Indian J. pure appl. Math., 26(6) : 485-501, June 1995

THE EVOLUTION OF THE D^2-STATISTIC OF MAHALANOBIS

SOMESH DASGUPTA

Indian Statistical Institute, Calcutta 700 035

The early works of Mahalanobis leading to the evolution of his celebrated D^2-statistic have been reviewed. A variety of statistical methods and some interesting conclusions related to the concrete problems that he considered in the process of the development of D^2-statistic have been presented. These classical papers are not easily available and the results are hitherto unknown.

1. THE BEGINNING

1.1. The early work of Mahalanobis in Statistics, and the development of the D^2-statistic, in particular, may be accounted to four sources of motivation, namely (i) the papers in Anthropometry published in the early volumes of Biometrika, (ii) the influence of Professor Brajendra Nath Seal, (iii) the data on the Anglo-Indians of Calcutta obtained through Mr. Annandale, and (iv) his own curiosity on the race origins and race mixture, especially with regard to the inhabitants of Bengal.

1.2. The publication of 'The Origin of Species' by Charles Darwin in 1859 amidst the ongoing industrial revolution in England gave rise, in particular, to the eugenics movement on one hand and the intellectual curiosity on the theory of evolution and natural selection on the other. The theory of evolution as espoused by Darwin embodies the existence of the so called racial differences. This particular theme was taken up in a series of papers published in the early volumes of the journal 'Biometrika', founded by Weldon, Karl Pearson and Galton. To start with, the relevant papers published before 1930 are worth mentioning; these are listed in References I.

These early papers provided statistical methods in order to analyse sets of anthropological data, and specifically to assess similarity or dissimilarity between two populations.

1.3. After his Tripos examination in 1915, Mahalanobis met his tutor W. H. Macaulay in the college library, who showed him some new bound volumes of Biometrika. Mahalanobis got so interested in these volumes that he purchased a set of Biometrika and brought it to India. The papers in Biometrika dealing with

Reprinted from *Sankhyā : The Indian Journal of Statistics* 1993, Special Volume 55, Series A, Pt. 3, pp. 442-459 dedicated to the memory of P. C. Mahalanobis with the permission of Indian Statistical Institute and the author.

biological and anthropological data had an immediate influence on Mahalanobis. During 1920-21, he wrote papers on anthropometric constants for Bengal caste data, a note on the criterion that two samples come from the same population, and another note on statistical constants for an Anglo-Indian sample. Besides being actively interested in statistical problems relating to agriculture, meteorology, and education, Mahalanobis was also deeply involved in the questions relating to racial mixture, racial origins and the assessment of group differences. During 1922-1936, he wrote 15 papers on these subjects leading to his celebrated paper on the generalized distance in 1936.

1.4. Mahalanobis' interest in the study of group divergence was also influenced by Professor Brajendra Nath Seal, who first stated the concept of group divergence in his address on "Race Origin" delivered before the Universal Races Congress in London in 1911. Professor Seal had stated, "If the groups requiring to be arranged vary in n characters, and if biometric measurements are complete, then composite mean of the groups may be taken as the point of origin, and the mean of the single characters for each group may be imagined as marked off on "n" coordinates, and the position in n-dimensional space of each group could be easily assigned". Professor Seal could visualize the future possibilities of statistics as a basic scientific discipline in India, and it was largely due to his encouragement that Mahalanobis continued statistical studies despite being trained in physics and applied mathematics.

1.5. In 1920, Mahalanobis met Dr. N. Annandale, then Director of the Zoological and Anthropological Survey of India. Dr. Annandale had taken anthropological measurements such as stature, head-length, head-breadth, nasal length, etc. of 300 Anglo-Indians in Calcutta. He gave the data-set, after omitting the doubtful and incomplete records, to Mahalanobis for statistical analysis. This was the first time that data relating to a true biologically mixed population were studied by statistical methods. Mahalanobis was able to develop deeper insight and better grasp in statistical methods through the analysis of this set of data.

1.6. Moreover, the intellectual fervour of the post-renaissance Bengal and the influence of the spirit of the Brahmo-samaj led Mahalanobis to inquire, in particular, into the nature of the Indian components among the Anglo-Indians, and later, on the formation and development of the different caste-groups in Bengal.

2. The Contents of this Article

The key papers of Mahalanobis leading to the formulation of the D^2-statistic will be briefly reviewed in this article. After 1936, Mahalanobis was deeply involved in concrete statistical problems relating to the socio-economic conditions in India and engaged in various sample surveys. He returned to the discriminatory problem in 1949 and wrote a joint paper with D. N. Majumder and C. R. Rao; this paper will also be reviewed in this article, since it reveals some of the important ideas of Mahalanobis.

The major focus will be given to the various other measures and the related statistical methods which he considered for the problem of discrimination. In particular, the major conclusions of Mahalanobis related to the concrete problems that

he considered in the process of the development of the D^2-statistic will also be presented. Lastly, all relevant references to Mahalanobis will be listed. These classical papers are not easily accessible, and they contain a variety of statistical methods and some interesting conclusions, hitherto unknown.

3. The Anglo-Indian Data

Mahalanobis wrote three papers on the Anglo-Indian (AI) data; it is this set of data with which he explored his different ideas of distance measures.

