AD NUMBER

ADB200462

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to DoD only; Administrative/Operational Use; Oct 1950. Other requests shall be referred to USAF School of Aviation Medicine, Ranfolph AFB, TX.

AUTHORITY

DTIC Form 55, dtd November 27, 2001, Control No. 1036003.

UNITED STATES AIR FORCE

School of AVIATION MEDICINE

DISCRIMINATORY ANALYSIS

I. SURVEY OF DISCRIMINATORY ANALYSIS

"DTIC USERS ONLY"

PROJECT NUMBER 21-49-004
REPORT NUMBER 1
(Formerly Project Number 21-02-105)

PROJECT REPORT

DISCRIMINATORY ANALYSIS

I. SURVEY OF DISCRIMINATORY ANALYSIS

Joseph L. Hodges, Jr., Ph.D.
Associate Professor of Mathematical Statistics
Statistical Laboratory
University of California, Berkeley

PROJECT NUMBER 21-49-004
REPORT NUMBER 1*
(Formerly Project Number 21-02-105)
* A University of California report under
Contract No. AF41(128)-8

"DTIC USERS ONLY"

USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS
OCTOBER 1950

Précis

OBJECT:

To survey, in as nontechnical a manner as possible, the extensive literature on discriminatory analysis and related topics.

SUMMARY:

The literature on discriminatory analysis and related topics is reviewed. A bibliography of over 250 references is appended. Mathematical research projects are suggested in relation to the medical and psychological problems of Air Force selection and classification programs.

CHAPTER I

Introduction.

The purpose of the present monograph is to survey, in as nontechnical a manner as possible, the extensive literature on discriminatory analysis and related topics which is listed in the bibliography (pages 89 - 115). It seems desirable to indicate briefly the point of view from which the topics were selected and discussed.

In a narrow sense discriminatory analysis may be identified with the finite multiple classification problem: an individual I is known to belong to just one of k specified categories or populations, and must be classified into one of these populations on the basis of whatever evidence is available about I and about the populations. The classification problem becomes statistical when we further specify that the available evidence about I consists of observed values of certain random variables, these random variables having different probability distributions in the different populations.

It did not seem reasonable, however, to place so strict an interpretation on the subject in preparing the present survey. The techniques employed in discriminatory analysis are intimately related to certain techniques, especially the

coefficient of racial likeness and the generalized distance, which were introduced earlier, and it was not possible to convey an adequate idea of the development of discriminatory techniques without first discussing its predecessors. We have therefore devoted Chapter II to the coefficient of racial likeness and Chapter III to the generalized distance. Extensive bibliographal listings are also given for these topics.

Until recently discriminatory analysis has been essentially no more than the application of the linear discriminant function. Correspondingly, a central place has been given to this topic. The discriminant function is introduced in Chapter V; in Chapter VI there is presented in tabular form a collection of its applicators to many scientific fields; and in Chapter VII some of its modifications and extensions are discussed.

The entire topic of multivariate analysis may be regarded as an extension of the discriminant function, but it did not seem reasonable to include in the present work a discussion of multivariate analysis. We have restricted ourselves to a brief indication of the connections between the two topics, given mostly in Chapters IV and VII.

In his invited address at the meeting of the Institute of Mathematical Statistics in Berkeley, California, June 16, 1949, Professor M. A. Girshick pointed out that the development of discriminatory analysis reflects the same broad phases as does the general history of statistical inference. We may

distinguish a Pearsonian stage, connected with the coefficient of racial likeness, followed by a Fisherian stage, connected with the linear discriminant function. Girshick further notes a Neyman-Pearson stage and a contemporary Waldian stage, which are discussed here in Chapters VIII and IX, respectively. These stages are marked by the introduction of the notions of probability of misclassification, and of risk.

As is indicated by the fact that the bibliography contains over 250 listings, it was impossible to give a thorough discussion to all of the literature. In making the selection of the papers to be presented at length, two principles have been followed. We have tried to present in some detail the ideas which marked important conceptual advances, rather than those which correspond to technical elaborations. And, other things being equal, we have preferred the simpler topics to the more complicated ones. This preference was of course dictated by the desire to have the monograph accessible to persons of limited training in mathematical statistics.

The bibliography was compiled by scanning recent volumes of the main statistical journals, by consulting bibliographic reference works such as Mathematical Reviews, Educational Index, Statistical Methodology Index, Psychological Abstracts, and Biological Abstracts, and by tracing back the bibliographic references in the papers themselves. Much of this work was done by the assistants, and I have particularly to thank Mr. Charles Kraft for doing most of the final checking for

accuracy. We tried to make the bibliography as complete as possible, and would appreciate having omissions brought to our attention. References in the text to the bibliography are made by giving author's name and date. A list of periodicals is given at the end of the report.

In conclusion I should like to thank my friends and colleagues, Dr. Evelyn Fix and Professor E. L. Lehmann, who have gone through much of the manuscript and have made many constructive changes. Our thanks are also due to the various scholars who have made available to us their unpublished manuscripts; in particular we thank T. W. Anderson, Z. W. Birnbaum, G. W. Brown, D. G. Chapman, C. L. Chiang, and M. A. Girshick.

CHAPTER II

The Coefficient of Racial Likeness.

Karl Pearson and his colleagues at University College, London, were deeply interested in the possibility that human crania might be used in the study of anthropology and evolution. They formed considerable collections of skulls, which were carefully measured and studied. Frequently the samples were quite small, so that it frequently became desirable to pool closely related samples. Hence there was need for a test of the significance of observed differences between the samples, which could be applied to determine whether such pooling would be appropriate. There were available tests for the significance of difference of two normal samples, in which each observation consisted of a single measurement, but in craniometric work it was usual to measure as many as 50 quantities on each skull. As Pearson saw, there was need for a test which would compensate for the smallness of the samples by the large number of quantities which might be measured on each individual.

As Pearson wrote later, he tackled this problem in 1919. The solution which he obtained was published in 1921, in a paper written by Miss M. L. Tildesley. Miss Tildesley wanted to know whether she should combine two small samples of

Burmese skulls so that the resulting larger sample could be used to give a more reliable estimate of Burmese cranial characteristics. To answer this question, she used the coefficient of racial likeness (which we shall hereafter denote by CRL). The CRL was given a number of slightly differing definitions but in a simple situation it might be defined as follows.

Suppose we have two samples, say a sample S_1 of n_1 individuals (skulls), and a sample S_2 of n_2 individuals. Suppose that on each individual of each sample we measure p traits. Denote the value of the ith trait measured on the jth individual of the αth sample by $x_{\alpha ij}$. From these measurements we compute for each sample and each trait the mean and standard deviation:

$$(1) \qquad \bar{x}_{\alpha i} = \frac{1}{n_\alpha} \sum_{j=1}^{n_\alpha} x_{\alpha ij}, \quad s_{\alpha i} = \sqrt{\frac{1}{n_\alpha} \sum_{j=1}^{n_\alpha} (x_{\alpha ij} - \bar{x}_{\alpha i})^2};$$

$$i=1,2,\cdots,p; \quad \alpha=1,2 .$$

Pearson then would define the CRL to be the quantity

$$(2) \qquad \frac{1}{p} \sum_{i=1}^{p} \left[\frac{(\bar{x}_{1i} - \bar{x}_{2i})^2}{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}} \right] - 1.$$

The motivation of Pearson's definition is approximately as follows. If the two samples do come from the same population, the expected value of $\bar{x}_{1i} - \bar{x}_{2i}$ is 0; and in any case an estimate of the variance of $\bar{x}_{1i} - \bar{x}_{2i}$ is given by

$\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}$. Since biological measurements are often approximately normally distributed, and since arithmetic means tend to be nearly normal even if the averaged quantities are not normal, we may think of

$$\frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}} \; ; \quad i = 1, 2, \cdots, p$$

as being, approximately, normal random variables of unit variance, whose expected values are 0 if the two samples come from the same population, but whose expectations would usually differ from 0 otherwise. Now if the p random variables were independent, a reasonable test of the hypothesis that the samples come from the same population would be provided by examining the sum of their squares:

$$(3) \qquad \sum_{i=1}^{p} \frac{(\bar{x}_{1i} - \bar{x}_{2i})^2}{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}$$

The quantity (3) would have approximately a chi-square distribution of p degrees of freedom, central if the hypothesis were true, non-central otherwise. From the point of view of modern theory, the use of the statistic (3) can be justified by, for example, the likelihood ratio principle. And since the CRL is a linear function of (3) the use of the CRL as a statistic for testing the hypothesis of homogeneity still seems reasonable, provided that the various assumptions mentioned above hold.

Pearson did not suggest that the chi-square distribution be used with the CRL, however. For most of the applications in craniometry, p would be large enough so that the chi-square distribution could be replaced by the normal with negligible loss of accuracy. Pearson gave the first two moments of the CRL (assuming the hypothesis true) and suggested that these be used in referring a computed CRL to a normal table. (It may be noted that the formula for the second moment given in Miss Tildesley's paper is wrong by a factor of 4. This mistake was repeated in a number of subsequent papers, and only corrected in 1926.)

Pearson was well aware that the theoretical justification for his coefficient rested on the assumption of the independence of the traits measured. The correlation of cranial traits had been the subject of much study by his school. Miss Tildesley wrote, "... we do know quite enough to assert that the correlation is never very high between cranial characters which do not have any portion in common, and which are not right and left measurements of homologous characters. It is indeed often wholly negligible." As Pearson pointed out (1926), it is easy in theory to allow for dependence of the traits, but when p is as large as 20 the resulting computations are overwhelming. He recommended that great attention be paid to the selection of traits little correlated with each other within the sampled populations.

From the point of view of the development of discriminatory

analysis it is of great interest to observe that from its inception, the CRL was employed for two rather different purposes. Properly speaking, the CRL is designed as a test statistic, large values of which are supposed to reflect high improbability that the two samples are drawn from the same population. In applying the test, one selects a critical value, say c, and rejects the hypothesis of homogeneity if the CRL exceeds c. The value of c is selected according to the level of significance which we desire our test to have; by increasing the value of c we decrease the probability of rejecting the hypothesis if it is true.

