Mahalanobis' Contributions to Sample Surveys

The Origins of Sampling in India

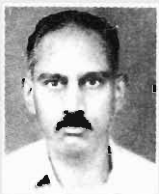
T J Rao

Introduction

Prasanta Chandra Mahalanobis was the chief architect of the post-independence statistical system in India. He was appointed as the Honorary Statistical Adviser to the Government of India in 1949. A Central Statistical Unit was set up in 1949 to function under his technical guidance. In 1951 the Central Statistical Organization (CSO) was established to coordinate the statistical activities in various ministries and other governmental agencies. Around the same time, the Standing Committee of the Departmental Statisticians as well as the National Income Committee felt the urgent need for quality data and recommended a sample survey covering rural areas of India. Thus the National Sample Survey started its operations in October 1950 under the leadership of Mahalanobis, the technical expertise being provided at the Indian Statistical Institute (ISI). Thus both the apex statistical bodies of independent India, the Central Statistical Organization (CSO) and the National Sample Survey (NSS) and the State Statistical Bureaus are the creations of Mahalanobis.

Contributions to Sample Surveys

According to C R Rao, "*the fame of Mahalanobis as a scientist will rest largely on his contributions to statistics. He viewed Statistics, or more generally collection and processing of information, as essential in seeking truth*". On the work done at the ISI, Fisher remarked: "... *The ISI has taken the lead in the original development of the technique of Sample Surveys, the most potential fact finding process available to the administration*". In particular, Mahalanobis' innovative techniques and methodology for large scale sample surveys are widely acknowledged throughout the world. On the complexity of the multipurpose, multi-subject framework of NSS, Deming



T J Rao is a Professor of Statistics in the Theoretical Statistics and Mathematics Division of the Indian Statistical Institute at Calcutta. His main research interests are in sample surveys theory and practice. He has more than 60 research publications in various journals and is a Member of the International Statistical Institute. Rao is on the Governing Council of the National Sample Survey Organisation and is the Managing Editor of *Sankhya*, Series B as well as a coeditor.

commented thus: "... *No country, developed, under-developed or over-developed, has such a wealth of information about its people as India....*".

During the early twenties, a civil servant, J A Hubback was dissatisfied with the inadequate and defective methods of collection of data on crop acreage and yields and devised crop-cutting experiments which he termed 'a random sampling method'. Mahalanobis was influenced by Hubback's method and adopted random cuts for estimating the acreage under jute crop in Bengal during the early forties. Mahalanobis's work was assessed by H Hotelling in his report submitted to the Indian Central Jute Committee as follows : "... *no technique of random sample 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...*". Mahalanobis' sample survey estimate of jute production was 7540 bales (1 bale = 400 lbs) obtained at a cost of 8 lakh rupees with a work force of 600 while the plot-to-plot enumeration yielded a figure of 6304 bales at an expenditure of 82 lakh rupees and 33,000 employees which turned out to be an underestimate by 16.6%. This was evidenced by the alternative customs and trade figure of 7562 bales.

Just as the large-scale commercial production of a chemical is a matter of chemical engineering rather than of pure chemistry, the organization of large scale sample surveys was equated to 'statistical engineering' rather than pure theory of sampling by Mahalanobis.

During the early period 1937–45, Mahalanobis introduced several innovative ideas and methodologies in what he called as 'experiments in statistical sampling'. He mainly dealt with the problems of organization which arise when a sample survey has to be carried out on a very large scale. Just as the large-scale commercial production of a chemical is a matter of chemical engineering rather than of pure chemistry, the organization of large scale sample surveys was equated to 'statistical engineering' rather than pure theory of sampling by Mahalanobis. These experiments covered wide ranging areas such as acreage and total production of important food and fibre crops, economic or demographic factors relating to indebtedness, unemployment, birth and death rates etc. of rural families, cost and level of living, consumption of food, clothes etc., preferences to particular commodities, public opinion, after-effects of Bengal famine,

traffic surveys, demand for currency coins and average life of currency notes, study of bark yield using regression estimates etc..

