fact that very often the field (map) does not change very abruptly from one survey to the next or even when it does (as can be verified by the repetition itself) there may be a fairly stable pattern in the change itself; furthermore the set of surveys itself may progressively reveal the nature of these patterns. If the first few surveys show that the field is random-like, it would probably be decided to use the design, previously pointed out, of a very small number of very big sample-units (each sample-unit consisting of a large number of elementary units) and also to continue to use the same design in subsequent surveys. It is evident that in this case we would either use a non-historical design from the beginning or change over from a historical to a non-historical scheme. On the other hand, in the case of repeated surveys of a non-random or patterned field, historical or sequential scheme of sampling, in general, should be more efficient than non-historical designs.

22. For repeated sampling, the problem of sample-design assumes a more general character, for it is possible, in fact, to consider the problem not merely as of deciding a suitable design for a single sample survey, or deciding separately suitable designs for successive repeated surveys, but as deciding a design for the series of successive surveys as a whole.

23. When several alternative sample-designs are open for choice, whether it be the case of a single sample survey or a series of repeated surveys, the objective is to choose that one which is most efficient (which may be called the optimum design). The most efficient or optimum design may be regarded as the one which (a) supplies the required information with a minimum margin of error (or uncertainty) at any assigned level of the cost; or alternatively, (b) to supply the required information within the assigned margin of error (or uncertainty) at a minimum cost.

24. In repeated sample surveys, the object now, as has been pointed out earlier, would be not merely to have an optimum design for the first survey alone, but to develop an optimum sampling scheme which would be most efficient over the whole series of successive surveys considered in its entirety. It would be, therefore,

often profitable in the long run to incur some extra expenditure in the earlier surveys for securing auxiliary information (not only relating directly to the variable x, under enquiry, but also to other operational factors such as costs related to the survey) which although not of immediate direct interest is likely to prove useful at later surveys in improving the efficiency of the sample-design either by way of providing supplementary information which leads to the use of better methods of estimation or to a more radical change of the sample-design itself. The need of securing such auxiliary information would be, in general, a continuing one and would be present at successive repetitions of the sample survey.

25. It is important to note that the nature of the auxilliary information need not necessarily be the same in every one of the surveys. In fact, the decision regarding what type of information is to be collected at each successive surveys may possibly depend upon the information provided by the earlier surveys. Thus a sequential procedure is also involved in the choice of the type of auxiliary information,

26. The optimum sample-design depends upon how we measure the margin of error (or uncertainty) which is the yard-stick used to compare the relative efficiencies of two sample-designs. The use of the standard error of the estimate as the yard-stick when a single character is to be estimated from a single sample survey is usually satisfactory. In case of repeated surveys for a single character (with added complications when several characters are to be estimated, as is usually the case) we need some sort of a 'composite' error defined in terms of the standard errors of the the several estimates calculated on the basis of the sample-design. The exact definition is a matter of details which need not be discussed in this general exposition. It may be possible, however, in certain situations to adopt Wald's decision function approach, involving a suitable risk function, to the problem of optimum sample-design.

27. Whatever may be the definition of optimum design, in attempting to settle it in any specific case, it would be usually necessary to make calculations relating to the expected margin of error (or uncertainty) and the expected cost of operations, on the basis of the prior information available in the frame. Such prior information, however, is not complete and/or is partly inaccurate. The results of any calculations made on the basis of such (incoomplete and/or partly incorrect) information must necessarily have a margin of uncertainty. To reduce the margin of uncertainty on this account it may therefore be sometimes worth-while to incur some expenditure with a view to improving the quality of information on the basis of which the sample-design is to be prepared.

28. Secondly, the optimum design must necessarily refer to the future instant of time, t, at which the survey would be actually conducted. In other words, the decision about the optimum design which has to be made earlier than the instant of time, t, must be made on the basis of calculations which have reference to the future values of x which would be assumed at the actual time of the survey. But in practice these calculations would have to be made either on the past values of the character (which are known from earlier surveys of the series or from some

other source) or on the projected values of the character for a future instant of time. Now fields in which one is interested are usually dynamic and change with time, that is, the value of $x\ (t)$ for each elementary unit is a function of time, and therefore in either case a further factor of uncertainty would enter into the decision relating to the optimum design.

29. For reasons explained above, the decision about the optimum design cannot be sharp or critically rigorous in a (non-probabilistie) mathematieal sense, (however, attempts may be made to devise rules for arriving at a rigorous decision in a probabilistic sense), and must have a margin of error or uncertainty. To explain the matter in another way I should like to emphasize that 'the' optimum design (or the 'true' optimum design) can only be defined precisely in terms of the x-values of the field at the time of the survey, the cost of various operations at the same time, etc. In actual practice we can never say unambiguously which is 'the true' optimum design (in fact, if we could there would have been no sampling problem) and all that we can do is to estimate or forecast "the" optimum design on the basis of prior information, and there is bound to be a margin of uncertainty in such estimates or forecasts. In other words, the decision about the optimum design must necessarily involve a margin of uncertainty. Refinements in the design which go beyond the actual margin of error or uncertainty are not of much use.

30. A programme of repeated sample surveys in which a historical sample-design is being used would thus always have the quality of pilot studies and would involve a sequential procedure in essence, and for such repeated surveys the real problem then would be to develop a sequential programme of the highest efficiency but there would always remain an element of uncertainty in the decision about the most efficient sequential programme, and it would never be possible to reach a mathematically rigorous solution.