P. C. Mahalanobis

C. R. Rao

Prasanta Chandra Mahalanobis (1893-1972) was born in a well-known family of Brahmos (a protestant theist movement within the fold of Hinduism) in Calcutta and had his early education in Brahmo Boys' School. After earning a BSc with honors in physics in 1912 from Presidency College, Calcutta, he went to Cambridge, where he took part I of the mathematics tripos in 1914 and part II of the physics tripos in 1915. Awarded a senior scholarship by King's College, Cambridge, he intended to work with the physicist C. T. R. Wilson at the Cavendish Laboratory, but upon returning to India for a short vacation he found so much to hold his interest that he remained there. Shortly after his return, Mahalanobis fell in love with the 16-year-old Nirmalkumari (or Rani as she is popularly known), and they were married in 1923.

Phenomenal activity

Mahalanobis was offered a teaching position in the physics department of Presidency College at Calcutta soon after his return from Cambridge. He joined the Indian Educational Service (IES) in 1915, held the post of professor of physics for a long time, and later became principal of Presidency College. He retired in 1948. During his life he held several distinguished posts, many of them simultaneously. He was the chief executive of the Indian Statistical Institute as secretary and director continuously from the founding of the institute in 1931; he was honorary statistical advisor to the government of India from 1949; and he was associated with the work of the Planning Commission as a member (1955–1967). He also served as head of the department of statistics of Calcutta University (1941–1945) and statistical advisor to the government of Bengal (1945–1948). He was a member of the UN Statistical Subcommission on Sampling from the time it was formed in 1946 and its chairman from 1954 to 1958; he was general secretary of the Indian Science Congress (1945–1948) and later its treasurer (1952–1955) and president (1949–1950); and he was president of the Indian National Science Academy (1957–1958) and editor of *Sankhyā: The Indian Journal of Statistics*, from its foundation in 1933.

An original thinker, Mahalanobis was suspicious of accepted knowledge, chose challenging problems irrespective of subject matter, and approached them in unconventional ways. According to him[1], 'the value of science to society lies in its unorthodoxy and ability to challenge accepted concepts and theories'.

P. C Mahalanobis (1893–1972)

When Mahalanobis founded the Indian Statistical Institute in 1931, it employed one part-time computing assistant and had an annual budget of $50. At the time of his death, forty years later, the institute had nearly 2000 employees and an annual budget of $2.5 million. The development of the institute as a teaching and research organization of international importance and recognition was itself a Herculean task.

Scientific contributions

While in Cambridge, Mahalanobis came across volumes of *Biometrika*, edited by Karl Pearson. He got so interested in them that he took a complete set back to India. He started reading them on the boat during his journey and discovered that statistics was a new discipline capable of wide application. Upon arrival in India, he looked for problems to which he could apply statistics. The resulting highly interesting problems in meteorology and anthropology marked the turning point in his career from a physicist to a statistician.

Mahalanobis distance

The first opportunity to use statistical methods came to Mahalanobis when N. Anandale (then director of the Zoological Survey of India) asked to analyse anthropometric measurements taken on Anglo-Indians (of mixed British and Indian parentage) in Calcutta. This study led to Mahalanobis's first scientific paper[2], and it was followed by other anthropometric investigations leading to formulation of the D^2 statistics[3,4], often known in the statistical literature as *Mahalanobis distance* and widely used in taxonomic classification[5].

Suppose each individual of a population is characterized by p measurements, perhaps anthropological, but the approach is generally applicable. The means (averages) of the p measurements can be represented as a point in a p dimensional space. Corresponding to k given populations, we have a configuration of k points in the p space. If $p=2$, the points can be represented on a two-dimensional chart, and the affinities between populations as measured by nearness of mean values can be graphically examined. We may find that some populations are close to each other in the mean values of the characters studied, thus forming a cluster. The entire configuration of points may then be described in terms of distinct clusters and relations between clusters. We can also determine the groups forming clusters at different levels of mutual distances. Such a description may help in drawing inferences on interrelations between populations and speculating on their origins.

Reprinted with permission.