Now suppose π_0, π_1, and π_2 are three populations from each of which we have a sample, say S_0, S_1, and S_2 respectively. Suppose we compute the CRL between S_0 and S_1 and find it to have the value C_1, and correspondingly find the CRL between S_0 and S_2 to have the value C_2. Suppose further that $C_1 > C_2$. We could then select a critical value c which lies between the two CRL's: $C_1 > c > C_2$. At the significance level corresponding to c, we should accept the hypothesis that π_0 and π_2 are identical and reject the hypothesis that π_0 and π_1 are identical. An examination of this situation makes it easy to understand why there is a temptation to say, in such cases, that " π_0 is nearer to π_2 than it is to π_1". If we succumb to this temptation, we shall be using the CRL not as a test statistic, but as a measure of some (as yet undefined) concept of relative degree

of resemblance or divergence in the totality of populations under study.

It should be clear that the temptation to use a test statistic as a metric is not confined to the CRL. If we have any statistic for testing whether two samples are drawn from the same population, the statistic being so constructed that large values are indicative of difference in the populations sampled, then it is rather natural to interpret larger significant values as indicative of greater differences.

For example, Miss Tildesley computes the CRL between French and English skulls (the value being 24.5), and also between Egyptian and Negro skulls (the value being 27.3), and then states "French and English are shown to be almost as far apart racially as Egyptians and Negroes." Both values of CRL are highly significant.

In the years following 1921, Pearson's school carried out many craniometric researches in which the CRL was the principle statistical tool. The chief contributor to this work was G. M. Morant. Morant commented in 1923 on the question of the use of the CRL as a measure of degree of resemblance, in the following terms (Morant 1923, p. 205):

> "the value of [(2)] computed from a number of mean characters of two races is the Coefficient of Racial Likeness between them and it is thus a measure of the probability of the two being random samples from the same population. It is not a true measure of absolute divergence, and must not for a moment

be considered as such, but nevertheless we shall speak of it, for convenience, as if it were an absolute measure of racial affinity."

In spite of this warning, however, Morant and others continued to use the CRL as a metric. The reason for this inconsistency was doubtless the fact that the craniometrists had need for such a metric, and the CRL was the only tool available to them for such a purpose.

Morant was by no means an uncritical user of Pearson's CRL. In 1924 he had this to say on the subject (Morant 1924, p. 12):

"Given two random samples each of ten individuals drawn from the same homogeneous population, the Coefficient of Racial Likeness deduced from the mean characters of the two samples will not differ significantly from zero, and if two samples each of a hundred individuals are drawn from the same population then their Coefficient will also be of the same order. But if two random samples each of ten individuals are drawn from two different populations and then two samples each of a hundred individuals are drawn from the same differing populations it will be found that the Coefficient between the first pair will be very distinctly less than that between the two samples of a hundred individuals each ... . It is for this reason that Coefficients of Racial Likeness may not be compared directly ..."

The reader may have been wondering what the CRL, whether viewed as a test or as a measure, has to do with discriminatory analysis. There is an obvious way in which a measure of divergence can be used for discrimination purposes. If we can measure the divergence of an individual (or a sample) from each of several populations, to one of which it is assumed that the individual (or sample) belongs, then it seems reasonable to assign the individual (sample) to that population from which the measured divergence is least. In a somewhat similar way a test of significance of difference can be used as a discriminator: we assign the individual to that population from which it is significantly different at the largest level of significance.

In 1926 Morant had occasion to deal with a discrimination question in craniometry (Morant 1926b). An ancient skull was discovered in 1888 in the commune of Chancelade in France. It was examined by an anatomist, Dr. Testut, who wrote, "Parmi les races actuelles, celle qui me paraît présenter la plus grande analogie avec l'homme de Chancelade est celle des Esquimaux." Most anthropologists agreed with Testut's conclusion, but some did not. In 1924 Sir Arthur Keith wrote, "...the Chancelade skull, while possessing a few superficial resemblances to Eskimo skulls, is in its essential character just as European as the people of England and France today,"

(quotations from Morant's paper). We have here a clear problem of discrimination and Morant approached this problem biometrically.

Before seeing what Morant did, let us examine the CRL more closely as a possible tool for discrimination. In practice, the CRL is usually employed in a form somewhat different from (2). Let $\sigma_{\alpha i}$ denote the standard deviation of the ith trait in the αth population. It is usually assumed that $\sigma_{1i} = \sigma_{2i}$; in fact, in craniometry it is customary to replace both σ_{1i} and σ_{2i} by a value σ_i obtained from a large standard sample, it being felt that the variation in standard deviation from one race to another is of less importance than the sampling error of the usual small samples. With the assumption $\sigma_{1i}=\sigma_{2i}=\sigma_i$, (2) simplifies to

$$(4) \qquad \frac{1}{p}\sum_{i=1}^{p} \frac{n_1 n_2}{n_1 + n_2}\left(\frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sigma_i}\right)^2 - 1.$$

Now if we wish to compare a single individual with each of several different races, we would compute (4) between a first sample, consisting of the single individual, and a second sample, consisting in turn of each of the races. Thus n_1 would be 1, $\bar{x}_{1i}$ would be the value, x_{1i}, of the ith trait for the individual, and (4) would become

$$(5) \qquad \frac{1}{p}\sum_{i=1}^{p} \frac{n_2}{n_2 + 1}\left(\frac{x_{1i} - \bar{x}_{2i}}{\sigma_i}\right)^2 - 1.$$

Finally, suppose that we have a large sample from the race;

then $\frac{n_2}{n_2 + 1}$ approximates to 1 and $\bar{x}_{2i}$ will tend in probability to the population mean value, say ξ_i. Thus, the CRL simplifies to

$$\frac{1}{p} \sum_{i=1}^{p} \left(\frac{x_{1i} - \xi_i}{\sigma_i} \right)^2 - 1. \tag{6}$$

We might then reasonably compute the value of (6), using the mean values ξ_i for each race in turn, and assign the individual to that race for which (6) is smallest.

Now let us consider what Morant actually did. To compare the Chancelade skull with male Eskimo skulls, he obtained the values of ξ_i and σ_i from large samples of modern male Eskimo skulls, and computed, for $p = 55$ traits, the values of the quantities

$$\frac{x_{1i} - \xi_i}{\sigma_i} . \tag{7}$$

If the Chancelade skull were Eskimo, we should have here observed values of 55 (supposedly independent) normal deviates, and might use these values to test the hypothesis that the Chancelade skull is Eskimo. The corresponding test might be made to determine whether the Chancelade skull resembles, say, modern English skulls. Morant actually makes two sets of such tests--by computing both the sample mean and standard deviation of the quantities (7) and comparing them with their "theoretical" values. Morant's conclusion was: "...from the evidence afforded by the skull and mandible, we may accept as

a reasonable working hypothesis the statement that the Chancelade individual was distinctly closer to the Eskino than to the modern English."

Since the standard deviation of the quantities (7) is a function of the form (6) assumed by the CRL in this situation, it turns out that one of the two tests made by Morant amounts to the use of the CRL as a discriminator. However, it is rather curious that the CRL is not explicitly mentioned by Morant; in fact, this is about the only craniometric work which Morant did in this period without mentioning the CRL. It is a rather curious historical fact that the connection of the CRL with discrimination did not come in the direct way just discussed, but only in the roundabout fashion outlined in the next chapters.

In 1926 Pearson published the first considerable theoretical work on the CRL. In this paper, Pearson deals with the independence assumption underlying his coefficient. In fact, he suggests an alternative form of the coefficient, which is suitable if the traits are not nearly independent, and if there are only a few of them. Let $r_{\alpha st}$ denote the sample correlation between the sth and t-th traits in S_α. Just as it is convenient to assume $\sigma_{1i} = \sigma_{2i} = \sigma_i$ it is convenient to assume $r_{1st} = r_{2st} = r_{st}$ Let

$$(8) \qquad y_i = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sigma_i} .$$

Let R denote the correlation matrix of $y_1, y_2, \cdots, y_p$:

$$(9) \qquad R = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \cdots & & & \\ r_{p1} & r_{p2} & \cdots & 1 \end{vmatrix}$$

and let R_{st} denote the cofactor of R at the sth row and t-th column. Then it is known that

$$(10) \qquad \frac{1}{|R|} \sum_{s=1}^{p} \sum_{t=1}^{p} R_{st}\, y_s\, y_t$$

will, if the samples are drawn from the same population and the matrix R is exact, have a chi-square distribution with p degrees of freedom. The quantity (10) may be considered to be a generalization of the original CRL (4), to which it reduces if the traits are independent.

Pearson points out the great labor involved in computing (10) when p is as large as, say, 20. He concludes that 'for the statistician, as for the statesman, the ideally best is not always the wisest course.

In 1928 Morant returned to the difficulty he had pointed out in 1924, that arises when one wishes to use the CRL as a measure of dispersion in cases in which the sample sizes differ widely. He suggested a corrective factor to be applied to reduce the CRL to a standard sample size. Morant's criticism and suggested correction are very similar to those offered at

about the same time by P. C. Mahalanobis, and we shall defer discussion till the next section. Finally in 1928 Pearson gave way before the arguments of Morant and Mahalanobis (K. Pearson 1928b), and sanctioned a corrective factor which in essence reduces the CRL to the D^2 statistic discussed in the next section.

After 1928 numerous papers applying the CRL to craniometric work continued to appear in Biometrika. Further theoretical work shifted into other lines, however. The D^2 statistic, introduced originally as a modification of the CRL, was studied extensively by the Indian school, with a steady development of the relevant distribution theory culminating in a paper by Bose and Roy in 1938. And in the West, work of Fisher and Hotelling on different but related problems prepared the way for the introduction of the linear discriminant function in 1935. In an important paper of Fisher in 1938, these various lines of development were brought together. We shall trace the important features of these researches in the next three chapters.

