APPENDIX A

Extracts from 'On Mahalanobis's Contributions to the Development of Sample Survey Theory'*

Earlier developments: To enable appreciation of Mahalanobis' contribution to sample surveys, it is proposed to give a very brief review of the developments in this field prior to 1930. Kiaer, who had realised the importance of the method of representative sampling almost 70 years ago, had conducted some surveys in Norway as early as 1900 and was instrumental in focusing the attention of the International Institute of Statistics on the need for application of the representative method in statistics. At about this time, Wright had conducted statistical surveys in the United States using this method (Seng, 1951). Zarkovic (1956; 1962) mentions the use of the sampling method in Russia as early as 1900. He has also drawn attention to Kowalsky's book on sampling methods published in 1924.

In India the need for random sampling in statistical surveys was perhaps first recognised by Hubback as early as 1923, or even earlier. In this report on 'Sampling for Rice Yield in Bihar and Orissa (1923–25)', published in 1927, Hubback (1946, p. 283) has stated:

'The only way in which a satisfactory estimate can be formed is by as close an approximation to random sampling as the circumstances permit, since that not only gets rid of the personal limitations of the experiment or but also makes it possible to say what is the probability with which the results of a given number of samples will be within a given range from the mean. To put this into definite language, it should be possible to find out how many samples will be required to secure that the odds are at least 20 : 1 on the mean of the samples being within one maund of the true mean.'

Mahalanobis (1946), who publicised the work of Hubback by republishing his report in *Sankhyā* and reviewing their work, states that Fisher's work at Rothamstead had been directly influenced by Hubback's paper. Fisher (1925) emphasized the need for randomization, replication and statistical control in scientific experiments. Clapham (1929) discussed the use of sampling method in the estimation of yield of cereal crops. Nemchinor (1932) presented, in detail, the methods used in the USSR for forecasting crop yield through extensive corp-cutting experiments.

The Committee on the 'Application of the Representative Method in Statistics', set up in 1924 by the International Statistical Institute, prepared its report in 1925 in which the theoretical and practical aspects of the random sampling method have been discussed (Jensen, 1926). Bowley's (1926) contribution to this report on the question of stratification and allocation is of particular importance.

Acquaintance with sample surveys: The earliest recorded reference to Mahalanobis' acquaintance with sampling theory, which I have come across, relates to his lecture on 'Elements of the Theory of Sampling' delivered on 8 July 1932 at the Statistical Laboratory, Presidency College, Calcutta. In this lecture, he discussed the concept of *population* in statistics by comparing it with that of *universe of discourse* in logic and stressed the distinction between the *population value* and *sample value* and the need for probability interpretation of the observed statistics.

Definition of sampling given by PCM: 'A *sample* consists of two or more *elementary units* drawn from a *population (universe or field)* in a *random* manner through the sample *frame* and the *inference* about the population is based on the *observations, measurements* and *experiments* on the *variates* (or *characteristics*) of the elementary units in the sample.'

From the beginning, Mahalanobis clearly recognized the immense potentialities of the method of random sampling and also the need for proper interpretation of sample survey results. In an editorial note on the paper 'Price of Rice at Bolpur' by Santipriya Bose, who had discussed the results of a survey without any reference to their margins of error, Mahalanobis (1936) pointed out that the margin of error in the results of the survey discussed in the paper is rather large and that the margin of error of even the key characteristics is of the order of 5 per cent for the sample of 15 holdings used in the survey. He further pointed out that to get a margin of 1 per cent it would be necessary to have a larger sample of about 500 holdings.

In another editorial note, on the paper 'The Future Trend of Population Growth in India' by B.P. Adarkar, which had given rise to considerable controversy, Mahalanobis (1937) commented that the data on which the paper was based were likely to be unreliable and stated:

'A census (or complete enumeration) method of studying differential fertility is impracticably in India owing to vast numbers. Recourse to a sample survey method is, therefore, inevitable. Fortunately, owing to recent developments in statistical theory of representative and stratified random sampling, the efficiency of this method has increased very appreciably. The cost is also not prohibitive.'

With regard to the question of crop statistics, Mahalanobis (1940a, p. 512) mentions: 'In March 1936, in reply to the enquiry from the Secretary, Imperial Council of Agricultural Research, I suggested exploring the possibilities of random sample method for estimating the area under different crops in Bengal.'