As p grows larger than 2, however, graphical examination becomes difficult or impossible. How might one sensibly measure the difference between the mean vectors of, for example, populations 1 and 2? Call these vectors δ_1 and δ_2. A direct approach — one that had been used by anthropologists — was to work with the squared length of the column vector $\delta_1 - \delta_2$,

$$(\delta_1 - \delta_2)' (\delta_1 - \delta_2),$$

but this approach pays no attention to internal population covariation among the coordinates, nor to heterogeneity of variances. A related disadvantage of this simple sum-of-squares distance is its noninvariance under linear transformations of the coordinate system.

Mahalanobis, perhaps motivated by his studies of mathematical physics, suggested a more useful kind of distance, providing that the dispersion (variance–covariance) matrix of the two populations is the same, say Λ. The Mahalanobis distance, traditionally given in squared form and called D^2, is

$$D^2 = (\delta_1 - \delta_2)' \Lambda^{-1} (\delta_1 - \delta_2),$$

so that D^2 is in a sense the standardized squared difference between δ_1 and δ_2, generalizing the univariate $(\mu_1 - \mu_2)^2/\sigma^2$, where σ^2 is the common variance and μ_1 and μ_2 are mean values.

If Λ is a common dispersion matrix for all k populations, one may then examine the $k(k-1)/2$ different D^2's and base on them various statistical analyses, including cluster analysis. One large-scale anthropometric cluster analysis was published by Mahalanobis and others[6] in 1949. These developments were closely related to research by R. A. Fisher and by Harold Hotelling. The approach is especially useful when the clusters of points for each population are, roughly speaking, ellipsoidal and with about the same shape from population to population.

Fuller and more precise statements of the above exposition will be found in Mahalanobis's writings listed in the bibliography and in chapter 9 of my monograph[7] *Advanced Statistical Methods in Biometric Research.*

Mahalanobis argued that inferences based on distances among populations might depend on the particular measurements chosen for study. The configuration may change, and even the order relations between distances may be disturbed, if one set of measurements is replaced by another set. Mahalanobis therefore laid down an important axiom for the validity of cluster analysis[4] called *dimensional convergence* of D^2.

When a comparison between two real populations is made, one should ideally consider all possible (relevant) measurements, typically infinite in number. Consequently, cluster analysis of a given set of populations should ideally be based on distances computed on an infinite number of measurements. If D_p^2 and D_∞^2 denote Mahalanobis distances between two populations based on p characters and all possible characters respectively, then it can be shown under some conditions that

$$D_p^2 \rightarrow D_\infty^2 \text{ as } p \rightarrow \infty$$

(naturally after making that expression precise; see Rao[8]). Since in practice one can study only a finite number p of characteristics, D_p^2 should be a good approximation to D_∞^2 if cluster analysis is to be stable. Mahalanobis showed that stability, in important senses, was possible if and only if D_∞^2 is finite.

Meteorological research

Sir Gilbert Walker (then director general of observatories in India) referred to Mahalanobis some meteorological problems for statistical study. This resulted in two memoirs and a note on upper air variables[9]. Correcting meteorological data for errors of observation, Mahalanobis established by purely statistical methods that the region of highest control for changes in weather conditions on the surface on the earth is located about four kilometers above sea level (a result rediscovered later by Franz Bauer in Germany from physical considerations).

Early examples of operations research

In 1922 a disastrous flood occurred in North Bengal. An expert government committee of engineers was about to recommend the construction of expensive retarding basins to hold up the flood water, when the question was referred to Mahalanobis for examination. A statistical study of rainfall and floods extending over a period of fifty years showed that the proposed retarding basins would be of no value in controlling floods in North Bengal. The real need was improvement of rapid drainage and not holding up the flood water. Specific remedies were recommended, many of which were implemented and proved effective[10].

A similar question of flood control in Orissa was referred to Mahalanobis, after a severe flood of the Brahmini River in 1926. An expert committee of engineers were of the opinion that the bed of the Brahmini had risen, and they recommended increasing the height of river embankments by several feet. The statistical study covering a period of about sixty years showed that no change had occurred in the river bed and that construction of dams for holding excessive flood water in the upper reaches of the river would provide an effective flood control[11,12]. Mahalanobis pointed out that dams could be used for the generation of electric power needed for the economic development of the region. He also gave first calculations for a multipurpose (flood control, irrigation, and power) scheme for the Mahanadi system in Orissa, which formed the basis of the Hirakud hydro-electric project inaugurated about thirty years later in 1957.