In the bibliography there is an extensive listing of papers pertaining to the CRL. Among these are Batrawi and Morant (1947), von Bonin (1931a, 1931b, 1936), von Bonin and Morant (1938), Cleaver (1937), Collett (1933), Dingwall and Young (1933), Goodman and Morant (1940), Harrower (1928), Hasluck and Morant (1929), Hooke (1926), Hooke and Morant (1926), Kitson (1931), Kitson and Morant (1933), Layard and Young (1935), Little (1943), Martin (1936), Morant (1923,

1924, 1925, 1926a, 1926c, 1927a, 1927b, 1928a, 1928b, 1929a, 1929b, 1931, 1935, 1936a, 1936b, 1937, 1939a, 1939b), K. Pearson (1926, 1928a, 1928b), Reid and Morant (1928), Risdon (1939), Stoessiger (1927), Stoessiger and Morant (1932), Tildesley (1921), Woo (1930), Woo and Morant (1932), and Young (1931). The bulk of these papers contain only routine applications of the CRL to craniology, and are devoid of theoretical interest. Of greater interest are certain papers which approach the CRL in a critical spirit. We have already mentioned some of the comments of Morant, and those of Mahalanobis will be further discussed in the next chapter. In this regard one may mention Pearl and Miner (1935), Fisher (1936a), and Seltzer (1937).

Certain other writers proposed coefficients similar to the CRL, independently of and sometimes earlier than Pearson. Joyce (1912) credits to H. E. Soper a "differential index" which resembles the CRL except that the terms are not squared; this reduces the statistical efficiency. A still more primitive coefficient is that of Aebly (1926), in which differences are not compared with their variabilities, but are summed directly.

CHAPTER III

The Generalized Distance.

In 1923-1925, P. C. Mahalanobis was engaged in an anthropometric study of the Anglo-Indians of Calcutta, and of their relations to other racial groups. He at first employed the then recently devised CRL as a principal statistical tool, but (as had been Morant) was disturbed by the influence of sample size on the CRL when it was used as a measure of the divergence of two populations. On what appear to have been rather intuitive grounds, Mahalanobis decided to drop the coefficient $\frac{n_1 n_2}{n_1 + n_2}$ and obtained in this way a statistic

$$(1) \qquad D^2 = \frac{1}{p} \sum_{i=1}^{p} \left(\frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sigma_i} \right)^2 .$$

This statistic, called at first the "caste-distance" and later the "generalized distance", was used by Mahalanobis in the presidential address delivered to the Anthropological Section of the Indian Science Congress in 1925 (Mahalanobis 1927), which was published in 1929.

The contrast between the CRL and D^2 is made clear if we consider what happens when the sample sizes n_1 and n_2 are increased. If there is in fact no difference between the

populations with regard to the means of the traits, the distribution of CRL will remain unchanged; the increase of $\frac{n_1 n_2}{n_1 + n_2}$ will serve precisely to counterbalance the tendency of $\bar{x}_{1i} - \bar{x}_{2i}$ to approach 0. On the other hand, if there is an actual difference in the population means, say $\delta_i \neq 0$, then $(\bar{x}_{1i} - \bar{x}_{2i})^2$ will tend in probability to the positive quantity δ_i^2 as the sample sizes are increased. Consequently the CRL will tend in probability to ∞. It is thus, as Morant saw in 1924, unreasonable to use the CRL as a measure of divergence unless the sample sizes are always the same. This difficulty does not arise in the case of D^2. If ξ_{1i} and ξ_{2i} denote the population means, then as the sample sizes are increased, D^2 tends in probability to

$$\Delta^2 = \frac{1}{p} \sum_{i=1}^{p} \left(\frac{\xi_{1i} - \xi_{2i}}{\sigma_i} \right)^2 . \tag{2}$$

We may therefore view the sample quantity D^2 as a point estimate of the corresponding population quantity Δ^2, and state that the estimate is consistent (i.e., tends in probability to the quantity being estimated as the sample sizes are increased).

Mahalanobis has stated (1949, p. 237) that he presented the foregoing argument to Karl Pearson in 1927, and that Pearson refused to admit its validity. In any case, Mahalanobis began to use his D^2 statistic, and in 1928 Morant published a very similar argument, together with numerical data showing

the tendency of the CRL to increase with the sample size. Morant suggested that CRL's based on widely different sample sizes could be made comparable by corrective factors. Pearson in the same year endorsed Morant's suggestion, whose effect is in practice to make the CRL very similar to D^2.

From 1930 to 1938 the Indian school devoted much effort to developing the distribution theory of the D^2 statistic. In reviewing the history of this research, it will be convenient to introduce some terminology to describe the various assumptions under which one may study the distribution of D^2 and related statistics.

The reader may have been disturbed by the way in which Pearson and his followers employ for the standard deviations σ_i quantities obtained from extraneous sources, and ignore the sampling variability of these estimates. Practically, if the samples are large, the variability of sample estimates of the variance will not make a major contribution to the distribution of the CRL or of D^2. In a sense, the values of the variances are of secondary importance to the values of the mean differences. But as the theory of statistics develops refinement, and its methods are applied to smaller samples, it becomes desirable to take into account the sampling fluctuation of σ_i. It was the great contribution of Student (1908) to recognize that the ratio of the mean deviation of a normal sample to the estimate of the standard deviation based on the sample, did not have a normal distribution. It seems reasonable to distinguish, therefore, be-

tween the "classical" and "Studentized" versions of the distribution theory problem. If it is considered that the quantities σ_i and r_{st} which are employed represent true population values, we shall say that the problem is being treated in its "classical" form; while if account is taken of the fact that these quantities are estimated from the samples, we shall refer to the problem as "Studentized."

The problem may be further characterized by either making or not making the assumption that the traits are independent. As with the CRL, the D^2 statistic was at first considered only in the case that the quantities $\bar{x}_{1i} - \bar{x}_{2i}$; $i = 1, 2, \cdots, p$, are independently distributed. By 1935, however, the obvious extension involving the addition of correlational terms had been made: the corresponding extension for the CRL was made by Pearson in 1926. The dependent version of D^2, using the notation of (10), Chapter II, is given by

$$(3) \qquad \frac{1}{|R|} \sum_{s=1}^{p} \sum_{t=1}^{p} R_{st} \frac{\bar{x}_{1s} - \bar{x}_{2s}}{\sigma_s} \frac{\bar{x}_{1t} - \bar{x}_{2t}}{\sigma_t} .$$

A third categorization of the distribution problem follows by observing that the distribution of D^2 may be sought either in case the populations sampled are the same (which we shall refer to as the central case), or in case the populations differ for at least one trait (which we shall refer to as the noncentral case). In summary, we may seek

the distribution of D^2 (or of the CRL) for the classical or Studentized, for the independent or dependent, and for the central or noncentral, cases. There are thus in total eight possible situations.

In the terminology just introduced, we may say that Pearson in 1921 gave an approximate distribution for the central, classical, independent CRL, and that in 1926 he gave the exact distribution for the central and classical CRL, which turned out to be the same (chi-square) regardless of independence.

In the same terminology, P. C. Mahalanobis considered the independent, classical case, both central and noncentral, in 1930. This was the first considerable paper on the theory of the D^2 statistic. By a method thought to be approximate (series expansion), Mahalanobis obtained the first four moments of D^2. From these, and from large scale sampling experiments, he was able to state: "We conclude therefore that the distribution of D^2 will conform generally to Type I of the Pearsonian family, except in the case of two groups (or samples) taken from the same population, when the distribution will pass into the Type III curve."

In the mid-1930's, the distribution problem of D^2 was attacked by R. C. Bose. In 1935 Mahalanobis had published the dependent form of D^2 mentioned above, and Bose first considered the classical D^2, in both the independent and dependent cases, both centrally and noncentrally. He was able

to obtain the exact distribution, and hence the moments. It was found that the results of Mahalanobis were exact, and were correct without the independence assumption.

R. C. Bose continued to work on the problem, trying to remove the assumption that the covariance matrix is known. In 1936 Mahalanobis defined explicitly the "Studentized" form of D^2, and reported that Bose had found the first four moments of D^2 in the noncentral Studentized case. Finally, in January, 1938, Bose and S. N. Roy were able to report to the first session of the Indian Statistical Congress that they had succeeded in solving the complete problem: they had found the distribution of D^2 in the Studentized case, whether central or noncentral, whether independent or correlated.

The chairman of the meeting was R. A. Fisher, and at the end of the paper of Bose and Roy, Fisher rose to point out that he had given (however, in connection with a quite different statistical problem) the distribution which they had obtained, in a paper published in 1928. It was also pointed out that Hotelling in 1931 had obtained, also in another connection, the Bose-Roy distribution for the central Studentized case. It is reported that Fisher remarked "that he, and Professors Hotelling and Mahalanobis had been unwittingly treading the same ground. He was glad to avail himself of the present opportunity to clear up this point." In the same year (1938) Fisher published in his journal a paper pointing out the close connection between several independent lines of development.

The Indian school has continued to develop the theory of D^2, usually without reference to parallel developments in the West. Roy and Bose (1940) have modified the D^2 statistic to permit the covariance estimates to be based on several samples while the mean differences are based on two. Bhattacharyya and Narayan (1941) have investigated the D^2 moments when the population variances are unequal. A. Bhattacharyya (1946) has extended the D^2 statistic to the measurement of divergence between multinomial distributions. P. K. Bose (1947a, 1947b, 1949) has developed recursion formulae with the aid of which he has tabled percentage points of the central and noncentral D^2 distribution, in both the classical and Studentized cases. Bose is apparently unaware of the relation of his distributions to the chi-square and F distributions, and as a result seems in some cases to have duplicated existing tables.

The D^2 statistic has recently been used as a major tool in a very extensive anthropological investigation (Mahalanobis, Majumdar, and Rao), which comprises Parts 2 and 3 of Volume 9 of Sankhya. The paper has several appendices in which various theoretical points are discussed.

CHAPTER IV

Beginnings of Multivariate Analysis.

The coefficient of racial likeness and the generalized distance share a feature which serves to distinguish them from much of the preceding work in statistics. Both of these techniques represent attempts to deal with inference problems in which the data consists of several correlated (normal) measurements, say $X_1, X_2, \cdots, X_p$, made on each individual or experiment considered. These statistics are therefore precursors of the theory of multivariate (normal) analysis, a prominent example of which is the linear discriminant function. Before discussing the linear discriminant function it will be useful to describe briefly some developments of multivariate analysis, most of which occurred between 1928 and 1938.