Systematic contributions to sample surveys: Mahalanobis' systematic contribution to sample surveys may be considered to have commenced with his scheme for sample census of jute crop in Bengal prepared as early as 1937. This scheme was devised to estimate the acreage under jute and yield of jute crop in Bengal and this was phased over five years. In a sense the approach adopted in this scheme was sequential. In this first year the survey was confined to two *thanas* (groups of about 100 to 200 villages) in Bengal and the survey was gradually expanded to culminate in a large scale sample survey in 1941 covering the whole of Bengal. The survey in its initial years was so planned as to provide data on variability of the characteristics under consideration and on costs of the different survey operations, which could be used to design the large scale sample survey efficiently and with minimum cost. His

*M.N. Murthy (1963), 'On Mahalanobis's Contributions to the Development of Sample Survey Theory and Method', in C.R. Rao et al. (eds) *Contributions to Statistics*, Pergamon Press.

work on the crop survey in Bengal was so successful that the Government of Bihar requested him to carry out a similar survey in Bihar. He has since carried out a number of surveys at the instance of the Central and the state governments, and other agencies...

He recognised the need for assessing and controlling non-sampling errors, especially in large scale surveys. During the sample census on jute crop in Bengal, he developed the technique of interpenetrating network of sub-samples and other techniques for assessing and controlling non-sampling errors. In short, randomization, statistical control and cost were his watch-words and throughout his work in sample surveys, he has been emphasizing these three most important factors in large scale surveys.

The pioneering work of Mahalanobis: Hotelling (1938) in his report to the Indian Central Jute Committee, prepared after a careful examination of Mahalanobis' scheme for a sample census in Bengal states, '... no technique of random samples has, so far as I can find, been developed in the United States or elsewhere, which can compare in accuracy or in economy with that described by Professor Mahalanobis.'

After briefly reviewing Mahalanobis' work in the field of sample surveys, Seng (1951) says:

> 'India can therefore safely claim to rank with the United States as amongst the foremost users of the sampling method in social and economic research. And it is a very happy combination, for in the United States we have the typical example of an industrial and highly developed country while in India the conditions approximate more nearly to those of a country not so highly developed, or more specifically, to the conditions of those countries, which, like China, have no genuine statistics, and where such statistics, if they are to be obtained at all, have to be obtained mainly by sample surveys, for which the experience of India will serve as a guide and as an example worthy of imitating.'

Yates (1951)... said in a talk on the work of the United Nations Sub-Commission on Statistical Sampling, 'He (Mahalanobis) consequently, recognized more clearly than most, that if more world censuses were to be properly carried out in the less-developed countries, the use of sampling method would be essential and it was he who proposed the setting up of the Sub-Commission on Statistical Sampling in order to assist the work of proper application of sampling methods.'

Fisher (1962) ... said at the first convocation of the Indian Statistical Institute, 'I need hardly say that I refer to the emergence of a statistically competent technique of Sample Survey, with which I believe Professor Mahalanobis' name will always be associated. What at first most strongly attracted my admiration was that the Professor's work was not imitative.'

Principles of Sampling Design: Professor Mahalanobis was perhaps the first person to recognize fully the implications of a joint consideration of sampling errors, non-sampling errors and cost aspects of statistical surveys based on the principle of random sampling. From a study of his earlier paper, it is clear that he had used many of the selection and estimation procedures currently being used in sample surveys. Examples of selection procedures used or considered by Mahalanobis in his surveys or papers are simple random sampling, systematic sampling, grid sampling, cluster sampling, varying probability sampling, stratified sampling, multi-stage sampling and multi-phase sampling. As regards the question of costs, he had made extensive studies to build up suitable cost functions for different situations. It can be safely said that he has contributed in a large measure to the rapid development of sampling theory by not only proposing new methods but also by actually demonstrating their potentialities and usefulness in practice.

From the beginning Mahalanobis appears to be very clear about the principles involved in designing a sample for a statistical survey. In his report on sample survey of acreage under jute in Bengal, Mahalanobis (1940, p. 513) gives a comprehensive formulation of the principle of sample designing as follows:

> 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 estimate. For this it is necessary to determine three things, namely,
>
> a) what is the best size of the sample 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 different districts, regions, or zones covered by the survey.

While elaborating on the principle of sample designing for a crop survey taking into account the cost aspect, Mahalanobis (1942b, p. 98) states:

> The total cost being given, we have to find what is the area of each sampling unit; how many such sampling units should be used altogether; how these should be allotted to different classes of land; and how the sampling units assigned to each class of land should be distributed geographically in order that the margin or error in the final results would be reduced to a minimum. Alternatively, we may define the objective in a slightly different way. The order of accuracy in the final results being given, we have to find what would be the best size and distribution of the sampling units in order to do the work at a minimum cost.