The above mentioned experiments in sampling stressed the importance of 'cost function' and 'variance function'. Obviously, as the sample size increases the variability in the estimates decreases but the cost of the survey tends to increase. Thus a balance has to be struck which leads to an 'optimum' sample size. 'Pilot surveys' played an important role in determination of sample size as well as for testing the schedules and field conditions, estimating the time and cost for survey and the variability.

When the units are of varying size, Mahalanobis was aware of sampling of units with probabilities of selection proportional to their sizes instead of equal probabilities in 1937 itself. He, however, abandoned this method of cumulating the sizes and recording the cumulative totals due to constraints of work load. The mathematical theory for Probability Proportional to Size Sampling (PPS) method was later given by Hansen and Hurwitz in 1943.

Mahalanobis in collaboration with D B Lahiri of the NSS presented a detailed analysis of errors in censuses and surveys in the Indian context.

Mahalanobis in collaboration with D B Lahiri of the NSS presented a detailed analysis of errors in censuses and surveys in the Indian context. The technique of Inter Penetrating Network of Subsamples (IPNS) developed by Mahalanobis during the thirties, consists in drawing the sample in the form of two or more sub-samples, selected according to the same sampling scheme so that each subsample provides a valid estimate of the parameter of interest. This technique helps in providing 'a means of control (i.e. appraisal) of the quality of the information', by way of securing information on non-sampling errors. This technique used by Deming as 'replicated sampling' was acknowledged to have 'simplicity in calculation of the standard error of the estimate' besides ability 'to detect gross blunders in selection, recording and processing'.

The Impact

The three notable contributions to sample survey methodology by Mahalanobis, namely 'pilot surveys, concept of optimum survey design, and inter penetrating network of subsamples (IPNS)' had a great impact on the present day sampling techniques in particular and statistical methods in general. For example, pilot surveys are acknowledged as a prelude to Abraham Wald's 'sequential analysis' which relates to decision making sequentially. 'Optimum survey design' stresses the Mahalanobisian philosophy that all the resources provided for a survey should be used optimally going beyond the mathematical propositions such as 'sampling error should be minimized for a fixed cost', or 'cost should be minimized for a fixed sampling error'. This can be considered to be a precursor to the present day 'operations research' philosophy. The IPNS technique, while being a tool for assessing and controlling the non-sampling errors in the survey, also 'permits evaluation of variances between investigators, coders and other workers in the various stages of processing'. Thus IPNS technique could be considered really as the curtain-raiser for 'resampling procedures' like Bootstrap.

IPNS technique could be considered really as the curtain-raiser for 'resampling procedures' like Bootstrap.

It is interesting to note that as early as in 1937 itself, Mahalanobis had considered the possibility of air surveys 'using specially sensitized films' for estimation of crop acreage. This is exactly what is done through 'remote sensing' satellites now-a-days.

The method of collection of data for the 'United Provinces Anthropometric Survey' of 1941 organized by Mahalanobis, D N Majumdar and C R Rao was described as follows : "*The random selection of a sample is not an easy task. To pick up a sample in a demonstrably rigorous fashion requires elaborate preliminary arrangements which are not often possible in practice. In this situation, rough and ready methods have to be used*". The anthropologist, D N Majumdar 'collected all healthy males between the ages of 18 and 48 (belonging to the caste or tribe under survey) who happened to be available, *arranged them in serial order just as they came and picked up either the odd or the even numbered individuals for*

measurement.' Summing up, Mahalanobis, Majumdar and Rao in their paper in *Sankhyā*, 1948 say "The present samples may therefore be treated as having been drawn, for all practical purposes, at random. As far as one can judge, the assumption of randomness is more true of the present material than any other series of anthropometric measurements so far available in India". Majumdar's selection involved a 'random start' from 1 and 2 and selection of a 'systematic sample' with the 'sampling interval' equal to two, a technique which has been developed by Madow and Madow in 1944. The 'arrangement' of individuals by Majumdar 'in a serial order just as they came in' makes systematic sampling equivalent to 'simple random sampling' as stressed by the authors. This was also formally shown by Madow and Madow later and led to the concept of 'random permutations model'. The use of random permutations model in drawing inferences in sampling from finite populations is very well known.

Mahalanobis defines 'group' as individuals belonging to the same caste, religion or tribe and living in the same district, giving rise to a two-way classification.