In his first paper (1922), Mahalanobis inquired whether the Anglo-Indians were more homogeneous or less homogeneous in comparison to the variability in other races, so far as stature is concerned. Let us denote the stature of an individual by X, the mean of X by $\bar{X}$, and the s.d. of X by s. Mahalanobis compared the Anglo-Indians with other caste groups by the following measures.

$$\frac{\bar{X}_{AI} - \bar{X}_{others}}{s.d.\ of\ group\ \bar{X}'s} \quad \text{and} \quad \frac{s_{AI} - s_{others}}{s.d.\ of\ group\ s.d.'s}$$

Furthermore, to find out whether the variability in stature is related to the average stature, he computed the correlation coefficient between $\bar{X}$ and s, and the correlation coefficient between $\bar{X}$ and the coefficient of variation. His findings led him to make the following conclusions :

- Inter-racially, taller races are more variable than shorter ones.
- The variability of stature among the Anglo-Indians is significantly greater than the corresponding variability in other castes, but it is not beyond the range of homogeneous variability.

  Mahalanobis also considered the Anglo-Indians in different age groups, and computed the correlation between age and stature. This led him to draw the following conclusion.

- The Anglo-Indians seem to be rather precocious in growth, and there is some indication of the arrest of growth occurring at an early age than in the case of European races.

  By comparing different age groups, it was revealed that

- the variability in smaller age groups is distinctly less showing a decrease of variability in time.

  This led him to conclude that

- Anglo-Indians of the younger generation are more homogeneous in their stature.

  The second paper (1931) on the Anglo-Indians dealt with the data on head-length, and an analysis similar to that in the first paper was carried out. His main conclusions were the following :

- Anglo-Indian variability in head-length as judged by s.d. or c.v. is definitely and significantly greater than that in other caste groups.

- The variability of head-length among the Anglo-Indians indicates recent inter-mixture.
- In spite of excessive variability, the variability of head-length in different age groups among the Anglo-Indians does not indicate a jump in the amount of variability.

It may be noted that the above work of Mahalanobis was clearly influenced by a proper by E. Tscherpourkowsky on the study of inter-racial correlation, published in Biometrika (1905).

In his third paper (1940), Mahalanobis considered seven anthropological measurements on the Anglo-Indians in Calcutta. Firstly, he deduced (by curve-fitting) that the marginal distribution of each of these measurements was approximately normal. Secondly, he found that all pairwise regressions were approximately linear. He was then led to conclude that the joint distribution of these measurements is approximately multivariate normal.

4. Analysis of Race Mixture in Bengal

In his 33-page long classic paper (1925), Mahalanobis made an attempt to get more insight into the nature of caste similarity or dissimilarity. Firstly, he posed the following questions :

"How are these 200 Anglo-Indians in Calcutta related to the different caste-groups of Bengal? Are they more closely allied with the Hindus or with the Mohammedans ? Do they show a greater affinity with the higher castes of Bengal or with the lower castes ?....".

To get some idea about the possible composition of the given sample of the AI's in terms of broader social and geographical divisions of the inhabitants of Bengal and its neighbourhood, Mahalanobis considered Risley's data (1891) which give anthropological measurements of individuals belonging to 30 typical castes of Northern India. These data represent about six geographical divisions (Bengal, Chotanagpur tribes, Bihar, New provinces and Oudh, Punjab, and Lepcha, Chakma and Magh in the Eastern districts), and four or five cultural strata (high castes, low castes, aboriginal tribes, Eastern tribes, Mohammedans).

The analysis presented in this paper was based on fifteen measurements of which ten are absolute anthropological measurements and five are anthropological indices.

Mahalanobis introduced a measure D of caste-distance as follows:

$$D = \frac{1}{p} \sum_{i=1}^{p} \frac{(m_i - m_i')^2}{s_i^2},$$

where m_i and m_i' are the means of the ith measurement in the two groups, respectively, and s_i^2 is the pooled variance of the ith measurement, the number of measurements being p.

However, Mahalanobis did consider the coefficient of racial likeness, introduced by Karl Pearson, which is given as follows :

$$C = \frac{1}{p} \sum_{i=1}^{p} \frac{nn'}{n+n'} \frac{(m_i - m_i')^2}{s_i^2} - 1,$$

where n and n' are the sample sizes in the two groups. But he emphatically made it a point that the coefficient C is influenced by sample sizes, and fails to measure the degree of divergence between the two groups. Later, Mahalanobis modified his measure D and suggested the following measure given by

$$D' = D - \frac{n+n'}{nn'},$$

and presented its (asymptotic) variance as

$$\sigma_{D'}^2 = \frac{4}{p}\left(\frac{n+n'}{nn'}\right)\overline{D} + \frac{2}{p}\left(\frac{n+n'}{nn'}\right)^2,$$

where $\overline{D}$ is the mean value of D'.

Most importantly, Mahalanobis introduced a concept called 'positional distance' to measure the relative position of a particular group in relation to a set of culturally similar or geographically close groups.

For example, to understand "geographical resemblance", one may consider the positional indices of Bengal Brahmins with respect to Bengal, Bihar, N.W.P., Punjab, etc. Similarly, the effect of "cultural affinity" may be studied by considering for example, the positional indices of Brahmins for high castes of Bengal, Bihar and Punjab, for low castes, for aboriginal tribes of Chotanagpur, etc.