Beginning with the publication of Student's revolutionary paper in 1908, the English school of statisticians have devoted much effort to obtaining analytical expressions for the distributions of commonly used statistics based on normal samples. Previously, in 1900, Karl Pearson had obtained the chi-square distribution as an approximate distribution for a test of goodness of fit. In Student's 1908 paper, the chi-square distribution was offered as the distribution of the sample variance of a normal sample, and a distribution

equivalent to what is now known as the Student t-distribution was given for the ratio of sample mean to sample standard deviation. The next major advance occurred in 1915, when R. A. Fisher, in finding the distribution of the correlation coefficient computed from a bivariate normal sample, introduced his method of geometrical argument in Euclidean hyperspace. However much this method may fall short of present-day requirements for rigor, in the hands of Fisher it was to produce in the next fifteen years revolutionary results. In 1921 Fisher applied his geometrical argument to find the distribution of the intraclass correlation coefficient. The distribution was labelled by Fisher with the letter z, a symbol now famous in statistics. It subsequently developed that the z-distribution had applications far more important than those to the intraclass correlation coefficient; in fact, it turned out to be the general distribution needed to establish the level of significance of all analysis of variance tests. In a transformed version it is now widely known as the F-distribution, having been so named by Snedecor in Fisher's honor. In a series of papers from 1921, Fisher and others gradually extended the statistical usefulness of the F-distribution. Kolodziejczyk (1936) reduced its use to the canonical form of tests of linear hypotheses.

In 1928 Fisher published another paper which is basic for the development of discriminatory analysis. Again employing the geometrical approach, he obtained the formula for the distribution of the multiple correlation coefficient

for normal variables. Although this was the immediate purpose of his work, it is rather two other results, given more or less as corollaries, which concern us. As a limiting form of the multiple correlation coefficient distribution, Fisher obtained a distribution, which he labelled (B); and as a variant form he obtained a third distribution labelled (C). The (B) distribution is today known as the noncentral chi-square distribution, and Fisher recognized that it "may be interpreted as the distribution of the sum of squares of n variates normally distributed with equal variance, but not with zero means." The distribution (C) is what is now known as the noncentral F-distribution, whose main present day use is in determining the power of analysis of variance tests. Needless to say, Fisher did not put his (C) distribution to such a use in 1928, but he did discuss one example (the distribution of a correlation ratio) which serves as precursor to the modern use.

It is interesting that the necessary analytic work had been done by 1928 for finding all of the eight distributions mentioned in connection with D^2 in the preceding chapter. In spite of this it took ten years for the statistical applications of these distributions to D^2 to be realized; and when they were, the realization came independently to two different investigators.

In 1931 the central Studentized case was obtained by Hotelling. Hotelling was interested in extending the work of Student to normal vectors. Student's t-distribution made it

possible to test the hypothesis that the mean of a normal population has a specified value, without assuming knowledge of the value of the variance. Suppose that instead of a sample from a univariate normal population we have a sample from a multivariate normal population, and wish to test simultaneously hypotheses specifying the values of the population means of all components of the normal vectors involved. Hotelling had previously studied problems of this kind while participating in an investigation of the flow of particles in protoplasm (Baas-Becking, et al, 1928). To deal with this testing problem, Hotelling suggested (apparently on intuitive grounds) a test statistic, termed by him T^2, and obtained its distribution. The T^2 statistic is a direct generalization of the Student t , and is, except for a constant multiplier, identical with the correlated form of the CRL, given by Pearson in 1926, where the variances and correlations are not assumed population values, but values estimated from the sample. The distribution of T^2 obtained by Hotelling is simply the central F-distribution first found by Fisher in 1921. Hotelling's great contribution was to show that Fisher's distribution was the appropriate one for a large class of testing problems, including one of interest to us. T^2 may be used to test the hypothesis that two multivariate samples have been drawn from the same normal population, assuming that the samples come from normal populations having the same covariance matrix. We proceed to describe this test in some detail.

Using the notation employed in Chapter II, let $x_{\alpha ik}$

denote the value of the i^{th} trait measured on the k^{th} individual from the α^{th} sample, where $\alpha = 1, 2$, $k = 1, 2, \cdots, n_\alpha$, and $i = 1, 2, \cdots, p$. Let $\bar{x}_{1i}$ and $\bar{x}_{2i}$ be the arithmetic means of the values of the i^{th} trait in the first and second samples, respectively, and define

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad n = n_1 + n_2 - 2,$$

$$n\, a_{ij} = \sum_{k=1}^{n_1} (x_{1ik} - \bar{x}_{1i})(x_{2jk} - \bar{x}_{2k}) + \sum_{k=1}^{n_2} (x_{2ik} - \bar{x}_{2i})(x_{2jk} - \bar{x}_{2j}).$$

Now form the matrix A of the quantities a_{ij}:

$$A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & & & \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix}$$

Next invert the matrix A, to obtain the matrix A^{-1} with elements a^{ij}. (It is this matrix inversion which begins to present great practical difficulties if p is very large). Hotelling's statistic is then

$$T^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} a^{ij}\, d_i\, d_j.$$

The hypothesis is rejected when T^2 is too large, the critical value for rejection being of course set according to the level of significance desired. Hotelling proved that $\frac{n+1-p}{n \cdot p} T^2$ has the distribution of F with p and n+1-p degrees of freedom. The critical values may therefore be taken from the widely available tables of percentage points of the F-distribution.

The test just discussed is pertinent to the discrimination problem, since there is no point in worrying about which of two populations an individual comes from unless the two populations are distinguishable. In the applications of the linear discriminant function (Chapters V and VI) it is customary first to employ the T^2 test to establish the difference of the populations involved.

The choice by Hotelling of the T^2 statistic seems to have been based on intuition. It is interesting that this particular statistic may be obtained by applying a general principle, and that it has certain optimum properties. Neyman and Pearson (1928) proposed the likelihood ratio criterion for obtaining statistical tests, and applied this criterion in 1930 and 1931 to obtain tests of the hypotheses that two or more univariate normal samples arose from the same population. Wilks (1932) obtained tests for a number of multivariate normal hypotheses by application of the likelihood ratio principle. In particular, Wilks found the likelihood ratio statistic for testing the hypothesis that k p-variate samples came from the same population, assuming that the samples arose from normal populations having the same (unknown) covariance matrix.

When $k = 2$, Wilks' result reduces to Hotelling's.

Wilks also found the likelihood ratio test for the hypothesis that several normal populations have the same covariance matrix. Since the assumption of equal covariance matrices underlies the linear discriminant function, the latter test is sometimes used as a preliminary to discrimination. Bartlett (1937) proposed a modification of the constant factors of the Wilks criterion and other modifications and applications of these procedures have been considered extensively, for example by Lawley (1938, 1939), Bishop (1939), and Bishop and Nair (1939). Exact tables are available only for the case $p = 1$ (Thompson and Merrington, 1943).

Work on T^2 is still continuing. Hsu (1938) investigated the noncentral T^2-distribution, that is, the distribution of the T^2 test statistic in case the sampled populations are in fact different. He found that the noncentral T^2-distribution coincides with the noncentral F-distribution investigated by Tang (1938) and with the (C) distribution of Fisher (1928). Because of the identity of T^2 and D^2, Hsu's result is equivalent to that of Bose and Roy (1938) discussed in Chapter III. In 1941, Simaika, following the lead of Hsu (1941), demonstrated that T^2 has the greatest power of any test whose power depends only on the distance (Δ^2) between the populations. Further optimum properties of the T^2 test are known. Wolfowitz (1949) showed that the T^2 test is the most stringent similar test, and Hunt and Stein showed the T^2 test to be most stringent and the uniformly most powerful

invariant test (see Lehmann, 1950). Further work has been done by Hotelling (1947) and Hsu (1945).

Two other papers by Hotelling (1933 and 1936) are also related to the problem of discrimination. In the earlier paper, he considered the problem of finding that rotation in p-dimensional space such that in the new coordinate system the coordinates would be independently distributed. These new coordinate directions were termed by Hotelling the "principal components" of the given multivariate normal distribution. It is interesting that the essential idea of Hotelling's work was anticipated by Pearson (1901). Girshick (1936) showed the equivalence of Hotelling's results with those obtainable from the maximum likelihood principle. This topic is discussed in greater detail in Part II of the present monograph.

In 1936 Hotelling considered the relations which may exist between two correlated sets of random variables. He showed how it was possible to rotate the sample space so that in the new coordinates, the variables of each set are independent among themselves, while between the sets there is dependence only between certain corresponding pairs of variates. These variates are called the "canonical variates" and the correlations between them, the "canonical correlations." This work is related to the linear discriminant function, since the latter may be viewed as a canonical variate. Waugh (1942) has illustrated the application of canonical variates to economic data.

CHAPTER V

The Linear Discriminant Function.

The first clear statement of the problem of discrimination, and the first proposed solution to that problem, were given by R. A. Fisher in the middle of the 1930's. As was the case with Karl Pearson's CRL, the ideas of Fisher first appeared in print in papers by other people [Barnard (1935), Martin (1936)], but it will be convenient to begin with a discussion of Fisher's own first work on the subject. This was contained in his paper, "The use of multiple measurements in taxonomic problems," which appeared in Annals of Eugenics in 1936.

In this paper, Fisher develops his theory largely by means of working out a numerical example, and he is not always careful to state precisely the assumptions which underlie his conclusions. In the exposition of his work which follows, it has been necessary at various points to infer what is meant.

The general situation studied by Fisher is as follows. There are, say, two populations, π_1, and π_2. From each population we have available a sample, say n_1 items from π_1 and n_2 items from π_2. There is then presented a new item, say I, which may have come from either π_1 or π_2. The decision problem is to assign I to one of the two

populations. The available information consists of measurements of, say, p quantities, $X_1, X_2, \cdots, X_p$, which are made on I and on each of the $n_1 + n_2$ sample items.

We may conveniently approach Fisher's solution by considering first the special univariate case, $p = 1$. We then have two univariate samples, whose values may be represented by numbers $x_{11}, x_{12}, \cdots, x_{1n_1}$ for the first sample, by $x_{21}, x_{22}, \cdots, x_{2n_2}$ for the second sample, and by x for the individual I. It seems reasonable to assign I to that group which it more nearly resembles as indicated by the measurements. We might, for example, compute the arithmetic means, say $\bar{x}_1$ and $\bar{x}_2$, of the two samples, and then see to which of these means x is closer. This is in fact the procedure which Fisher proposes. (It may be noted that Fisher's rule implies that the two possible errors of classification are treated symmetrically. This matter is discussed at length in Chapters VIII and IX).