An important feature of Mahalanobis' scheme for sample surveys is the phasing of the work relating to the survey by first conducting exploratory studies to get a very rough idea of the nature of the variability of the characteristics under consideration and the cost of different survey operations, and gradually expanding the survey to result in a full-fledged survey, planned efficiently on the basis of the data collected and the experience gained during the exploratory stage of the survey. For instance, in the scheme for a sample survey of jute crop in Bengal as has been explained before, the survey was phased over five years (1937–41) with the full-fledged survey covering the whole of Bengal taking place in 1941. Wald (1947) has drawn attention to this survey in this introduction to his books *Sequential Analysis* and had stated:

> The occasional practice of designing a large scale experiment in successive stages may be regarded as a forerunner of sequential analysis. The idea of such chain experiments was briefly discussed by Harold Hotelling (1941). A very interesting example of this type is the series of sample censuses of area of jute in Bengal carried out under the direction of P.C. Mahalanobis (1940a). Mahalanobis (1952) has discussed the advantages of sequential sampling by considering what he terms, *historical* and *non-historical designs*.

It is of interest to note that Mahalanobis had not only considered estimation of certain population parameters, such as average, population total, ratio, etc. but had also given considerable attention to the question of estimating, what he terms, *abstract distribution* and *space distribution* of the variety with the help of sample surveys. He terms the latter problem as *mapping surveys.*

Theoretical formulation: It is proposed to give here briefly the theoretical set-up developed by Mahalanobis (1944) in connection with the sample survey of different corps in Bengal. In his theoretical formulation he had followed a generalized approach to the basic problems of sample surveys.

Let us consider a geographical region of finite area consisting of a finite number, say N_0, of basic cells of smallest area measured in suitable areal units. The basic cell, termed *quad* and denoted by the symbol □, can be considered to be the elementary unit. A quad can be uniquely identified by a pair of coordinate numbers (i,j) where i = 1, 2, 3, ... l and j = 1, 2, 3, ... , m; where lm = N_0 □ = A, geo-graphical area of the region. Let z(i,j) be the value of the variety z for the quad □ corresponding to the pair of coordinate number (i,j). Mahalanobis defined a *field* as a field in the sense of the present paper thus consists of a finite number, say N_0, basic cells arranged in a definite space or a geographical order together with a single value (or a set of values in the multivariate case) of z, for each basic cell'.

For a particular field, where the value of z for each basic cell is uniquely determined, the set of N_0 values of z is said to constitute an *abstract set* of z. In the case of an actual field that N_0 values of z have a definite space distribution and the statistical properties of this space distribution are of great importance. The distribution between abstract distribution of z and the space distribution of z and be easily understood by noting that corresponding to any given abstract distribution of z, there are $N_0!$ space distributions, if all the N_0 values of z are different.

With this set-up, Mahalanobis has considered the following four types of sampling—unitary unrestricted, unitary configurational, zonal unrestricted and zonal configurational—which, in modern terminology, correspond to unrestricted simple random sampling, unrestricted cluster sampling, stratified simple random sampling and stratified cluster sampling, respectively.

Unitary unrestricted (unrestricted simple random) sampling relates to selection of n basic cells with equal probability. In unitary configurational (unrestricted cluster) sampling, a sample of groups of cells, where the cells in each group are arranged in any particular geographical configuration, which can be termed a *grid*, is selected. For instance, square shaped blocks of 4, 9, 16, ..., m^2 adjoining basic cells are grids of the corresponding number of quads. In case of zonal (stratified) sampling, sampling of cells or groups of cells (grids) is done separately in each of the zones or strata in which the whole field is sub-divided.

Mahalanobis (1944), while discussing the properties of fields, has considered in detail the correlation function and has also extended the theoretical set-up evolved in case of uni-stage sampling, to the case of multi-stage sampling. He has also discussed the nature of occurrence of crop patches in random and non-random fields giving results of suitable empirical studies.

Selection procedures: In his exploratory studies on crop acreage and production, Mahalanobis had used different selection procedures which are commonly being used now. For instance, in 1937 he conducted exploratory studies on crop statistics in 20 *mauzas* (villages) randomly selected with equal probability from *thanas* (groups of villages) and in the plots selected systematically from some of the selected villages. In case of crop yield surveys, he used a multi-stage sample design with *mauzas* (villages), grids within villages, plots within grids and cuts of different sizes and shapes as the different stage units.

It is clear that he was familiar with the method of probability proportional to size sampling even as early as 1937, since in his report on experimental crop census conducted in 1937, he states that for obtaining valid estimates, it would be necessary to select the plots using the cumulative totals of their areas, since the area of plots varies considerably. Proceeding, he states that it would be impracticable to select the plots with this procedure as the workload involved would be excessive. Because of this consideration and due to the fact that the cost of journey between plots widely scattered over the region would be high, he recommended the use of grids, the size of which, he pointed out, should be decided on the bias of empirical studies taking into account the variance and cost functions.