Let us now define the positional index following Mahalanobis.

Consider a group G in relation to a given list of n groups. Compute the distance D of the group G from each of the groups in the list, and rank these measures in accordance to their values. Let C be a collection of m groups selected from the given list excluding G, and r be the average of these rank-values of the groups in C. Then the positional index of G for C is defined to be

$$P = \frac{n+1-2r}{n-m} \times 100$$

It is clear that P varies from +100 to −100, and a high value of P indicates that the group G is relatively closer to the given collection C of groups.

Let us quote the summary of his analysis for Bengal castes.

"Summing up we find that intermixture within Bengal, i.e. intra-provincial intermixture has varied with the degree of cultural proximity, so that for Brahmins the amount of intermixture with other castes has been in proportion to the social standing of the caste concerned. Influence from outside Bengal, i.e., inter-provincial intermixture has followed two well-defined and clearly distinguished streams, one from the castes of Northern India (chiefly from Bihar and the Punjab) and the other from the aboriginal tribes of Chotanagpur. The influence of the Northern Indian castes decreases and that of the aboriginal tribes of Chotanagpur increases as we go down

to the social scale ... . None of the castes analysed here show much resemblance with any of the aboriginal tribes of the east ... . Mohammedans (also) show a highly mixed character. They appear to be originally largely derived from Bihar but have intermixed extensively in Bengal; they do not show any resemblance with the Punjab Pathans."

Referring back to his original question on the Anglo-Indians of Calcutta, Mahalanobis gave the following answers :

"The Anglo-Indians included in the present sample are derived (on the Indian side) mainly from the Bengal castes. They show a certain amount of admixture with Bihar and also possibly with the Punjab, but not with N.W.P. They are singularly free from contact with the Chotanagpur tribes, but appear to have intermixed to some extent with the Lepchas of Darjeeling. So far as the present analysis goes, we also see that intermixture between Europeans and Indians in Bengal appears to have occurred more frequently among the higher castes than among the lower. Evidently cultural status played a considerable part in determining Indo-European Union". The most striking conclusion of his analysis is given in the general summary.

"If we assume that physical resemblance is the result of actual intermixture, and that also more or less in quantitative proportion, then we may give a coherent interpretation to our results and thus obtain a broad view of the general tendency of social history in Bengal.

We find that movements of caste-synthesis are proceeding on every side under our very eyes. Social barriers and caste restrictions have not been able to suppress it completely. The peoples from the north west have fused with the indigenous stock in Bengal and the aboriginal tribes of Chotanagpur have intermingled with them. Intermixture within the province has gone on slowly and steadily even if imperceptibly and a larger Hindu Samaj has evolved which is not only identical with the traditional society of Vedic or Classic times but is in many respects even antagonistic. Sectarian obstacles have not proved insurmountable; the Mohammedans who came originally as immigrants have contributed their share and have received back their own contributions from the other castes. The process has not stopped here; it has gone on even after the advent of the Westerners with their totally different culture, history and tradition.

Yet equally striking is the fact that intermingling has not been altogether chaotic. It presents a gradual and well-ordered character in which cultural affinity and cultural selection has played an important part ... . The Hindu community of Bengal does not on one hand conform to the orthodox scheme of a logically perfect system of rigidly exclusive castes between which no intercourse is ever possible, on the other hand neither does it present an amorphous or chaotic character."

Although Mahalanobis tried to devise suitable statistical methods in order to analyse some given sets of anthropometric data, he did not accept any such set of data without proper scrutiny of field records and possible measurement errors. He used these data only after appropriate corrections. In particular, he examined Risley's data quite rigorously and reconstructed the set (1933, 1934).

5. Analysis of Chinese Data and Swedish Data

In the process of developing his ideas on measure of group divergence, Mahalanobis considered, in particular, two sets of data :

(a) Anthropological measurements of different parts of head for Chinese samples.

(b) Anthropological measurements for Swedish samples.

The data set (a) was taken by S. M. Shirokogroff in 1908-12 and 1923-24 and later published by the North China Branch of the Royal Asiatic Society in the form of two reports. The data set (b) was taken from "the Racial characters of the Swedish Nation" edited by H. Lindeberg and F. J. Linders, and published by the Swedish State Institute for Race Biology, Uppsala, in 1926.

In his paper entitled "A statistical study of the Chinese head" (1928), Mahalanobis tried to ascertain the degree of similarity or dissimilarity among the various provincial groups. Although he computed Pearson's CRL for different pairs of groups, he used his measure D^2 (given in Section 4 as D) to form a smaller number of groups or clusters. His conclusions are summarized as follows :

"We thus see that all the Chinese groups from the northern provinces, e.g. the Chinese of Manchuria, Northern Chinese, and Eastern Chinese are closely associated with one another, and all show fairly close resemblance with both Manchus and Koreans who also come from the North. The Southern Chinese on the other hand are clearly differentiated from practically all the northern groups, with the single exception of the Chinese from the eastern provinces (which are adjacent to Kwangtung) with whom they show fairly close association.

Koreans and Manchus, although both show appreciable resemblances with all the Chinese groups from the North, are distinctly differentiated from each other".

He described his findings by a simple relationship between geographical proximity and physical resemblance.