Following the U P Anthropometric Survey, D N Majumdar and C R Rao carried out the Bengal Anthropometric Survey in 1945 on similar lines. In the foreword to the paper on the statistical study published in *Sankhyā* in 1958 by D N Majumdar and C R Rao, Mahalanobis defines 'group' as individuals belonging to the same caste, religion or tribe and living in the same district, giving rise to a two-way classification. He then points out that any further sub-division on the basis of sub-caste, clan, endogamy etc. would have resulted in a large number of such groups for which a much larger survey would be necessary to obtain sufficient number of individuals under each sub-group for *proper* statistical analysis. This also arises in the problems of *small domain estimation*, a technique which is of prime importance in the present day context in view of decentralized planning and data dissemination in several countries.

Even from late forties, controversies existed regarding the size of cuts to be used for crop cutting experiments followed by the Indian Council of Agricultural Research (ICAR) and ISI. Mahalanobis experimented on different sizes of cuts and preferred

circular cuts of radius 4 feet for yield surveys, while ICAR under Panse's guidance was using rectangular cuts of size 33feet × 16.5feet for surveys conducted by state agencies. Panse emphasized that any sampling method must fit well into the existing administrative set up, while Mahalànobis stressed on well-trained investigators specially recruited for the survey. B P Adhikari connects this to the observation that Panse and Mahalanobis belonged to two different social backgrounds – one from a place where the Moghul system of revenue collection had its influence and the other where the British system of revenue collection was in vogue. Also joint studies of the Ministry of Agriculture, CSO and ISI conducted in 1960-61 and the studies by a technical committee set up by the Planning Commission conducted on four crops during 1963-66 did not reveal any significant differences between the two types of cuts.

Mahalanobis insisted that, "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"

Towards Human Welfare

Mahalanobis believed that the ultimate analysis of statistics has one single aim : '*to improve the efficiency of action programmes for the welfare of humanity*'. In 1971 he observed: "The use of sample surveys is spreading rapidly in underdeveloped countries .... But the danger still remains of much waste of resources in work which is highly imitative of advanced countries ...". He was always conscious of costs in a survey and insisted that "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." In contrast to Neyman, Mahalanobis derived the optimum allocation in stratified sampling when costs varied from stratum to stratum. Lahiri comments that Mahalanobis was of the view that Neyman was not conscious about costs because the latter was possibly not thinking of stratification for purposes of field surveys where costs might differ from stratum to stratum, but only for sample selection in office for sample tabulation in which case the costs would not vary from stratum to stratum.

On the occasion of the birth centenary of Mahalanobis in 1993,

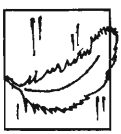
the Government of India released a postage stamp bearing his picture and the Institute he founded in a fitting recognition to his fundamental contributions to statistics towards human welfare and national development. C R Rao thinks that, perhaps, this is the only instance where a statistician has been honoured with the issue of a postal stamp bearing his or her picture. As per the norms, government buildings are rarely named after individuals but for a few rare exceptions; it is indeed a worthy exception to name the building of the National Sample Survey Organisation at its Calcutta Head Quarters as 'Mahalanobis Bhavan'.

Suggested Reading

[1] **D B Lahiri, Prasanta Chandra Mahalanobis and Large Scale Sample Surveys, *Sankhyā*, 35(suppl.), 27–44, 1973.**

[2] **C R Rao, Statistics must have a purpose, the Mahalanobis dictum, *Sankhyā*, 55, 331–349, 1993.**

Address for Correspondence
T J Rao
Theoretical Statistics and Mathematics Division
Indian Statistical Institute
Calcutta 700 035, India.



" ... But, as in all human endeavors, it is best to remember the pioneers. Just as today's laser disk technology provides recorded music infinitely superior to the scratchy sound of a nineteenth-century phonograph, so, too, has modern probability theory shortened and simplified Bernoulli's proof of the law of large numbers. Yet we still revere Thomas Edison ... . It is only fitting that we accord the same respect to Jakob Bernoulli for the golden theorem of which he was so justly proud."

William Dunham, in *The Mathematical Universe*, John Wiley & Sons, 1994.