For estimating the yield of cinchona bark Mahalanobis used a multi-phase sample design which he termed multiple survey or multiple sample. In the first phase a number of physical measurements such as girth, height, number of shoots, surface area of the plant, etc. were made on a number of standing plants selected at random. A sub-sample of these plants were uprooted for determining the yield of bark. An estimate of the yield of bark was obtained with the help of regression analysis technique using the measurements of the plants collected in the first phase of the survey.

Another example of multi-phase sampling given by Mohalanobis relates to the crop production survey, where acreage data were collected from a large sample of grids and yield data were collected from a relatively small sample of cuts located at random in a sub-sample of fields taken up for acreage survey. Further in case of crop-cutting surveys the green weight of the harvested crop (say, jute or paddy) was obtained immediately after harvesting, whereas in only a fraction of these case (of the order of 10) was soaked, rested and the dried fibre extracted for weighing or the paddy dried, husked and the weight of rice determined.

It is of interest to note that as far back as 1937, Mahalanobis (1945, p. 97) had considered the possibility of air-surveys using specially sensitized films for estimation of crops acreage.

Grid sampling: As mentioned earlier, Mahalanobis recommended the use of grids as sampling units in crop surveys, since the selection of plots with probability proportional to area after cumulating their areas would be impracticable and since surveying sample plots spread over a large area would be uneconomical. Grids of square shape are located with the help of randomly selected points on the cadastral maps showing plots (or field) boundaries. The random point locates a specified corner (such as south-west) of the grid with its sides in north-south and east-west directions. The location of the random point on the map is done by a simple instrument, termed *coordinatograph*, designed by Mahalanobis.

Since the grid is an artificial sample unit, it may cut across the plot (or field) boundaries. To obviate this difficulty, Mahalanobis considered the possibility of including in the sample all those plots having a major part of their area included within the grid and termed this the *half-unit method*. He recognized that this procedure would be biased and the carried out studies to ascertain the value of this bias for grids of different sizes...

It is clear that the bias and the increase in variance are considerable for smaller grids and these decrease as the size of the grid is increased. Actually, in later surveys the exact method was used instead of this half-unit method, as smaller grids were usually found to be more economical in practice. The exact method in this case consists in considering only those plots or parts of plots which fall inside the selected grid as in the sample. However, in observing the crop proportion for border plots in the earlier crop surveys, the crop proportion observed for the plot as a whole was applied to the part of the plot inside the grid to facilitate field work.

Variance function: As has been pointed out earlier, Mahalanobis has stressed the importance of studying the variance function on the basis of exploratory studies and model sampling experiments for designing sample surveys efficiently. For instance, in case of the jute survey in Bengal he had carried out extensive studies on the variance and the cost functions with a view to determine the optimum size of the sampling unit, which is the grid in the case of crop acreage survey and which is the sample cut in the crop yield survey and the optimum geographical density of sample points.

Crop acreage survey: Mahalanobis clearly realized that the variance function in crop acreage surveys would depend on p, the proportion of the area under the crop in the region, x, the grid size and y, geographical density of sample units and that the variance would decrease with increase in x and y. One the basis of empirical studies he conjectured that the variance function of the estimate of the proportion of the area under the crop in case of sampling of one grid be of the form

$$Vslubx = \frac{pq}{(b.x)^g}$$

where b and g are constants.

It is of interest to note that the variance function assumes the familiar binomial form pq, when $x = 1/b$. This gives physical interpretation to the constant b, which according to Mahalanobis, is the reciprocal of the largest possible unit area reporting 1 or 0 for the crop proportions. The relative variance of the grid sampling, compared to the binomial case, is given by

$$E_x = \frac{1}{(bx)^g} = \frac{a}{x^g}$$

where a and g are constants ...

The value of g, in a number of empirical studies turned out to be less than 1, though it varied from zone to zone. Mahalanobis pointed out that the value of g is determined by the degree of correlation between the crop proportion in neighbouring plots. In fact the intra-gid correlation and the parameter g are connected by the relation

$$\rho(n_0 - 1) = (n_0)^{1-g} - 1$$

It was also found that the value of g decreased with an increase in the value of p, the crop proportion.

Crop yield survey: As regards the question of shape and size of crop cuts in estimation of yield rates in crop surveys, Mahalanobis has conducted extensive studies and a comprehensive account of these studies is given by Mahalanobis and Sengupta (1951). It is of interest to note that in the earlier exploratory surveys, a sample cut was usually harvested in the form of sub-cuts, which helped in studying the behaviour of the variance with changes in the size of the cut. The shapes of cuts considered are triangular, rectangular, square and circular. Mahalanobis, following Hubback, has recommended the use of small cuts in preference to large cuts in crop yield surveys from the point of view of both operational convenience and accuracy. He has recognized the possibility of getting over-estimates of yield rate from very small cuts and has demonstrated by actual experimentation that the bias becomes negligible for circular cuts as the radius is increased to 4′...