In this paper, Mahalanobis tried to ascertain the role of each measurement separately for discriminating different provincial groups. He proposed, for this purpose, to compute the ratio of inter-class variance to intra-class variance. His findings indicate that nasal breadth, minimum frontal diameter, height of head, internal ocular breadth are constant or 'family characteristics', ... while the variation within the family is strongly marked in external ocular breadth, ear-length, morphological face length, bigonial diameter, physiological face length and head length.

In his statistical study of certain anthropometric measurements from Sweden (1930), Mahalanobis tried to compare the various Swede groups by Pearson's method of CRL instead of comparing the groups with respect to each of the characters separately. The sample consists of 46, 983 individuals, classified into five regional groups, and individuals in each territory were classified again into four occupational groups.

Since Pearson's CRL depends on sample sizes, Mahalanobis modified these coefficients assuming that the means are based on 100 individuals (or same number of individuals); the s.d. of CRL was modified accordingly.

Based on his findings, Mahalanobis represented the various groups in a diagram, given here as Fig. 1. Besides these, Mahalanobis tried to grade the different characters by their role for measuring the "significance of the differences". He used the ratio of inter-class variance to intra-class variance for this purpose. He wanted to diagnose the characters which show markedly significant territorial differences and characters which show significant occupational sequence within the territories.

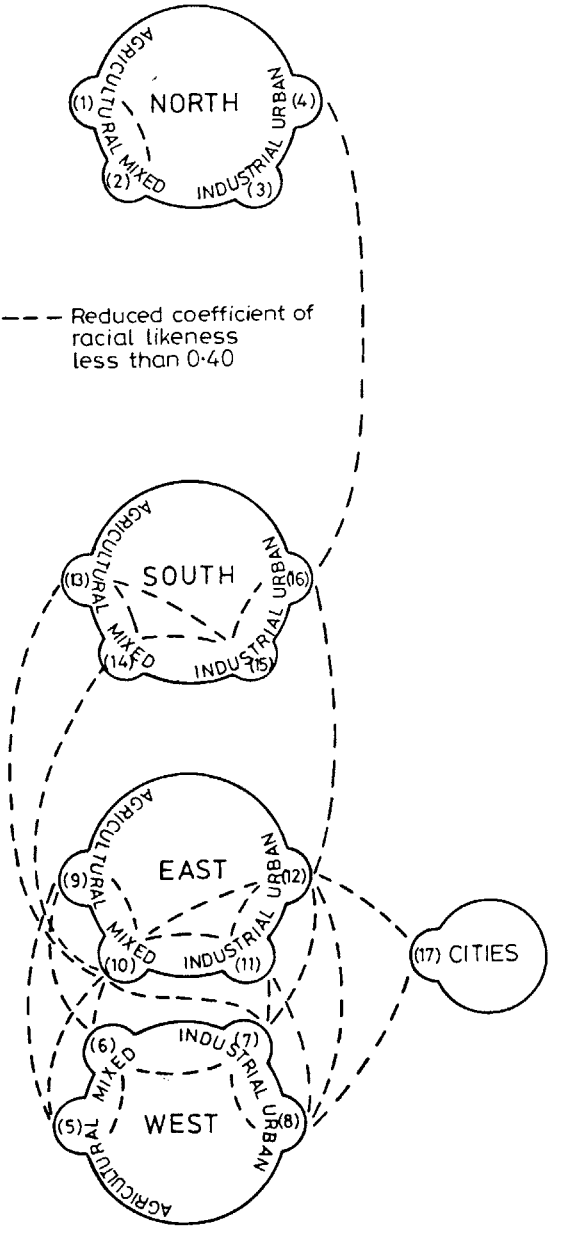

FIG. 1. Inter-relationship of various groups of the population of Sweden.

6. The First Theoretical Paper

The analysis of different sets of anthropological data by Mahalanobis gradually led him to develop some theoretical concepts. His first theoretical paper on test and measures of group divergence (1930) was 48 pages long, and presents an enormous bulk of theoretical calculations.

Firstly, Mahalanobis formulated measures of divergence not only with respect to means, but also with respect to variability, skewness, and kurtosis.

Secondly, he made an attempt to formulate a measure of divergence from axiomatic viewpoint. He proposed the following general measure to assess divergence in terms of means :

$$U^2 = f\left(\sum_{i=1}^{p} \frac{(m_{iq} - m_{iq'})^2}{k_i^2} \right)$$

where m_{iq} and $m_{iq'}$ are the means of the ith character in the qth and q'th groups, respectively, and k_i, is a suitable multiplier having the same dimension as in m_{iq} or $m_{iq'}$ p being the total number of characters. In particular, he suggested that k_i could be taken as the intra-class s.d. or the inter-class s.d. or the familial s.d.. Next, he evaluated the approximations to the first four moments of

$$\frac{1}{p} \sum_{i=1}^{p} \frac{(m_{iq} - m_{iq'})^2}{k_i^2}$$

for large sample sizes, and specialized his results for

$$D^2 = \frac{1}{p} \sum_{i=1}^{p} \frac{(m_{iq} - m_{iq'})^2}{\overline{\sigma}_i^2} - \frac{1}{p} \sum \left(\frac{1}{n_{iq}} + \frac{1}{n_{iq'}} \right),$$

where n_{iq} is the number of observations on the ith character in the qth group, and $\overline{\sigma}_i^2$ is a reliable constant value for the variance of the ith character. By assuming that $n_{iq} = n_q$ for all i, and defining