Fisher deals with the multivariate problem by reducing it to the univariate problem just stated. This is accomplished by replacing, for each individual, the p measurements by a single measurement, say Y. There are of course many different ways in which p quantities may be combined to produce a single quantity, but Fisher considers only <u>linear</u> combinations,

$$Y = \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_p X_p.$$

We may here use any set of coefficients $(\lambda_1, \lambda_2, \cdots, \lambda_p)$,

and the major accomplishment of Fisher is to give a reasonable solution to the problem of choosing the coefficients in the most advantageous way.

Let us denote the measured value of the j^{th} quantity on the k^{th} individual in the i^{th} sample by x_{ijk}; $i = 1,2,$; $j = 1,2,\cdots,p$; $k = 1,2,\cdots,n_i$; and denote the measured value of the j^{th} quantity on I by x_j. Correspondingly, let

$$(1) \qquad y_{ik} = \lambda_1 x_{i1k} + \lambda_2 x_{i2k} + \cdots + \lambda_p x_{ipk}$$

and

$$(2) \qquad y = \lambda_1 x_1 + \cdots + \lambda_p x_p.$$

The appropriatness of the choice of values for $\lambda_1, \lambda_2, \cdots, \lambda_p$ may be measured by the relative ease of classifying I through use of the numbers y and y_{ik}. If the two y samples are widely spaced and each is tightly clustered about its own mean:

x xxxx x x o ooooo o o

it will in general be easier to make a correct decision about I than if the y samples overlap:

x xoxo x ox o o

Fisher introduces a numerical measure of the ease of distinguishing between the two populations. This is the ratio:

$$\frac{\text{difference of sample means}}{\text{standard error within samples}}.$$

He then is able to suggest a reasonable criterion for determining appropriate values of $\lambda_1, \lambda_2, \cdots, \lambda_p$: "what linear function of the ... measurements ... will maximize the ratio of the difference between the [sample] means to the standard deviations within [samples]?"

Mathematically, the problem is to maximize the ratio

$$(3) \qquad \frac{|\bar{y}_2 - \bar{y}_1|}{\sqrt{\sum_1 (y_{1k} - \bar{y}_1)^2 + \sum_2 (y_{2k} - \bar{y}_2)^2}}$$

where $\bar{y}_i = \frac{1}{n_i} \sum_i y_{ik}$, and $\sum_i$ denotes summation over the i^{th} sample. We do not need to divide the denominator by the constant $n_1 + n_2 - 2$, since constant factors do not affect the maximization problem, and we may equally consider the square of (3), since this is more convenient mathematically and since the non-negative quantity will be maximized when its square is maximized.

A little computation shows

$$\bar{y}_1 - \bar{y}_2 = \sum_{j=1}^{p} \lambda_j d_j$$

where $d_j = \frac{1}{n_2} \sum_2 x_{2jk} - \frac{1}{n_1} \sum_1 x_{1jk}$ is the difference in sample means for the j^{th} quantity, and

$$\sum_1 (y_{1k} - \bar{y}_1)^2 + \sum_2 (y_{2k} - \bar{y}_2)^2 = \sum_{j=1}^{p} \sum_{m=1}^{p} \lambda_j \lambda_m s_{jm}$$

where S_{jm} is the pooled sum of products of deviations from the sample means of traits j and m:

$$S_{jm} = \sum_{i=1}^{2} \sum_{i} (x_{ijk} - \bar{x}_{ij})(x_{imk} - \bar{x}_{im}).$$

Here the quantities d_j and S_{jm} are computed from the sample measurements.

Our problem then is to determine the values of the λ_j for which

$$(4) \qquad \frac{\left(\sum_{j=1}^{p} \lambda_j d_j \right)^2}{\sum_{j=1}^{p} \sum_{m=1}^{p} \lambda_j \lambda_m S_{jm}}$$

is maximized. Since (4) is not altered when all of the λ's are multiplied by the same quantity, there will be many equally good solutions, differing only by a constant factor. Ordinary methods of the calculus give the solution. If we differentiate (4) with respect to λ_r and set the derivative equal to 0, $r = 1,2,\cdots,p$, we obtain the equations

$$(5) \qquad \frac{\sum_{m=1}^{p} \sum_{j=1}^{p} \lambda_j \lambda_m S_{jm}}{\sum_{j=1}^{p} \lambda_j d_j} \cdot d_r = \sum_{j=1}^{p} \lambda_j S_{jr}, \quad r=1,2,\cdots,p$$

Since we are only interested in solution up to proportionality, we may ignore the factor $\frac{\sum\sum \lambda_j \lambda_m S_{jm}}{\sum \lambda_j d_j}$ which

is the same for all equations, and obtain as a solution the roots λ_j of the equations

$$
\begin{aligned}
s_{11}\lambda_1 + s_{12}\lambda_2 + \cdots + s_{1p}\lambda_p &= d_1 \\
s_{21}\lambda_1 + s_{22}\lambda_2 + \cdots + s_{2p}\lambda_p &= d_2 \\
\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \\
s_{p1}\lambda_1 + s_{p2}\lambda_2 + \cdots + s_{pp}\lambda_p &= d_p .
\end{aligned}
$$

We have thus in practice to solve a set of p simultaneous linear equations; and as was pointed out by Karl Pearson in 1926 this places a practical limitation on the value of p.

Having obtained the appropriate λ's, we can now compute the corresponding quantities $\bar{y}_1$, $\bar{y}_2$, and y, according to (1) and (2). The problem becomes a univariate one, and we can, for example, classify I into π_1 if and only if y is closer to $\bar{y}_1$ than to $\bar{y}_2$.

It will be appropriate to give a numerical illustration, and for historical reasons it seems desirable to use the illustration employed by Fisher. Fisher considers the problem of distinguishing between species of Iris plant on the basis of four measurements made on each plant: sepal length, sepal width, petal length and petal width. He has samples of 50 from each of 2 species, I. setosa and I. versicolor (a third species, I. virginica, is included in making genetical applications). Fisher's example is unfortunate, in that a single

one of these characteristics will serve to do all of the discriminating that anyone would ever need. Thus, the 50 Iris setosa plants have petal lengths running from 1.0 to 1.9 cm while the 50 Iris versicolor plants have petal lengths from 3.0 to 5.1 cm. Clearly no refined statistical technique is needed to distinguish between such populations!

An excellent illustration of the linear discriminant function may however be obtained if we ignore the figures on petal length and width, and pretend that only the figures on sepal length and width are available. Figure 1 shows the two samples, Iris setosa and versicolor, plotted for the sepal measurements. An inspection of this diagram shows just wherein the value of the linear discriminant function lies. If we considered sepal length and sepal width separately (see Figure 1) it would be quite difficult to make an accurate discrimination because of the large degree of overlap of the two samples. But if we compute the linear discriminant function, the discrimination becomes very good.

The figures involved are the following, letting π_1 be Iris setosa and π_2 be Iris versicolor:

$$d_1 = 0.930 \qquad d_2 = -0.658$$

$$S_{11} = 19.1434 \qquad S_{21} = S_{12} = 9.0356 \qquad S_{22} = 11.8658 .$$

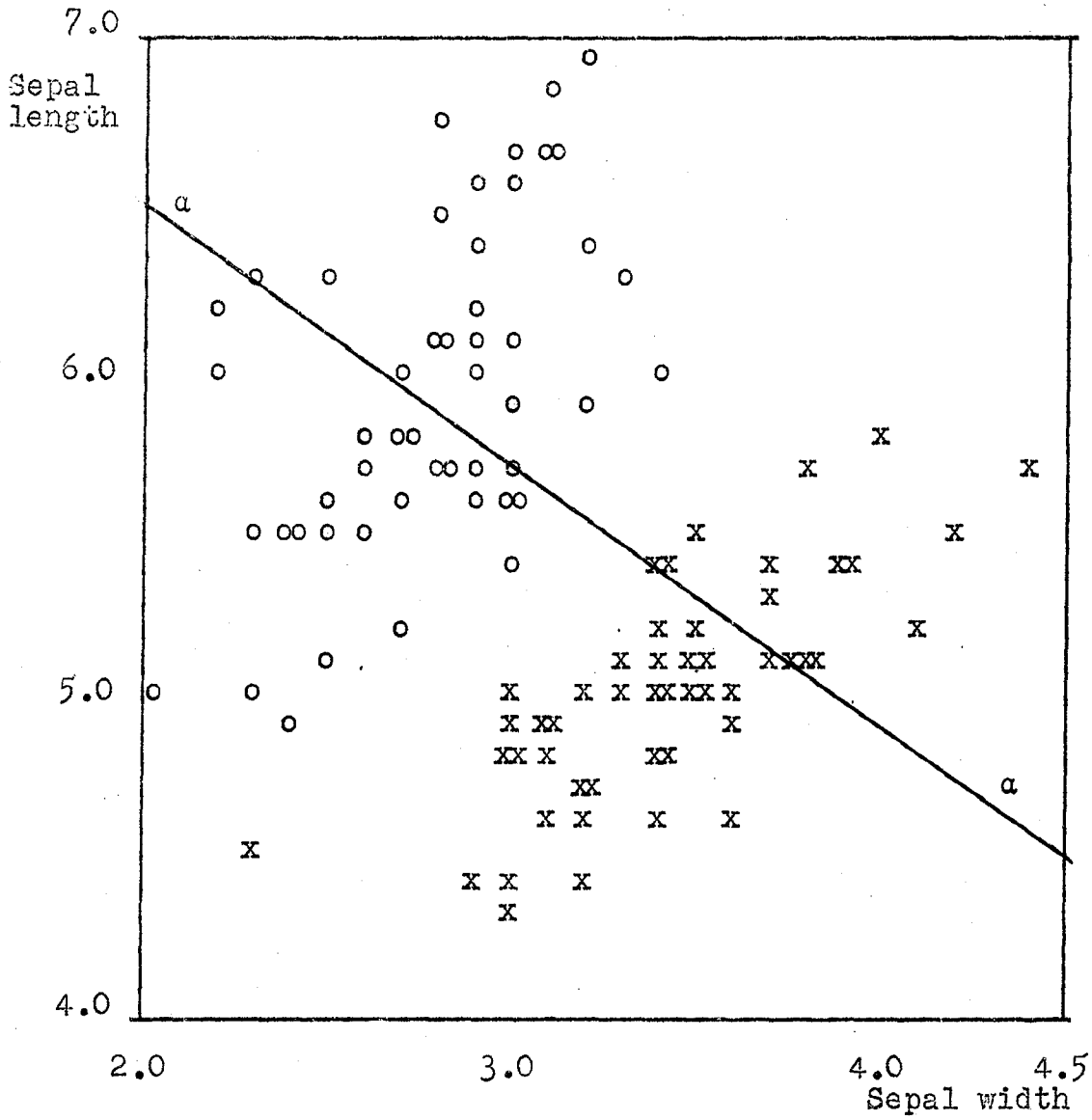

Figure 1.