Large-scale sample surveys

Large-scale sample survey techniques as practiced today owe much to the pioneering work of Mahalanobis in the 1940s and 1950s. He saw the need for sample surveys in collecting information, especially in developing countries, where official statistical systems are poor and data are treated as an integral part of the administrative system regulated by the principle of authority. A sample survey, properly conducted, would provide a wealth of data, useful for planning and policy purposes, expeditiously, economically, and with a reasonable degree of accuracy, and at the same time ensure objectivity of data.

The methodology of large-scale sample surveys was developed during 1937–1944 in connection with the numerous surveys planned and executed by the Indian Statistical Institute. The survey topics included consumer expenditure, tea-drinking habits, public opinion and public preferences, acreage under a crop, crop yields, velocity of circulation of rupee coins, and incidence of plant diseases. The basic results on large-scale sample surveys were published in 1944 and also presented at a meeting of the Royal Statistical Society[13] (1961 *a*).

The *Philosophical Transactions* memoir on sample surveys[14] is a classic in many ways, touching on fundamental problems: what is randomness? what constitutes a random sample? can different levels of randomness be identified? It gives the basic theory of sample surveys and estimation procedures.

In the 1944 paper, Mahalanobis[14] described a variety of designs now in common use, such as simple random sampling with or without replacement, *stratified*, *systematic*, and cluster sampling. He was also familiar with multistage and multiphase sampling, and with ratio and regression methods of estimation. He was, in a sense, conscious of selection with probability proportional to size (area); he pointed out that selection of fields on the basis of cumulative totals of the areas of millions of fields was difficult.

Mahalanobis made three notable contributions to sample survey techniques: pilot surveys, concepts of optimum survey design, and interpenetrating network of samples.

A pilot survey provides basic information on operational costs and the variability of characters, which are two important factors in designing an optimum survey. It gives an opportunity to test the suitability of certain schedules or questionnaires to be used in the survey. A pilot survey can also be used to construct a suitable frame for sampling of units.

From the beginning, Mahalanobis was clear about the principles of good sample design. He wrote[15] 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 and for given precision in the estimates, the minimum possible total expenditure....
>
> For this it is necessary to determine three things, viz. (a) what is the best size of the sample units?, (b) what is the total number of sample units which should be used to obtain the desired degree of accuracy in the final estimates?, and (c) what is the best way of distributing the sampling units among different districts, regions, or zones covered by the survey?

Mahalanobis constructed appropriate variance and cost functions in a variety of situations and used them in designing actual surveys.

As a physicist, Mahalanobis was aware of instrumental errors and personal bias in taking measurements, and consequently he stressed the need for repeating measurements with different instruments, different observers, and under different conditions. He maintained that a statistical survey was like a scientific experiment and that the planning of a survey required the same discipline and rigor as do other investigations. He advocated built-in cross-checks to validate survey results. For this purpose, he developed the concept of interpenetrating network of samples (i.p.n.s.).

A simple design using i.p.n.s. is as follows. Suppose that a given area is divided into four strata and that we have four investigators for the field work. The normal practice is to assign one stratum to each investigator to cover all units (randomly) chosen from that particular stratum. With i.p.n.s, however, each investigator works in all the strata and covers a random quarter of its units. Thus the i.p.n.s. design provides four independent (parallel) estimates of the characteristic under study for the region as a whole, corresponding to the four different investigators. The validity of the survey will be in doubt if the four estimates differ widely. In such a case, it may be possible, by further data analysis, to take the differences properly into account when reporting the final estimate. Such a camparison or critical study would not have been possible if the four different strata had been assigned to four different investigators, because stratum and investigator differences would have been confounded.

Fractile graphical analysis

Fractile graphical analysis is an important generalization of the method and use of concentration (or Lorenz) curves. A Lorenz curve for wealth in a population tells, for example, that the least wealthy 50 per cent of the population owns 10 per cent of the wealth. (If wealth were equally distributed, the Lorenz curve would be a straight line.) The comparison of Lorenz curves for two or more populations is a graphical way to compare their distributions of wealth, income, numbers of acres owned, frequency of use of library books, and so on.