$$\frac{2}{\overline{n}} = \frac{1}{n_q} + \frac{1}{n_{q'}} ,$$

he obtained the following results for large sample sizes :

$$E(D^2) = \overline{D}^2 = \frac{1}{p} \sum_{i=1}^{p} \frac{(\mu_{iq} - \mu_{iq'})^2}{\overline{\sigma}_i^2}$$

$$\mu_2 (D^2) \sim \frac{8(\delta + 1)}{p \overline{n}^2}$$

$$\beta_1(D^2) \sim \frac{2}{p}\frac{(3\delta+2)^2}{(\delta+1)^3},$$

$$\beta_2(D^2) \sim 3+\frac{12}{p}\frac{2\delta+1}{(\delta+1)^2},$$

where $\delta = \bar{n} D^2$, and μ_{iq} is the population mean of the ith character in the qth group.

Based on the values of β_1 and β_2, Mahalanobis suggested to use a Pearson's Type III curve for approximating the distribution of D^2 when δ is small in comparison to 1; when δ is significantly different from 0, he suggested to use a Pearson's type I curve. He also carried out model sampling experiments (i.e., simulation), and found that the above approximations were fairly satisfactory in most of the cases. In this study, the characters were assumed to be independent normal variates. He made similar calculations with other choices of k_i.

Mahalanobis proposed to use f_i^2, the ratio of the inter-group variance to the average intra-group variance for the ith character, for measuring the amount of differentiation existing within a given collection of groups with respect to the ith character. He defined a coefficient f^2 of familial differentiation as

$$f^2 = \frac{c(k)}{p}\sum_{i=1}^{p} f_i^2$$

where $c(k)$ is a suitable numerical constant depending on k, the number of groups under consideration. It has been shown that f^2 is a linear function of the average value of all pairwise measures of divergence. Furthermore, Mahalanobis derived the approximate values of the first four moments of f^2 and showed that the distribution of f^2 is approximately normal for large sample sizes, large k, and when the k populations are the same.

In this paper, Mahalanobis cited a number of comparisons for which Pearson's CRL C^2 and his D^2-measures gave widely different results. Mahalanobis claimed that, in critical cases, values of D^2 were more in accordance with the known anthropological facts. He noted that "the magnitude of C^2 determines the degree of certainty with which the *existence* of divergence can be asserted, but does not supply any information regarding the magnitude of such divergence".

The draft of this 1930 paper was shown to Karl Pearson; but Pearson did not accept the views of Mahalanobis on the relative merits of C^2 and D^2.

7. The Generalized Distance

In his 1930 paper, Mahalanobis noted that "the most important of the restrictions which requires further considerations is the neglecting of the correlation between

different characters". At least, in 1936, he was able to formulate a distance measure incorporating the correlations between different characters. His frame of reference for this development was of course the multivariate normal distribution. The square of the generalized distance measure between two multivariate normal populations $N_p(\mu_1, \Sigma)$ and $N_p(\mu_2, \Sigma)$ was defined by Mahalanobis as follows :

$$\Delta^2 = \frac{1}{p}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2).$$

Mahalanobis noted that the above formula is the analog of ds^2 in the restricted theory of relativity.

The sample analog of Δ^2 when Σ is known was given by

$$D_1^2 = \frac{1}{p}(\bar{X}_1 - \bar{X}_2)' \Sigma^{-1}(\bar{X}_1 - \bar{X}_2),$$

where $\bar{X}_1$ and $\bar{X}_2$ are the mean vectors obtained from random samples of sizes n_1 and n_2, respectively, from the above two normal populations. Later, he modified the above D_1^2 as follows in order to estimate Δ^2 unbiasedly :

$$D^2 = \frac{1}{p}(\bar{X}_1 - \bar{X}_2)' \underset{\sim}{\Sigma}^{-1}(\bar{X}_1 - \bar{X}_2) - \frac{2}{\bar{n}},$$

where

$$\frac{2}{\bar{n}} = \frac{1}{n_1} + \frac{1}{n_2}.$$

Mahalanobis derived the first four moments of D^2, which are given below :

$$E(D^2) = \Delta^2, \quad \mu_2(D^2) = \frac{8}{p\bar{n}}\left(\Delta^2 + \frac{1}{\bar{n}}\right)$$

$$\mu_3(D^2) = \frac{32}{p\bar{n}^2}\left[3\Delta^2 + \frac{1}{\bar{n}}\right], \quad \mu_4(D^2) = \frac{192}{p\bar{n}^2}\left[\left(\Delta^2 + \frac{2}{\bar{n}}\right)^2 + \frac{4}{p\bar{n}}\left(2\Delta^2 + \frac{1}{\bar{n}}\right)\right].$$

When both the mean vectors and the dispersion matrix are unknown, the sample version of Δ^2 was suggested as follows :

$$D_2^2 = \frac{1}{p}(\bar{X}_1 - \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2),$$

where S is the pooled estimate of Σ.

When Σ is diagonal, the pooled estimate of Σ is also a diagonal matrix, and correspondingly D_2^2 is modified; R. C. Bose derived the first four moments of D_2^2 in this case, and the results are given in the appendix of this 1936 paper.