Another illustration of the study of the variance function in case of crop yield survey with a multi-stage design is provided by the survey of 'aman' paddy in West Bengal in 1949–50. The design was a three-stage one with locality (defined as circular area of radius one mile demarcated round a centre located at random within a police station), grids of size 2.25 acres and concentric circular cuts of radii 2′, 4′ and 5′ 8″ as first, second and third stage units, respectively.

... the decrease in the standard deviation in increasing the radius of the circular cuts from 2′ to 4′ is more than that achieved by increasing the radius from 4′ to 5′ 8″. Further, the mean yield rate is considerably higher for circular cuts of radius 2′ than that for circular cuts of radii 4′ and 5′ 8″, which agree fairly closely. Thus, Mahalanobis and Sengupta (1951) showed that, though the cuts of very small size possibly give rise to an over-estimate of yield, the bias becomes negligible as the radius of the circular cut is increased to 4′.

Even a brief description of Mahalanobis' work relating to the development of crop yield survey in India would be incomplete without reference to the long standing technical controversy in India on the issue of the shape and the size of cuts to be used in crop yield surveys. After considerable experimentation with cuts of different shapes and sizes, Mahalanobis recommended the use of concentric circular cut of radius 4′ for yield surveys and the Indian Statistical Institute (ISI) has been using the circular cuts in the National Sample Survey and other surveys.

As against this, the Indian Council of Agricultural Research (ICAR) has been using the rectangular cuts of size 33′ × 16.5′ in their crop yield surveys conducted through the state agency. Mahalanobis has shown keen interest in resolving the technical controversy on this issue and has been suggesting joint studies by both the ISI and the ICAR. In this connection Mahalanobis (1946b, p. 279) states in his paper on sample surveys of crop yield in India:

> I may mention, however, that for some considerable time I have been pressing on ICAR authorities the need for carrying out crop cutting work by both ISI and ICAR methods in the same region with a view to studying the relative efficiencies of the two systems. I have submitted definite schemes for this purpose which are under consideration by the Government of India.

He reiterated his suggestion for joint studies during the Fifth Session of the United Nations Sub-Commission on Statistical Sampling held in Calcutta in 1951.

Cost function: One of the main considerations, which led Mahalanobis to use the grid as the sampling unit in crop acreage surveys and a multi-stage design with cuts of small size as ultimate sampling units in case of crop yield surveys, was cost. From the beginning he has emphasized the importance of cost studies in statistical surveys and has himself carried out extensive studies to evolve suitable cost functions for different situations. In crop surveys, he considered the cost function to be made up of four components:

i) 'enumeration', consisting of identifying the sample girds and noting down the proportion of area under the crop;

ii) journey', comprising all journeys for the survey;

iii) 'miscellaneous', comprising preliminary arrangement, copying of field records, attending to supervisory officers, despatch of data collected, etc.; and

iv) 'indirect', which is the residual category comprising recreation, sleeping, etc.

As mentioned earlier he recognised that the total cost would depend on the grid size x and density of grids y. Further, the deduced and empirically verified that the expected value of the distance between n points selected at random in a region of A units of area would be proportional to

$$\sqrt{A}\left(\sqrt{n} - \frac{1}{\sqrt{n}}\right)$$

Taking cost functions for the first three cost components, which directly relate to the survey operation as

$$E = a_1 + b_1 y, \quad J = a_2 + b_2\sqrt{y,} \quad M = a_3 + b_3 y,$$

where E, J and M are the cost (or time) per square mile, the total cost (or time) of field work can be written as

$$w = A + B\sqrt{y} + Cy,$$

where w is the total cost (or time) per square mile and A, B and C are constants, which depend on the grid size and density of grids.

It can be seen that reduction of cost in decreasing grid size from 9 acres to 6 acres or 4 acres is substantial, whereas reduction in cost becomes marginal for decreases in grid size below 4 acres. Further, from Table 6, giving the cost per square mile for different levels of error and grid sizes, it is clear that the optimum grid size is 4 acres.

It is to be noted that Mahalanobis considered the cost of statistical operation as important as the field cost and took into account both field and tabulation costs in arriving at optimum grid size and density of grids.