One of Mahalanobis's contributions in this domain was to stress the extension of the Lorenz curve idea to two variables[16,17]. Thus one can consider, for example, both wealth and consumption for families, and draw a curve from which it may be read that the least wealthy 50 per cent of the families consume 27 per cent of total consumption, or a certain quantity per family on the average. Or, treating the variables in the other direction, one might find, perhaps, that the 20 per cent least-consuming families account for 15 per cent of the wealth or a certain value per family. (The numbers in these examples are hypothetical and only for illustration.) Such bivariate generalized Lorenz curves can, of course, also be usefully compared across populations.

Economic planning

For Mahalanobis, statistics—and science generally—were meaningful only as they helped in understanding the problems of the real world and particularly the problems of poverty in India. This led him to develop and devour statistics relating to the Indian economy with a passion displayed by few economists. He never studied traditional economic theory, and he believed that his lack of formal education in economics enabled him to work on economic problems of the country with a sense of realism. He said[18], 'The sophisticated economic theories which may be appropriate for advanced countries acted for a long time as thought barriers to economic progress in India.'

Soon after independence, in 1949, Mahalanobis was appointed by Jawaharlal Nehru as honorary statistical advisor to the cabinet—a post that he held until his death. He believed that, before any worthwhile thinking about the economic problems of India could be done, it was necessary to have the basic national income statistics. On his advice, the government of India appointed in 1949 the National Income Committee 'to prepare a report on the national income and related estimates, to suggest measures for improving the quality of available data and for the collection of further essential statistics and to recommend ways and means of promoting research in the field of national income.' Mahalanobis was appointed chairman of this committee. The committee's reports laid the foundations of systematic and continuous work on national income in India.

The first major task was development of a national statistical system for collection and tabulation of official data at the state and federal levels. To this end, Mahalanobis advised the government of India to set up the central statistical unit that became the Central Statistical Organization (CSO). Simultaneously, plans were made to establish statistical bureaus in all the states to ensure efficient collection and reporting of data at the state level. But the work of the CSO and the state bureaus needed to be supplemented by extensive field surveys covering the entire country. For this, Mahalanobis recommended the establishment of the independent National Sample Survey (NSS) organization to operate with its own field staff and technical management.

Beginning in 1950, the NSS started turning out vast amounts of data relating to many important but dark areas of the Indian economy, for example, consumer expenditure, distribution and land holdings, unemployment, economics of cultivation, vital statistics, household enterprises, and amenities like postal, educational, and health services available in rural areas. Soon Mahalanobis was called upon by Jawaharlal Nehru to help in government planning. Setting himself the twin objectives of doubling the national income and reducing unemployment considerably over a period of twenty years, he produced what are known as two- and four-sector models for economic development with investment in various sectors of the economy as the centre piece. In his 1950 presidential address 'Why Statistics?', delivered to the Indian Science Congress at Poona, Mahalanobis first expounded his ideas about using capital-output ratio to derive estimates of the level of investment required for a stipulated increase in income. Subsequently he developed independently a forward-looking Harrod-Domar type of model for economic growth[19-21].

Mahalanobis's association with the Planning Commission, and his involvement in the formulation of its Second Five-Year Plan brought him into contact with problems of wide national and international importance. He wrote extensively on such subjects as (1) the priority of basic industries, (2) the role of scientific research, technical manpower, and education in economic development, (3) the industrialization of poorer countries and world peace[22], and (4) labour and unemployment.

Honors and awards

Mahalanobis was elected as a fellow of the Royal Society of London in 1945, of the Econometric Society, USA, in 1951, and of the Pakistan Statistical Association in 1961; as an honorary fellow of the Royal Statistical Society, UK, in 1954 and and of King's College, Cambridge, in 1959; as honorary president of the International Statistical Institute in 1957; and as a foreign member of the USSR Academy of Science in 1958. He received the Weldon medal from Oxford University in 1944, a gold medal from the Czechoslovak Academy of Sciences in 1963, the Sir Deviprasad Sarvadhikari gold medal in 1957, the Durgaprasad Khaitan gold medal in 1961, and the Srinivasa Ramanujam gold medal in 1968. He received honorary doctorates from Calcutta, Delhi, Stockholm, and Sofia universities and one of the highest civilian awards, Padmavibhushan, from the government of India.