Later, the exact distribution of D^2 and its moments were obtained by R. C. Bose (1936). The moments of D^2 were also obtained by S. N. Bose (1936, 1937) without using its distribution explicitly.

The distribution of the studentized D^2-statistic, given by D_2^2, was obtained by R. C. Bose and S. N. Roy (1938); their objective was to verify (and probably provide an alternative derivation) the corresponding result obtained earlier by Hotelling (1931) and Fisher (1936); incidentally, it may be noted that Fisher's derivation was not correct.

One year later, in a joint paper (1937) with R. C. Bose and S. N. Roy, Mahalanobis introduced a concept of dimensional convergence. Mahalanobis stated that, if D_p^2 denotes the generalized distance based on p characters between two groups, then the sequence D_p^2, D_{p+1}^2, D_{p+2}^2, ... tends to a definite limit in whatever order the characters may be chosen. Methods for judging the significance of an additional character have been discussed later by Rao (1948).

8. Last Paper on D^2 by Mahalanobis

The first phase of the work of Mahalanobis on group divergence ended in 1936, and he earned international reputation for his generalized distance measure. From the late 1930's, Mahalanobis shifted his interests to the development and application of statistical methods in relation to various socio-economic issues in India. In particular, he was deeply involved in various sample surveys seeking information on agricultural production, and socio-economic condition in India. Research on group divergence and the generalized distance measure, in particular, was continued by some of his colleagues and students in the Indian Statistical Institute; the works of R. C. Bose, S. N. Roy, and C. R. Rao, in particular, led to significant development.

Later in 1949, Mahalanobis participated in the programme for statistical study of anthropometric survey in the U.P., in collaboration with D. N. Majumdar and C. R. Rao. Their development and findings came out in *Sankhya* (1949). This set of anthropometric data was collected in the field survey by D. N. Majumdar in connection with the population census in India in 1941. Anthropometric measurements of about 12 characters for 2836 individuals belonging to 22 castes and tribes in U.P. (a province in India) had been collected.

The generalized distance measure of Mahalanobis was extensively used in this paper mainly for the following issues and questions : (a) To classify all the groups into a number of clusters so that groups within a given cluster have a smaller D^2 value among themselves than those obtained from groups belonging to other clusters. (b) What are the best t ($< p$) linear combinations of the p variates which would make the sum of all possible D^2's arising out of a number of populations as calculated with these t linear combinations a maximum ?

This paper, among other results, presents a clustering method and some methods for selection of characters.

The analysis presented in the paper led to suggest three basic clusters — namely, the Brahmin group, the Artisan group and the Tribal groups, besides some other clusters in between the above basic clusters.

In order to assess the variations of the mean values $\mu_1, ..., \mu_k$ of k p-variate normal populations with the common dispersion matrix Σ, Mahalanobis, Majumdar and Rao (1949) proposed the following measure, called the generalized variance :

$$V_p = tr(\Sigma^{-1} B),$$

where

$$B = \sum_{\alpha=1}^{k} (\mu_\alpha - \bar{\mu})(\mu_\alpha - \bar{\mu})',$$

$$\bar{\mu} = \frac{1}{k} \sum_{\alpha=1}^{k} \mu_\alpha .$$

Next, they suggested that "the relative gain in using $p + 1$ characters can then be measured conveniently by the ratio of the maximum (or average) value of V_{p+1} to the maximum (or average) value of V_p."

With reference to the set of data considered, Mahalanobis and Rao (1949) demonstrated that the "analysis brings out that no additional information is obtained in the problem of classification by the inclusion of the indices together with the original characters". Appendix 4 of this paper contains a development by C. R. Rao on the problem of representing p dimensional data in lower dimensions. To resolve the question (b) cited earlier in this section, Rao has obtained the transformed variates $l_{(1)}' X, ..., l_{(t)}' X$, where $l_{(1)}, ..., l_{(t)}$ are the eigenvectors corresponding to the largest t eigenvalues of B in the metric of Σ, where $B = \sum_{\alpha=1}^{k} (m_\alpha - \bar{m})(m_\alpha - \bar{m})'$, $\bar{m} = \sum_{\alpha=1}^{k} m_\alpha / k$ with m_α as the mean-veetor in the αth group, and Σ as the common dispersion matrix, k being the total number of groups.

Appendix 3 of this paper contains an important general development by Rao on the distance between two populations. Rao introduced a new geometry of the space of distribution functions induced by the quadratic differential metric; the idea was developed earlier by Rao (1947).

In his last paper on D^2, Mahalanobis also gave a historical note on the D^2-statistic, presenting a summary of the process of evolution of this statistic.

9. Concluding Remarks

The process of the evolution of Mahalanobis' generalized D^2-statistic was initiated by the problem of classification of some specific groups of individuals through anthropometric measurements and the associated problem of devising

appropriate statistical methods to assess similarities or dissimilarities among various groups. Although Mahalanobis was greatly influenced by the related statistical work in England in early part of this century, especially led by Karl Pearson and his associates and later developed by Barnard (1935) and others, he considered anthropometric data on the living (rather than on skulls) and focussed on the question of evolution and admixture of different caste-groups in India.