Fifty Iris setosa plants (x) and fifty Iris versicolor plants (o) plotted for sepal width and sepal length.

We have then to solve two linear equations in two unknowns:

$$19.1434\ \lambda_1 - 9.0356\ \lambda_2 = 0.930$$

$$9.0356\ \lambda_1 + 11.8658\ \lambda_2 = -0.658\ .$$

The roots are easily found to be:

(6) $\lambda_1 = +0.1167$ $\lambda_2 = -0.1443$

Any pair of numbers proportional to these would serve as well.

A simple geometrical interpretation may be given to the LDF. On figure 1 is drawn the line a whose slope is $\frac{\lambda_2}{\lambda_1} = -0.8086$ If we use, not the coefficients (6) but the proportional coefficients

$$\lambda_1' = \frac{\lambda_1}{\sqrt{\lambda_1^2 + \lambda_2^2}} = 0.6288, \quad \lambda_2' = \frac{\lambda_2}{\sqrt{\lambda_1^2 + \lambda_2^2}} = 0.7776$$

then the LDF

$$y = \lambda_1' x_1 + \lambda_2' x_2$$

amounts to projecting the points (x_1, x_2) onto the line a. The line a is so directed that projecting the samples onto it provides the maximum possible separation of the samples.

We may note in passing that in this particular example,

excellent discrimination could be obtained by using the ratio of sepal width to sepal length; this amounts to a projection through the origin onto a vertical line. In other situations, however, the ratio would be a worthless discriminator. The great virtue of the LDF is that it always projects the samples in the direction which gives the greatest possible separation.

It is interesting to note that a trait which of itself provides little or no discrimination, may still be worth measuring in that it enhances the discriminatory power of other traits. An exposition of this situation has been given by Cochran and Bliss (1948).

In his paper, Fisher makes several interesting comments on the relation of the LDF to other statistical techniques. On the one hand, the LDF corresponds to an analysis of variance, with $(\bar{y}_2 - \bar{y}_1)^2$ corresponding to variance between species and $\sum_1 (y_{1k} - \bar{y}_1)^2 + \sum_2 (y_{2k} - \bar{y}_2)^2$ corresponding to variance within species. On the other hand, the LDF can be considered as the solution of a regression problem. This is done by giving to each population a different value of an artificial variable, say z, and then regressing z on the measurements $x_1, \cdots, x_p$. Through these considerations, Fisher is led to suggest a test of the hypothesis that the two populations are in fact identical. This test is identical with the T^2 test proposed by Hotelling in 1931, which has been discussed in Chapter IV.

In conclusion, it should be pointed out that Fisher makes no attempt to justify on probabilistic grounds his definition of optimum separation, nor his restriction to linear combinations of the measurements. We shall see later, in Chapter VIII that when the two populations are normal and have the same covariance matrix, then the LDF has certain optimum properties. Otherwise it is not optimum.

CHAPTER VI

Application of the Linear Discriminant Function.

Since 1935 the LDF has been applied to an amazing variety of problems. To indicate the diversity of the published applications, we present here in tabular form some thirty-two papers. In each case we give the nature of the groups being discriminated, and the nature of the observed quantities on the basis of which the discrimination is effected. We have purposely omitted from the list papers in which previously published data is reanalyzed, such as Bartlett (1947), Brown (1947), Fisher (1938b, 1940), Garrett (1943), Park and Day (1942), and Penrose (1947).

Not all of the applications in this list are of the simple type described in Chapter V; some involve modifications and extensions of the LDF such as are discussed in Chapter VII. However, in general the applications follow a set pattern: the nature of the groups and observations are described, the LDF is computed, and the significance of the discrimination is tested. Sometimes there is an enquiry into the accordance of the data with the assumptions which underlie the LDF, or an appreciation of the relative discriminatory value of the different variables measured.

It may be noted that the thirty-two papers listed appeared in twenty-one different periodicals, most of which were not specifically statistical in nature.

Authors	Date	Groups Discriminated	Nature of Observed Quantities
Baggaley	1947	Harvard freshmen classified by major subject	Nine scores on a psychological examination
Barnard	1935	Skulls from four periods of Egyptian history	Seven craniometric measurements on the skulls
Baten	1943	Two species of spores	Length and width of spores
Baten	1945	Potatoes from three states	Judges' ratings as to color, mealiness, texture, flavor.
Baten and DeWitt	1944	Coal from two mines	B.T.U., percent volatile material, percent fixed carbon, percent ash
Baten and Hatcher	1944	Students taught cooking by two different methods	Scores on three tests of cooking ability
Beall	1943	The two sexes	Scores on four tests: pictorial absurdities, paper form board, tool recognition, vocabulary
Brier, Schott and Simmons	1940	Ewes to be kept for breeding or not to be kept	Score on body type, and body weight

Authors	Date	Groups Discriminated	Nature of Observed Quantities
Cochran and Bliss	1948	Effects of different doses of insulin on rabbits.	Blood sugar determination
Cox and Martin	1937	Soils with or without Azotobacter	Available phosphate content, total nitrogen content, pH.
Day	1937	Cost categories of transporting logs	Observations on carloads of logs
Day and Sandomire	1942	Four age classes of deer	Seven physical measurements such as dressed weight, length of body, length of hind foot
Durand	1941	Good and bad credit risks	Such factors as length of residence, age, sex, time in present job
Fisher	1936	Three species of Iris	Sepal length and width, petal length and width
Hazel	1943	Sheep to be kept or not to be kept for breeding	Weight, gain on feeding trials, judges' scores on form, etc.
Johnson	1950	Individual engineering students	Letter grades in courses

Authors	Date	Groups Discriminated	Nature of Observed Quantities
Kossack	1945	Students who do or do not succeed in intermediate algebra	Mathematics placement test score, high school mathematics record, and Army general classification test
Lorge	1940	Ninth grade students taking or not taking an industrial arts course	Eleven factors such as age, intelligence and occupational choice
Martin	1936	Sexes of mandibles	Six measurements such as length of condyle, length of corpus and mandibular angle.
Maung	1941	School children classified by county of residence	Scores on eye color
Panse	1946	Hens to be kept or not to be kept for breeding	Egg production, etc.
Rao	1948	Two ethnic groups in prehistoric Britain	Seven cranial measurements
Rao	1948	Ethnic groups in India	Stature, sitting height, nasal depth, nasal height

Authors	Date	Groups Discriminated	Nature of Observed Quantities
Rao and Slater	1949	Neurotic and normal Army officers	Thirteen discreet variates, such as ill health, alcoholism, and unstable work record
Selover	1942	Students classified by major subject	Scores on a battery of tests
Smith, C.	1947	Normal and psychotic persons	Scores on a psychological test and an intelligence test
Smith, H. F.	1937	Varieties of wheat to be developed or to be dropped	Five measurements such as yield, ear number, weight of straw
Stevens	1940	Quality classes of rubber	Scores on deterioration after tests
Tintner	1946b	Consumers and producers goods	Observations on price behavior, such as median length and amplitude of price cycle
Travers	1939	Air pilots and engineer apprentices	Scores on intelligence, dynamometer, and perseveration tests

Authors	Date	Groups Discriminated	Nature of Observed Quantities
Wallace and Travers	1938	Successful and unsuccessful salesmen	Scores on psychological tests
Yardi	1946	Plays classified into three creative periods of Shakespeare	Numbers of redundant final syllables, full split lines, unsplit lines with pauses, speech lines

CHAPTER VII

Some Modifications and Extensions of the Discriminant Function.

During the fifteen years in which the LDF has been in use, a number of papers have been published which are concerned with modifications of the LDF, designed to simplify its application, or with extensions of the LDF to problems somewhat different from the classification problem which led Fisher to its invention. Some of these results are briefly described in the present chapter.

If p, the number of traits measured, is small (say 2, 3, or 4), then there is no special difficulty in solving the linear equations which determine the LDF, even if no computing machine is available. But if p is even as large as 6 or 8, the labor involved begins to be practically prohibitive, and with p greater than 10 few persons will care to tackle the problem aided only by a desk calculator. The labor involved in computing the coefficients S_{ij} increases as p^2, and the labor involved in solving the equations increases about as $p(p!)$.

For this reason there has been a good deal of effort expended in seeking out simple and reasonably satisfactory approximate solutions. There is of course a large general literature on the solution of linear equations, which we shall

not consider here. We do however wish to discuss some work aimed specifically at the equations arising in discriminatory analysis.

The first suggested approximation seems to have been that given by Karl Pearson (1926). He pointed out that if the traits are all independent, then we may replace the system of p linear equations in p unknowns by p equations each involving a single unknown:

$$(1) \qquad s_{ii}\,\lambda_i = d_i; \quad i = 1,2,\cdots,p.$$

These equations present no difficulty even if p runs into hundreds. Of course, the Pearson method is only reasonable if in fact the correlations between traits are not too large. Pearson suggested that the traits to be measured might be chosen with this in mind.

Beall (1945) has investigated the accuracy of approximation (1) and of other approximations for three sets of data, computing in each case the discriminant ratio obtained. In one of his examples (data from Travers 1939) the correlations are mostly small; ranging from -0.41 to $+0.38$, with 10 out of the 15 being between -0.1 and $+0.1$. In this case, the simple equations (1) give a discriminant ratio of 1.27, which may be compared with the ratio of 1.31 obtained by using the correct LDF. But on another example (data from L. S. Penrose), where the correlations run from 0.31 to 0.57, Beall finds that (1) yields a discriminant ratio of 0.94, as compared with the LDF ratio of 1.25.