Errors in surveys: The most important contribution of Mahalanobis to the field of statistics in general and to sample surveys in particular, which he himself considered to be of considerable value, is the development of the techniques for assessing and controlling errors in censuses and surveys. He was perhaps the first person to fully realise the implications of sampling and non-sampling errors in statistical surveys. The theory of sampling, developed till 1936, related mainly to the study of sampling errors. He has considered the total margin of errors in the results of a sample surveys to consist of the three components:

i) sampling errors arising due to variation from sample to sample;

ii) error due to physical fluctuations in observation, measurement and tabulation; and

iii) mistakes such as deliberate wrong recording of the data.

Differentiating the components (ii) and (iii), Mahalanobis (1945, p. 37) states: '... mistakes or false entries are more dangerous as they are not amenable to statistical treatment. Special precautions have to be taken and suitable statistical checks and controls have to be incorporated in the design of the survey to detect (and hence to discourage) such dishonest work'.

It is proposed to consider briefly, giving illustrations, the following techniques of assessing errors in surveys, which have been proposed and developed by Mahalanobis:

i) interpenetrating (net-work of) sub-samples;

ii) duplicated samples;

iii) use of variance function; and

iv) inspection, scrutiny and verification.

It may be pointed out the Mahalanobis has always emphasized the need for controlling sampling and non-sampling errors rather than their complete elimination, since reduction of errors beyond a certain stage may be not only costly but also unnecessary as long as the margin of error in the estimate is within what he has termed *permissible error* for the purpose in view.

Interpenetrating sub-samples: Since 1936, Mahalanobis has been using the technique of interpenetrating (net-work of) sub-samples in assessing and controlling the errors in surveys. This technique, in its general form, consists in drawing the sample in the form of two or more sub-samples so as to be able to get a valid estimate of the population parameter under consideration on the basis of each sub-sample and arranging to get the work of data collection and tabulation for these sub-samples done by different parties of investigators and computers. This procedure helps in analyzing the total variation in the results into its components, namely, sampling error, ascertainment error and tabulation error. Variation between k sub-sample estimates provides an idea of the total margin of error of the estimate including both sampling and non-sampling errors. In fact, in case of estimators for which the usual variance estimator is rather complicated, the estimate of variance based on the sub-sample estimates is the simplest way of getting an idea of the total margin of error. Further, minimum and maximum estimates based on the k sub-samples provides a confidence interval for the median (and the mean, in case of a symmetric distribution) of the estimator with a confidence coefficient of $1 - (1/2)^{k-1}$.

Discussing the question of assessment of errors in surveys, Mahalanobis (1945, p. 37) describes the technique of interpenetrating sub-samples as follows:

> One type of control has proved extremely useful in Bengal. The total number of grids in each zone is divided into equal portions say (A) and (B). The grids allotted to sub-sample (A) are scattered at random all over the zone; and, in the same way, the grids allotted to sub-sample (B) are also scattered over the whole area; the two sub-samples (A) and (B) are thus completely mixed up and inter-penetrate into one another. The information on the two sub-samples is collected by two entirely different sets of field investigators who work independently and at different times in the same zone so that they never meet. How far they are in agreement immediately furnishes a good idea about the reliability (or otherwise) of the results.

This technique can well be applied to bring out the variation between investigators, methods of data collection, tabulation procedures and variation over time. Mahalanobis has used linked sub-samples to study the differential bias of investigators, as it would increase the precision of comparison of estimates supplied by different investigators. In the exploratory surveys, linked pairs of grids were located on the maps in the form of dumb-bell shaped figures, one end of each figure represented the grid belonging to sub-sample 1 and the other end belonging to sub-sample 2. The two sub-samples were surveyed by two different sets of investigators independently.

The technique of interpenetrating sub-samples may be used as a broad check on the different operations involved in a large-scale sample survey. For instance, if estimates based on *k* different sub-samples surveyed and tabulated by different sets of investigators and computers agree fairly closely, it can be safely assumed that the survey operations have been under statistical control. However, it may be noted that if there is some systematic error which is common to the different sets of investigators and computers, it would not be shown up by this procedure. On the other hand, if one of these estimates differs substantially from the other $k - 1$ estimates, it is clear that this sub-sample estimate is unreliable and hence it is necessary to check up the primary data and verify the calculations for that sub-sample, on which this particular estimate is based. Thus, it is seen that it is possible to take corrective action on the basis of the sub-sample estimates, thereby increasing the accuracy and the utility of the final results.

Another illustration of the use of the technique of interpenetrating sub-samples for estimating the total margin or error is given by the Bengal Labour Enquiry (Mahalanobis, 1946a), where the field work was carried out in the form of a Latin square arrangement with five blocks (groups of houses) and five investigators. Here the estimator for cost of living index, for which the usual variance estimator is quite complicated, is considered and the standard error of this estimator is calculated on the basis of the sub-sample results.