Life with a mission

The 79 years of Mahalanobis's life were full of activity. His contributions were massive on the academic side as the builder of the Indian Statistical Institute, founder and editor of *Sankhyā*, organizer of the Indian Statistical Systems, pioneer in the applications of statistical techniques to practical problems, promoter of the statistical quality control movement for improvement of industrial products, architect of the Indian Second Five-Year Plan, and so on.

Statistical science was a virgin field and practically unknown in India before the 1920s. Developing statistics was like exploring a new territory; it needed a pioneer and adventurer like Mahalanobis, with his indomitable courage and tenacity to fight all opposition, clear all obstacles, and open new paths of knowledge for the advancement of science and society. (See Rao[23].)

The Mahalanobis era in statistics, which started in the early 1920s, ended with his death in 1972. It will be remembered as a golden period of statistics, marked by an intensive development of a new (key) technology[24], and its application for the welfare of mankind.

1. Mahalanobis, P. C., *Sci. Cult.*, 1960, **27**, 101–128.
2. Mahalanobis, P. C., *Rec. Indian Mus.*, 1922, **23**, 1–96.
3. Mahalanobis, P. C., *J. Proc. Asiatic Soc. Bengal, New Series*, 1930, **26**, 541–588.
4. Mahalanobis, P. C., *Natl. Inst. Sci. India, Proc.*, 1936, **2**, 49–55.
5. Mahalanobis, P. C., *Sankhyā*, 1940b. **4**, 594–598.
6. Mahalanobis, P. C., Majumdar, D. N. and Rao, C. Radhakrishna, *Sankhyā*, 1949, **9**, 89–324.
7. Rao, C. Radhakrishna, *Advanced Statistical Methods in Biometric Research*, Hafner, New York; Reprinted with corrections in 1970 and 1974.
8. Rao, C. Radhakrishna, *Int. Stat. Inst. Bull.*, 1954, **34**, 90–97.
9. Mahalanobis, P. C., *Nature*, 1923, **112**, 323–324.
10. Mahalanobis, P. C., Report on Rainfall and Floods in North Bengal, 1870–1922, 2 vols. Submitted to the government of Bengal, vol. 1, text. vol. 2, 28 maps.
11. Mahalanobis, P. C., Statistical Study of the Level of the Rivers of Orissa and the Rainfall in the Catchment Areas During the Period 1868–1928. Report submitted to the government of Bihar and Orissa.
12. Mahalanobis, P. C., *Sankhyā*, 1940c, **5**, 1–20.
13. Mahalanobis, P. C., *Experiments in Statistical Sampling in the Indian Statistical Institute*, Asia Publ. House and Statistical Publ. Society, Calcutta; also in *J. R. Stat. Soc.*, 1961a, **A109**, 325–378.
14. Mahalanobis, P. C., *Philos. Trans.*, 1944, **B231**, 329–451.
15. Mahalanobis, P. C., *Sankhya*, 1940a, **4**, 511–530.
16. Mahalanobis, P. C., *Bose Res. Inst., Trans.*, 1958b, **22**, 223–230.
17. Mahalanobis, P. C., *Econometrica*, 1958c, **28**, 325–351; also in *Sankhya*, 1961, **A23**, 41–64.
18. Mahalanobis, P. C., *Operations Res. Soc. Jpn J.*, 1961b, **3**, 98–112
19. Mahalanobis, P. C., *Sankhyā*, 1953, **12**, 307–312

20 Mahalanobis, P. C., *Sankhyā*, 1955a, **16**, 63-90.
21. Mahalanobis, P. C., *Sankhyā*, 1955b, **16**, 3-62.
22. Mahalanobis, P. C., *Bull. Atomic Sci.*, 1958a, **15**, 12 17, also in *Sankhyā*, 1960, **22**, 173-182.
23. Rao, C. Radhakrishna, *Biogr. Mem. Fellows*, 1973, **19**, 455-492.
24. Mahalanobis, P. C., *Am. Statistician*, 1965, **19**, 43-46.

C. R. Rao is in Statistics Department, Pennsylvania State University, University Park, 417 Classroom Bldg., Pennsylvania, 16802, USA.