Fortunately, certain anthropological data sets were made available to him, and he, in his turn, was also looking for reliable data sets on which he could try his ideas. Besides considering Indian data, he also analysed Swedish and Chinese data. But he questioned the reliability of any data set before considering it for statistical analysis. He pointed out the lack of standardization of anthropometric data and suggested certain standardization techniques. He also questioned the reliability of field records and subsequently subjected the records to close scrutiny and corrected them whenever necessary. His works are good reminders to all statisticians regarding the first phase of any statistical analysis.

Mahalanobis was also a pioneer in design of experiments and he questioned the relative relevance of a given character for the problem under consideration. Not only he tried to assess the significance of a given character, he even devised equipment in order to measure appropriate anthropometric characters accurately (1937). He also pointed out the inappropriateness of some anthropometric indices which were then popular among anthropologists.

In the process of devising appropriate statistical methods for assessing group divergence, he started with the only one measure then available, namely Pearson's CRL. To begin with, he could recognize that this coefficient was inappropriate. This led him to formulate his classical D^2-statistic, free from sample sizes. He made himself familiar with the ongoing theoretical development at that stage, and computed (approximately) the moments of this statistic in order to assess its sampling distribution and the standard error, in particular.

His concept of "positional distance" is a novel idea giving information on the evolution of different groups and the (relative) degree of their admixture. He tried to assess the difference between two populations not only with respect to their means but also with respect to the variability, skewness and kurtosis measures. One may recall that these statistical measures were the only ones to consider during that time, following the development of Karl Pearson.

Mahalanobis did recognize that his classical D^2-statistic was not an appropriate measure since it ignored the correlations among the characters. However, he was not aware of the paper by Hotelling (1931) on the generalization of Student's ratio. Eventually, he came out with his formulation of generalized distance incorporating the covariances among the characters. Fisher (1936) noticed the connection between the works of Hotelling and Mahalanobis, and developed a similar concept from the viewpoint of discrimination.

However, the generalized distance measure was developed by Mahalanobis in order to assess group divergence, and not as a test statistic.

It is interesting to note that Mahalanobis did not reanalyze the Anglo-Indian Data or Risley's data through his measure of generalized distance. The conclusions, in that case, could have been somewhat different. However, even in the normal case, the inference using the generalized distance measure may not be uniformly better than the one using the classical D^2-statistic. Mahalanobis was aware of the fact that any method of inference should more or less agree with the known anthropological facts, and any statistical inference, however nicely formulated, does not provide the last word. It would be interesting now to consider these classical data sets from the viewpoint of modern statistics and the current knowledge in anthropology.

Mahalanobis tried to determine the distribution of the classical D^2-statistic by curve-fitting as well as by simulation. It appears from his development that non-central chi-square distribution could be reasonably approximated by a Pearson's type I curve. It would be interesting to study the conditions under which such a simple approximation would be reasonably valid.

Mahalanobis and Rao suggested a measure called as the generalized variance in their (1949) paper for the problem involving more than two groups. Using the standard statistical methods one may now test the significance of the role of an additional character with respect to this measure; this is not available in the literature.

Mahalanobis, and later Mahalanobis and Rao considered the problem of ranking different pairwise distance measures for a set of K groups. The main tool of Mahalanobis was the (large sample) standard error. This is indeed an important problem and should be addressed now with rigour. On the other hand, nothing is known about the property of the validity of the clustering methods suggested by Mahalanobis. Moreover, the K-group classification problem has been explored very little.

Lastly, it may be mentioned that Mahalanobis' pioneering work not only led to considerable research work on D^2 and related measures, but also to a vast body of statistical analysis of a variety of data in different scientific fields. In particular, his measure of distance has led to a rich development of distance measures, initiated mainly by C. R. Rao.

We conclude this article by stating a comment of Mahalanobis (1956) which, in particular, was exemplified by his work leading to the development of the generalized distance.

"Once the purpose for which the statistics is required has been made clear it becomes possible to construct suitable concepts, definitions and standards. Also, as information begins to be collected and utilized it becomes possible to make necessary improvements in the methods of collection and of processing the data and in the utilization of the information more and more effectively".

ACKNOWLEDGMENTS

The author is thankful to Professor C. R. Rao for his encouragement to write this article and to Sri Chitta Bhattacharya, In-charge, I.S.I. Library, for providing the reprints of the early papers of Mahalanobis.

REFERENCES

I. *Early papers in Biometrika*

PEARSON, K. (1909-10). Darwinism, biometry and some recent biology. *Biometrika*, **VII**, 368-387.

BENINGTON, R. C. and PEARSON, K. (1911-12). A study of the Negro skull with special reference to Congo and Gaboon crania. *Biometrika*, **VIII**.

SNOW, E. C. (1913). The intensity of natural selection in man. *Biometrika*, **IX**, 58-68.

RYLEY, K. V., BELL, J. and PEARSON, K. (1913). A study of the nasal bridge in the Anthropoid Apes and its relation to the nasal bridge in man. *Biometrika*, **IX**, 391-445.

ELDERTON, E. M. and PEARSON, K. (1914-15). Further evidence of natural selection in man. *Biometrika*, **X**, 488-506.

PEARSON, K. (1914-15). On the probability that two independent distributions of frequency are really samples of the same population with special reference to recent work on the identity of trypansome strains. *Biometrika*, **X**, 85-143.

— (1918-19). Inheritance of psychical characters. *Biometrika*, **12**, 367.