The results of the enquiry into the tea-drinking habits of middle class Indian families in Calcutta (Mahalanobis, 1943) where the sample, selected in the form of four interpenetrating sub-samples was surveyed by two sets of investigators.

In the National Sample Survey (NSS) the sample is usually selected in the form of four or more sub-samples, half of which are surveyed by the Central agency and the other half by the State agency and within each agency half the number of sub-samples are surveyed by one party of investigators and the other half by another party of investigators. The data collected are also tabulated separately by sub-samples by the two agencies. Since two or three years the sub-samples allotted to the Central agency are also being tabulated independently in different tabulation centres at the instance of Mahalanobis.

In the 8th round in the NSS, conducted during 1954–5, the Central and the State agencies surveyed four and eight sub-samples respectively for the land holdings enquiry. The design of the survey was a stratified two-stage one with villages and households as first and second stage units respectively.

The technique of interpenetrating sub-samples is also very useful in getting an idea of the margin or error in the case of estimation of distributions such as concentration curves and mapping problems. Mahalanobis (1958, 1960) proposed that use of the area between the two estimated distributions based on the two interpenetrating sub-samples as a measure of the total margin of error in the estimated distribution based on the combined sample.

Mahalanobis recognized that the use of the technique of interpenetrating sub-samples in practice requires additional care at different stages of the survey and that the cost of the survey could also increase to some extent. As regards the cost, he has pointed out that only the journey cost in field work is likely to increase and that this additional cost, which would be a small portion of the total cost including both field and tabulation costs, would be worthwhile considering the advantages of this technique. The various aspects of this technique have been discussed in detail by Mahalanobis (1940a, 1944, 1945, 1946a, 1958, 1960), Mahalanobis and Lahiri (1961) and Lahiri (1954, 1958a and 1958b).

In the preface to his book, in which he has extensively used the technique of interpenetrating sub-samples for estimation of errors, Deming (1960) states: 'In respect of new methods, the chief contribution is replication by procedures that maintain efficiency, yet facilitate the estimation of standard error and of any other errors that exist in the estimator used. The method is essentially based on Mahalanobis' interpenetrating sub-samples, which he introduced in 1936 in his surveys in Bengal'.

Duplicated samples: Another technique, which is commonly used to assess non-sampling errors is to resurvey the sample units using better trained personnel. In a sense this may be considered as a particular case of the technique of interpenetrating sub-samples, since in this case the interpenetrating sub-samples are identical and surveyed independently by two sets of investigators. In the exploratory stage of the crop survey in Bengal, this method was extensively used to bring out the importance of non-sampling errors. This method was also used in the Bihar Crop Survey 1943–4, where 25 per cent of the samples was re-enumerated independently by another set of investigators.

> ... it can be seen that both the net and gross errors are fairly large for most of the comparisons, though the gross error is, in most cases, considerably larger than the net error. This shows that the non-sampling errors are fairly large and that proper precautions and suitable statistical controls are necessary to reduce such errors in sampling.

Control through variance function: Another method used by Mahalanobis in assessing and controlling errors in surveys relates to the use of the variance function. For instance, in the case of crop yield surveys the variance of the estimate of yield rate based on sample cuts should decrease with an increase in the size of the cut and this result is made use of in assessing the quality of field work by instructing the investigators to harvest the cut in the form of sub-cuts. That is, while adopting a circular cut of radius 5′ 8″, the crop within a circular cut of radius 2′ is harvested first, then the annular portion between circular cuts of radii 2′ and 4′ and finally the annular portion between circular cuts of radii 4′ and 5′ 8″. Discussing this type of control Mahalanobis (1946b, p. 276) states:

> A comparison of mean values based on different sizes of cuts shows whether or not any bias had arisen in using different sizes of cuts in the process of harvesting. Cuts of different sizes would, however, in general show decreasing variance with increase in size. A comparison of variances for cuts of different sizes thus supplies another valuable control. For example, it is sometimes found that mean values based on cuts of different sizes are in excellent agreement (often much better than one could expect from probability considerations), and variance of yield rates for cuts of different sizes are also about equal. In such cases it is practically certain that cuts of only one size had been actually harvested and records for cuts of other sizes were obtained by paper calculation. Multiple cuts of other shapes and sizes have also been used in crop-cutting work, and found to be quite useful.