My friendship with Ramanujan in England

P. C. Mahalanobis

Prof. P. C. Mahalanobis, the eminent statistician, recalls the days of his friendship with Ramanujan in England.

Berry's class

I joined King's College, Cambridge, in October 1913. I was attending some mathematical courses at that time including one by Professor Hardy. A little later, we heard that S. Ramanujan, the mathematical prodigy, would come to Cambridge. I used to do my tutorial work with Mr Arthur Berry, Tutor in Mathematics of King's College. One day I was waiting in his room for my tutorial when he came in after having taken a class in elliptic integrals. He asked me: 'Have you met your wonderful countryman, Ramanujan?'. I told him that I had heard that he had arrived but I had not met him so far. Mr Berry said: 'He came to my elliptic integrals class this morning.' (This was some time after the full term had begun, and I knew Mr Berry had already given a few lectures on that subject.) I asked, 'What happened? Did he follow your lecture?'. Mr Berry said, 'I was working out some formulae on the black board. I was looking at Ramanujan from time to time to see whether he was following what I was doing. At one stage, Ramanujan's face was beaming and he appeared to be excited. I asked him whether he was following the lecture and Ramanujan nodded his head. I then enquired whether he would like to say anything. He then got up from his seat, went to the black board and wrote some of the results which I had not yet proved'. I remember Mr Berry was greatly impressed. He said that Ramanujan must have reached those results by pure intuition as Professor Hardy had advised Ramanujan to attend the lectures on elliptic integrals because Ramanujan had not studied that subject before.

I was fortunate in becoming good friends with Ramanujan very soon. It came about in a somewhat strange way. Within a few days of his arrival I had managed to get acquainted with him and was meeting him from time to time. One day I went to see Ramanujan in his room in Trinity College. It had turned quite cold. Ramanujan was sitting very near the fire. I asked him whether he was quite warm at night. He said he was feeling the cold; he was sleeping with his overcoat on and was also wrapping himself up in a shawl. I went to his bedroom to see whether he had enough blankets. I found that his bed had a number of blankets but all tucked in tightly, with a bed cover spread over them. He did not know that he should turn back the blankets and get into the bed. The bed cover was loose; he was sleeping under it, with his overcoat and shawl. I showed him how to get under the blankets. He was extremely touched. I believe this was the reason why he was so kind to me.

In my second year, I got rooms in King's College; one term I had rooms in the staircase overlooking Queen's (just below where J. M. Keynes, at that time, a brilliant young fellow of King's, had his rooms). On Sunday mornings Ramanujan and I often went out for long walks. One Sunday it had been arranged that we would both have our breakfast in my room and then go out for a walk. It was a cold morning with some snowfall. I was a bit late in getting up and was shaving in my bedroom when he arrived. I asked him to wait in the sitting room. When I came out I found that he was reading Loney's *Dynamics of a Particle* with great interest. Seeing me, he put back the book on the table and said it was very interesting. Evidently he had never studied dynamics but had got interested in what he was reading.

A problem from *Strand* Magazine

On another occasion, I went to his room to have lunch with him. The First World War had started some time ago. I had in my hand a copy of the monthly *Strand* Magazine which at that time used to publish a number of puzzles to be solved by the readers. Ramanujan was stirring something in a pan over the fire for our lunch. I was sitting near the table, turning over the pages of the *Strand* Magazine. I got interested in a problem involving a relation between two numbers. I have forgotten the details but I remember the type of the problem. Two British officers had been billeted in Paris in two different houses in a long street; the two numbers of these houses were related in a special way; the problem was to find out the two numbers. It was not at all difficult; I got the solution in a few minutes by trial and error. In a joking way, I told Ramanujan, 'Now here is a problem for you'. He said, 'What problem, tell me', and went on stirring the pan. I read out the question from the *Strand* Magazine. He promptly answered 'Please take down the solution' and dictated a continued fraction. The first term was the solution which I had obtained. Each successive term represented successive solutions for the same type of relation between two numbers, as the number of houses in the street would increase indefinitely. I was amazed and I asked him how he got the solution in a flash. He said, 'Immediately I heard the problem it was clear that the solution should obviously be a continued fraction; I then thought, which continued

Reprinted from *Sankhyā*.