These results suggest that if most of the correlations are small (say between -0.2 and 0.2) with none of them very large (say an absolute maximum of 0.6), then the simple solution (1) may be used without much loss in discrimination.

Another interesting approximate solution has been given by Jackson (1943). He postulates that all of the correlations have a common value, whose estimate is, say, r, and correspondingly replaces the quantity S_{ij} by the quantity $r\sqrt{S_{ii}\,S_{jj}}$.

If we divide the i^{th} of the linear equations for the discriminant function coefficients by S_{ii}, and let $\mu_j = \lambda_j S_{jj}$, $e_i = \dfrac{d_i}{\sqrt{S_{ii}}}$ we obtain the system

$$
\begin{aligned}
&\mu_1 + r\mu_2 + r\mu_3 + \cdots + r\mu_p = e_1 \\
&r\mu_1 + \mu_2 + r\mu_3 + \cdots + r\mu_p = e_2 \qquad (2) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&r\mu_1 + r\mu_2 + r\mu_3 + \cdots + \mu_p = e_p .
\end{aligned}
$$

These equations may be readily solved. Summing them, we find

$$
[1 + (p-1)r] \sum_{i=1}^{p} \mu_i = \sum_{i=1}^{p} e_i . \qquad (3)
$$

The j^{th} of equations (2) may be written

$$
(1-r)\mu_j + r \sum_{i=1}^{p} \mu_i = e_j . \qquad (4)
$$

Combining (3) and (4), we have

$$(5) \qquad \mu_j = \frac{(1-r)\, e_j + pr(e_j-\bar{e})}{(1-r)\,[1+(p-1)r]}$$

where $p\,\bar{e} = \sum_{i=1}^{p} e_i$. Since we require a solution only up to proportionality, we may use

$$(6) \qquad \sqrt{s_{jj}}\,\lambda_j = (1-r)\, e_j + p\, r(e_j-\bar{e}).$$

There still remains the problem of determining an average value of r. Jackson and Beall suggest various estimates, which are not very dissimilar. A reasonable one is Jackson's, given by Beall as:

$$r = \left\{ z_o^2 - \sum_{i=1}^{p} z_i^2 \right\} \Big/ \left\{ \left(\sum_{i=1}^{p} z_i \right)^2 - \sum_{i=1}^{p} z_i^2 \right\}$$

where

$$(n_1+n_2)z_i^2 = \sum_{j=1}^{n_1} (x_{ij}^1-u_i)^2 + \sum_{j=1}^{n_2} (x_{ij}^2-u_i)^2$$

$$(n_1+n_2)u_i = \sum_{j=1}^{n_1} x_{ij}^1 + \sum_{j=1}^{n_2} x_{ij}^2$$

$$(n_1+n_2)z_o^2 = \sum_{j=1}^{n_1} \left[\sum_{i=1}^{p} x_{ij}^1 - v \right]^2 + \sum_{j=1}^{n_2} \left[\sum_{i=1}^{p} x_{ij}^2 - v \right]^2$$

$$(n_1+n_2)v = \sum_{j=1}^{n_1} \sum_{i=1}^{p} x_{ij}^1 + \sum_{j=1}^{n_1} \sum_{i=1}^{p} x_{ij}^2$$

The computations here are not heavy if a desk calculator is available.

In all of the examples considered by Beall, the results obtained in using (6) compare very favorably with those obtained from the LDF. Where the LDF gives discrimination ratios of 5.03, 1.31, and 1.25, the Jackson method (6) gives 5.00, 1.30, and 1.24, respectively. It should be remarked however, that in all three examples the correlations are not widely divergent.

In using the Jackson technique, one should, where possible, give the scales of measurement a common orientation, so that the correlations will at least tend to have the same sign. Thus, if the measurements are all related to intelligence, then a high score on all tests used should have the same meaning--either high intelligence in all cases or low intelligence in all cases. This result can be obtained by appropriate choice of sign.

In conclusion, we may state that the problem of approximate solutions of the LDF equations deserves further study, both empirical and theoretical. Empirically, more studies of the kind carried out by Beall would be of interest. Theoretically, one might seek mathematical bounds for the loss in discriminatory power which results from using various approximations. Further approximations might also be studied, an obvious one being a combination of those of Pearson and Jackson.

Pending such studies, the experimenter may use the following rules of thumb:

(1) If p is not too large, or if the importance of the problem and accuracy of the data warrant the extra work, the accurate LDF should be found exactly.

(2) If the correlations are believed to be mostly small, the equations (1) should be used.

(3) If the correlations are sizeable but not too divergent in size, the equations (6) should be used, after, however, so taking the orientation of scales that the correlations are all of the same sign.

Theoretically the LDF is designed to solve the problem of assigning an individual to the proper one of two populations. However, from the very beginning (Barnard 1935) the technique has been employed with more than two populations. It is clear that a single linear function will do a good job with more than two populations only when these populations are collinear --that is, when the changes in the means of the p traits, from one population to another, are proportional. As is customary in applied statistics, an assumption which underlies a theoretical result need not be exactly satisfied for the result to be usable. But if the populations are not at least approximately collinear, useful information will be lost if classification is carried out through use of the LDF. It is possible to test the hypothesis of collinearity--tests have been proposed by Fisher (1938), Bartlett(1947b), and Rao (1948b). The two latter authors reexamine Barnard's (1935) data, and

find that the linearity assumption is not reasonable in that case. A visual inspection of Barnard's data will lead to the same conclusion. One might almost make it a postulate that if the samples are large, a test of collinearity will lead to rejection. This may still not preclude the reasonableness of using the LDF, if the departure from linearity, while significant, is not large. If it is large, one may employ more than one discriminant function. This procedure is discussed by Rao (1948c) and Brown (1947), as well as in the papers just cited. For a different approach, see Day and Sandomire (1942).

In the practical applications, after the LDF has been found, it is natural to enquire whether some of the variables contribute enough to the discrimination to warrant their continuance in further studies. The problem is complicated by the fact, mentioned in Chapter V, that the contribution of a variable to the discrimination may be indirect. The problem of omitting a variable from a discriminant function is not essentially different from that of omitting a variable from a multiple regression. Aside from empirical discussions (such as that in Barnard, 1935 and other applicational papers), various authors have proposed tests of the additional discriminatory power contributed by a particular trait or traits. For discussion of the numerical problems involved in dropping a variate, see Cochran (1938) and Quenouille (1949a, 1949b).

The LDF has proved to be a valuable tool in fields of application other than that for which it was originally intended. There is a tendency in the literature to term any linear combination of measurements, in which the coefficients are adjusted to achieve some optimum effect, a "discriminant function," even though the effect sought is not the specific discrimination of groups. This extension is not of course directly pertinent to the problem of classification, and will be dealt with briefly.

An early example of such an extended use of the LDF is provided by H. F. Smith (1937), who found that linear function of several observed characteristics of wheat which correlated most highly with a compound of the corresponding qualities representing economic value. Further examples of extension of the LDF arise when one seeks to assign scores to qualitative characters in such a way as to maximize some effect. Examples of this process may be found in Fisher (1925-1946, pp 289-295), Fisher (1946), Maung (1941) and Johnson (1950).

The extended LDF has even been used to effect a general attack on problems of multivariate analysis (Rao, 1948b). Recall that in Chapter V we introduced the LDF as that (linear) reduction of a multivariate problem to the corresponding univariate problem, which would effect the best separation of the univariate samples. More generally, in performing multivariate tests of significance, we may seek that linear reduction of the data which makes greatest the

apparent significance oeing tested. The tests obtained in this way cannot in general be dealt with through solving systems of linear equations, but the test statistics obtained are functions of the roots of a determinantal equation of the form $|A - \lambda B| = 0$, where A and B are $p \times p$ sample covariance matrices. The sampling theory of these roots and of the test statistics which depend on them is very complicated and will not be dealt with here. In general, the distributions involved have not been tabled, but large-sample approximations are available. References to some of the literature are given in the bibliography. See Anderson (1946, 1948), Anderson and Girshick (1944), Anderson and Rubin (1949), Bartlett (1938, 1941, 1947a), Fisher (1938b, 1939, 1940), Girshick (1939), Hsu (1939, 1940, 1941a, 1941b, 1941c, 1941d), Rao (1946, 1948b), and Roy (1939c, 1940a, 1940b, 1942a, 1942b, 1945, 1946a, 1946b). In his lectures at Columbia, Anderson (1947) has given a thorough treatment from the likelihood ratio viewpoint. General surveys have also been made by Bartlett (1947b) and by Tukey.

CHAPTER VIII

Classification from the Point of View of Probability of Error.

The distinguishing feature of the modern theory of statistical inference is the focusing of attention on the probabilistic behavior of statistical procedures. The approach of the linear discriminant function to the classification problem is essentially intuitive rather than probabilistic: we ask, what linear combination of the measurements best separates the samples? The philosophy underlying the LDF is very similar to that which motivated the development of the analysis of variance by Fisher in the 1920's.

The development of a theory of statistical tests, as distinct from a collection of special examples, may be said to have begun with the introduction of the notion of types of error by Neyman and Pearson in 1928 and 1933. Correspondingly, the initiation of a theoretical attack on the classification problem may be said to have begun when the Neyman-Pearson ideas were adapted to the discriminant function by Welch in 1939. Welch's results were published in a brief note, but the ideas involved are of sufficient importance to warrant a rather full discussion.

Welch considers only the problem of classifying an individual into one of two populations, say π_1 and π_2, and

further restricts the problem by assuming that the probability density function of the measured quantities is completely known within each of the populations. Let $f_1(x_1,x_2,\cdots,x_p)$ denote the probability density of the observable quantities $X_1,X_2,\cdots,X_p$ in π_1, and let $f_2(x_1,x_2,\cdots,x_p)$ be the corresponding density in π_2.

Welch observes that <u>any</u> method of classifying an individual I into one of the two populations on the basis of observations on $X_1, X_2,\cdots, X_p$, amounts to a partition of the p-dimensional "sample space" of the X's into two exhaustive and mutually exclusive regions, say R_1 and R_2, with the rule that I will be assigned to π_1 if the random point with coordinates $(X_1, X_2,\cdots,X_p)$ falls into R_1, and will be assigned to π_2 if $(X_1, X_2,\cdots, X_p)$ falls into R_2. The choice of a rule for classification or discrimination is thus equivalent to the choice of a partitioning of the sample space into the regions R_1 and R_2.