Inspection, scrutiny and verification: Mahalanobis has also attached considerable importance to the conventional forms of error control through inspection, scrutiny and verification. For instance, even in the case of such a large scale survey as the NSS, the ratio of inspectors to investigators is 1 : 4. He has emphasized the salutary effect that the very idea of inspection, scrutiny and verification has on the quality of work. He had tried out various forms of inspection such as inspection prior to and after data collection by investigators and accompanied inspection. He had made it a policy of the Institute to completely verify all computations after introduction of dummy mistakes to minimize the error at the tabulation stage. The quality of work of the verifiers is judged on the basis of the proportion of dummy mistakes they have been able to spot out. In the NSS, a two-tier scrutiny system has been developed, in which the data collected are scrutinized once in the field offices and then again at the tabulation stage. The aim of field scrutiny is to find out mistakes and discrepancies which need verification in the field and which can possibly be rectified by further investigation, whereas the scrutiny at the tabulation stage aims at detection of errors which can be rectified without the need for further field work and at studying the general quality of the data collected with a view to improve upon the format of schedules, instructions and methods of data collection in future surveys.

Bengal crop survey: As has been pointed out earlier, the main difference between the development of sample survey methods in India by Mahalanobis and in other countries has been the scale of operations with which the theory of sampling has been applied to practical problems. Starting with exploratory surveys confined to an area of a few square miles in Bengal in 1937, Mahalanobis was perhaps the first person to organize and carry out an objectively designed large scale sample survey covering the whole of Bengal (about 59,000 square miles) in 1941. The magnitude of the scale of operations and the difficulties involved in organising the jute acreage survey in Bengal in 1941 can easily be imagined if we note that a total of about 102,000 sheets of maps covering about 50,000 square miles had to be collected and used for selecting about 58,000 sample units (grids) after suitable stratification and that a staff of about 350 field investigators and about the same number of statistical workers were employed in this project. As mentioned earlier Mahalanobis has termed the task of organizing and carrying out such large scale surveys as *statistical engineering*.

A comprehensive description of the stages of work involved in the Bengal jute survey plan, as given by Mahalanobis (1940a, p. 512), is as follows:

> In this plan the whole of the area to be surveyed (in this case roughly 55,000 square 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 conveniently be 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 form the grids within the zone can be taken to be the representative figure of the zone. Multiplying by the total area of each zone (which is known) it is then possible to estimate that 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.

Exploratory studies: Recognizing fully the difficulties in organizing and carrying out large scale sample surveys, Mahalanobis had been strongly advocating the need for exploratory and pilot studies before taking up a large scale survey. During the exploratory stage, studies can be carried out to evolve suitable sample design taking into account both variance and cost aspects and to evolve suitable field and statistical organisations. As has been mentioned earlier, the Bengal Crop Survey was taken up in a phased manner since 1937 which helped in carrying out the sample survey on a large scale covering the whole of Bengal in 1941. The experience gained in the Bengal Crop Survey was of considerably help in organizing and conducting efficiently the crop survey in Bihar in 1943–44.

Popularization or large scale surveys: Having foreseen the potentialities of large scale sample surveys, Mahalanobis helped considerably in the development of sample survey theory and methods by popularizing the random sample method through demonstration of its utility in practical problems.

The scores of statistical surveys presently being conducted annually in India and the facilities that are now available in this country for training and research in sample surveys stand testimony to the extent of Mahalanobis' success in propagating the idea of using properly designed sample surveys to meet the growing needs for statistical data to be used in planning for national development and for other purposes.

It will not be out of place to give here two important illustrations of how Mahalanobis proceeded about popularizing large scale sample surveys by actual demonstration of its usefulness.

At the request of the Indian Central Jute Committee, Mahalanobis prepared a scheme in 1936 for a sample survey of jute crop in Bengal which envisaged a gradual expansion of the surveys with a view to having a large scale sample survey in 1941 covering the whole of Bengal. The Jute Census Committee laid down three tests for the success of the scheme:

i) that the margin of error of area under jut should not exceed 5 per cent

ii) that the results be available by the first week or second week of September; and

iii) that the cost of sample survey should not be excessive.

Mahalanobis amply demonstrated that it was possible to successfully meet these tests laid down by the Jute Census Committee by carrying out the sample survey of jute crop in Bengal such that

i) the margin of error was only 2.8 per cent well below the stipulated margin;

ii) the results were made available on 27 August, well ahead of the target date; and

iii) the cost of the sample survey was Rs 1.14 lakhs which compared very favourably with the cost of Rs 15 lakhs for a complete enumeration.

This success of Mahalanobis is of considerable importance, as this demonstration of the usefulness of random sampling went a long way, especially in India, in popularizing the method.

Another demonstration relates to the comparision of the results of a well designed sample survey of jute crop in Bengal conducted under Mahalanobis' guidance and direction and those of a complete enumeration conducted by the official agency with the very reliable trade figures which became available subsequently.