PEARSON, K. and DAVIN, A. G. (1920-21). On the seasmoids of the knee-joint II : evolution of the sesamoids. *Biometrika*, **13**, 350-400.

TILDESLEY, M. L. (1921). A first study of the Burmese skull. *Biometrika*, **13**, 247-251.

PEARSON, K. (1926). On the coefficient of racial likeness. *Biometrika*, **18**, 105-117.

MORANT, G. M. (1926). A first study of craniology of England and Scotland from neolithic to early historic times, with special reference to Anglo-Saxon skulls in London museums. *Biometrika*, **18**, 56-98.

— (1928). A preliminary classification of European races based on cranial measurements. *Biometrika*, **20**. 301-375.

PEARSON, K. (1928). The application of the coefficient of racial likeness to test the character of samples. *Biometrika*, **20**, 294-300.

II (a). *Papers by Mahalanobis on D^2-statistic and related areas*

(1922). Anthropological observations on the Anglo-Indians of Calcutta. Part I. Analysis of male stature. *Rec. Indian Museum*, 23, 1-96.

(1925). Analysis of race mixture in Bengal. *Jour. Asiatic Soc. Bengal*, **23**, 301-333.

(1928). A statistical study of Chinese head. *Man in India*, **8**, 107-122.

(1928). On the need for standardization in measurements on the living. *Biometrika*, **20A**, 1-31.

(1930). A statistical study of certain anthropometric measurements from Sweden. *Biometrika*, **22**, 94-108.

(1930). On tests and measures of group divergence. *Jour. Asiatic Soc. Bengal*, **26**, 541-588.

(1931). Anthropological observations on the Anglo-Indians of Calcutta. Part II, Analysis of Anglo-Indian head length. *Rec. Indian Museum*, **23**, 97-149.

(1933). Revision of Risley's anthropometric data relating to the tribes and castes of Bengal. *Sankhyā*, **1**, 76-105.

(1934). A revision of Risley's anthropometric data relating to Chittagong Hill tribes. *Sankhyā*, **1**, 267-276.

(1936). On the generalized distance in statistics. *Proc. Nat. Inst. of Science. India*, **2**, 49-55.

(1937). On the accuracy of profile measurements with a photographic profiloscope. *Sankhyā*, **3**, 65-72.

(1937). (with R. C. BOSE and S. N. ROY) Normalization of statistical variates and the use of rectangular coordinates in theory of sampling distributions. *Sankhyā*, **3**, 1-34. (Appendix by P. C. Mahalanobis, pp. 35-40).

(1937). (with B. N. DUTTA). Note on the foot and stature correlation of certain Bengali castes and tribes. *Sankhyā*, **3**, 245-248.

(1940). Anthropological observations on the Anglo-Indians of Calcutta. Statistical analysis of measurements of seven characters. *Rec. Indian Museum*, **23**, 151-187.

(1943). (with C. BOSE). Correlation between anthropometric characters in some Bengal castes and tribes. *Sankhyā*, **5**, 249-260.

(1949). (with D. N. MAJUMDAR and C. R. RAO). Anthropometric survey of the United Provinces, 1941 : a statistical study : *Sankhyā*, **9**, 90-324.

II(b). *Related short notes and abstracts*

(1920). On the stability of anthropometric constants for Bengal caste data. *Proc. Asiat. Soc. Bengal*, **16**.

(1920). Statistical notes on anthropometric sub-group constants for Bengal caste data. *Proc. Asiat. Soc. Bengal*, **16**.

(1921). Note on the criterion that two samples are samples of the same population. *Proc. Asiat. Soc. Bengal*, **17**.

(1921). The statistical constants of an Anglo-Indian sample. Part I, II : head length and head breadth. *Proc. Asiat. Soc. Bengal*, **17**.

(1934). On the statistical divergence between certain species of Phytophthora. *Proc. Asiat. Sci. Acad. (Bombay)*, **21**.

(1935). Relation between heights and weights of Bengali women (with G. NANDI, J. DHAR and S. SEN). *Proc. Indian Sci. Cong. (Calcutta)* **22**.

(1938). On the distribution of Fisher's taxonomic coefficient. *Proc. Ind. Sci. Cong. (Calcutta)* **52**.

III. *Other references cited in this article*

BARNARD, M. M. (1935). The secular variations of skull characters in four series of Egyptian skulls. *Ann. Eug.*, **6**, 352-371.

BOSE, R. C. (1936). On the exact distribution and moment coefficients of the D^2-statistic. *Sankhyā*, **2**, 143-154.

BOSE, R. C. and ROY, S. N. (1938). The distribution of Studentised D^2-statistic. *Sankhyā*, **4**, 19-38.

BOSE, S. N. (1936) On the complete moment coefficients of the D^2-statistic. *Sankhyā*, **2**, 385-396.

— (1937). On the moment coefficients of the D^2-statistic, and certain integral and differential equations connected with the multivariate normal populations. *Sankhyā*, **3**, 105-124.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eúg.*, 7, 179-188.

HOTELLING, H. (1931). The generalisation of Student's ratio. *Ann. Math. Statist*, **2**, 360-378.

MAHALANOBIS, P. C. (1956). Statistics must have a purpose. *Presidential address, Third Pakistan Statistical Conference, Lahore.*

RAO, C. R. (1947). The problem of classification and distance between two populations. *Nature*, **159**, 30.

— (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58-79.