Welch further proposes a criterion on the basis of which the various possible partitions may be compared as to their desirability. He suggests that a partition (or rule for classifying I) be judged on the basis of the <u>probabilities of misclassification</u> which arise when the rule is employed.

Two forms of the problem are treated. First, Welch supposes that there exist <u>a priori</u> probabilities that I comes from the two populations, say probability p_1 that I does in fact belong to π_1, and probability p_2 that I belongs to

π_2. Here of course $p_1 + p_2 = 1$. Using the method of Bayes, we may compute the a posteriori probabilities that I belongs to π_1 and to π_2. These values are

$$\frac{p_1 f_1(x_1, x_2, \cdots, x_p)}{p_1 f_1(x_1, x_2, \cdots, x_p) + p_2 f_2(x_1, x_2, \cdots, x_p)},$$

$$\frac{p_2 f_2(x_1, x_2, \cdots, x_p)}{p_1 f_1(x_1, x_2, \cdots, x_p) + p_2 f_2(x_1, x_2, \cdots, x_p)}$$

respectively. We may then assign I to that population whose a posteriori probability is greatest. This procedure coincides with that which is obtained if we compute the likelihood ratio

$$\lambda = \frac{f_1(x_1, x_2, \cdots, x_p)}{f_2(x_1, x_2, \cdots, x_p)}$$

and assign I to π_1 if $x > \frac{p_2}{p_1}$, otherwise assigning I to π_2. Welch asserts (as is easily shown) that these equivalent rules lead to the minimum possible probability of misclassification.

The solution obtained by Welch under the assumption of the existence of a priori probabilities had an historically interesting precursor. In 1898, Heincke was led, in his study of the races and varieties of herring in the North Sea, to attempt a probabilistic solution of the species problem. Heincke noticed that whereas each of several observable traits of the herring would provide some information as to the variety,

none of these traits considered alone would enable him to make a sufficiently accurate classification. He thus sought a method which would enable him to combine the information obtained from several observed traits. The distinguishing features of his work were, first, that the variables he considered were primarily discrete instead of continuous, and secondly, that he made the assumption of equal a priori probabilities. That is, if there were three possible varieties from which a given herring might have come, Heincke assumed that there was a 1/3 chance that the herring came from each. Heincke's principle of classification, granting his assumption, has a distinctly modern sound: "Das Individuum muss schliesslich derjenigen Rasse zugezahlt werden, fur die das Produkt der Wahrscheinlichkeiten aller Eigenschaften ein Maximum ist."

Heinecke's assumption of equal a priori probabilities corresponds to the ancient "principle of insufficient reason." However, from the frequency interpretation of probability here adopted, this assumption would be reasonable only if, say, the herring had been drawn at random from a master population in which the three varieties were mixed in equal proportions. In general, the validity of the assumption of a priori probabilities seems to be restricted in applications. An interesting example in which there existed known a priori probabilities was considered by Martin (1936). Here, skulls and jawbones were recovered from a large grave, but in the recovery process the jawbones became disassociated from the skulls. In the sexing of the material, it is considerably easier to attach

the correct sex to a skull than to a jawbone. Thus (considerably simplifying the problem for purposes of illustration) we might say that we know the proportion of male and female jawbones, and can use these proportions as known a priori probabilities. The example is exceptional, however, and on the whole a solution of the problem which does not involve the assumption of known a priori probabilities is more frequently needed. We may remark that it is easy to show that a formulation in which there are assumed to exist a priori probabilities which are however unknown, does not essentially differ from a formulation in which no a priori probabilities are assumed to exist. (In the language of Wald's theory, this amounts to saying that the class of Bayes solutions is complete. This point is discussed further in Chapter IX.)

Heincke's work was the stimulus of a line of research on the European continent which seems to have been rather independent of the researches which are the main subject of this paper. Of this European work we may mention that of Zarapkin (1934), Kozminski (1936), and Cavalli (1945). Zarapkin modified the Heincke method, and Cavalli considered the relative merits of the methods of Heincke and Zarapkin. These researches do not seem to have contributed much to the main stream of discriminatory analysis.

The biological problem of species, has, naturally, been the stimulus of a great deal of work on the classification problem. We have already seen that Karl Pearson began with the problem of human racial classification, and Fisher's

first paper on the discriminant function was concerned with taxonomy. Heincke, Kozminski, Zarapkin, and Cavalli were similarly motivated. In this connection there is a wealth of material on mathematical definition of species, mostly not of a probabilistic nature. See, for instance, Joyce (1912), where an idea of Soper's is used; Williams (1929); and Ginsburg (1938), who uses the notion of probability of misclassification to define degrees of biological dissimilarity.

Welch also considers a second form of the problem, in which no a priori probabilities are postulated. Here there are two probabilities of misclassification to be considered:

(1) The probability of classifying I into π_2 when in fact I belongs to π_1,

(2) The probability of classifying I into π_1 when in fact I belongs to π_2.

Let us denote these probabilities by $P(R_2 | \pi_1)$ and $P(R_1 | \pi_2)$, respectively. Welch states (as again is easily shown) that to minimize these two probabilities of misclassification, subject to the condition that they are equal, one again employs a partitioning of the likelihood ratio kind. The region R_1 consists of those points for which $\lambda > k$, the value of k being chosen so that $P(R_1 | \pi_2) = P(R_2 | \pi_1)$.

The reader acquainted with the Neyman-Pearson theory will recognize the foregoing as a slight modification of the fact that the most powerful test of a simple hypothesis against a simple alternative is that based on the likelihood ratio

principle (Neyman and Pearson, 1933a). The only novelty in Welch's work is that the two types of error are treated symmetrically, whereas in the usual formulation of the hypothesis testing problem, the two types of error are treated differently: we place a preassigned limit (called the level of significance) on the probability of one error, and then seek to minimize the probability of the other error. Symmetrization of the problem does not alter its essential mathematical nature.

Welch concludes his brief note by considering an example. He supposes that $X_1, X_2, \cdots, Xp$ have a joint normal distribution with a known covariance matrix $\|\sigma\|$ which is the same in the two populations, the two populations thus differing only in the (known) expectations. Formally,

$$f_k(x_1,x_2,\cdots,x_p) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \cdot \frac{1}{\sqrt{|\sigma|}} \cdot e^{-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(x_i-\theta_{ik})(x_j-\theta_{jk})},$$

$$k = 1, 2.$$

When we form the likelihood ratio the constant factors cancel and the exponential factors combine to give

$$\lambda = e^{-\frac{1}{2}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}\{(x_i-\theta_{i1})(x_j-\theta_{j1}) - (x_i-\theta_{i2})(x_j-\theta_{j2})\}\right]}$$

Thus λ is a monotone function of the double sum in the exponent, which may be simplified. We obtain

$$-\log\lambda = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\sigma^{ij}(\theta_{i1}\theta_{j1}-\theta_{i2}\theta_{j2}) + \sum_{i=1}^{p}\sum_{j=1}^{p}{}^{*}\sigma^{ij}(\theta_{j1}-\theta_{j2})\,x_i.$$

The first term on the right is independent of the sample point, so λ is a monotone function of the second term. But the latter is that to which Fisher's LDF simplifies if the population expectations and covariance matrix are known. Thus Welch's work puts a theoretical basis under the LDF, at least in a special case.

It is important to observe the essential nature of the assumption that the two populations have the same covariance matrix. Without this assumption, the likelihood ratio does not simplify as much as before, and we find that λ is a monotone function of a quadratic function of the sample values. We are then led to a quadratic rather than to a linear discriminant function. Smith (1947) has introduced and employed these quadratic discriminators. The theory of the quadratic discriminant functions has not yet been extensively developed.

From the applicational point of view, Welch's results are obtained under rather severe restrictions. Two of these were removed in 1945 by von Mises. Von Mises considered the problem of classifying the individual into one of several populations, say $\pi_1, \pi_2, \cdots, \pi_k$, instead of only two; and further, he was able to remove the rather undesirable restriction, imposed ab initio by Welch, that the two probabilities of misclassification should be equal. If there are k populations, then the number of possible errors of classification is $k(k-1)$, since the individual may belong to any of the k populations, and then may be misclassified into any of the $k-1$ remaining populations. Thus, the two population problem

gives $1.2 = 2$ errors, the three population problem gives $2 \cdot 3 = 6$ errors, etc. The problem thus becomes very rapidly more complicated with increasing k. We can effect a considerable simplification if we focus attention not on the misclassifications but on the correct classifications, for there are only k of the latter. In the case $k = 2$, we get the same results whether we consider misclassifications or correct classifications, since there are two of each and their probabilities are complementary by pairs. In the general problem, however, a real simplification is implied by considering the correct classifications. This amounts to treating all errors alike for a given true population, but permitting the errors to be considered differently as the true population is changed. We may extend our former notation, letting the sample space be partitioned into k regions $R_1, R_2, \cdots, R_k$, with the rule that I shall be assigned to π_i if and only if the sample point $x=(x_1, x_2, \cdots, x_p)$ falls into R_i. Again, $P(R_i \mid \pi_j)$ will denote the probability that I will be assigned to π_i, given that I belongs to π_j.

Von Mises formulated the problem in the following terms: what classification procedure will maximize the minimum of the probabilities $P(R_i \mid \pi_i)$ of correct classification? (It may be noted that this formulation of the problem amounts to a completely symmetric way of viewing the errors.) As did Welch, von Mises considers that the random variables to be observed have, within each of the k populations, known

density functions, which we may denote by $f_i(x_1,x_2,\cdots,x_p)$, $i = 1,2,\cdots,k$. Using the methods of the calculus of variations, von Mises obtains the results in the following terms: "The partition of the x-space that solves our problem is characterized by two properties: (1) for all k regions R_i the value of $P(R_i|\pi_i)$ is the same; (2) along the border between R_i and R_j the ratio $f_i(x)/f_j(x)$ is constant." Thus, Welch's assumption of equal probabilities of incorrect (and hence of correct) classification comes out as a consequence in von Mises work, and again the optimum partition of the sample space is that given by the ratio of the likelihoods.

The reader who is acquainted with recent developments in the theory of statistical decision functions will have recognized that von Mises' formulation of the problem (i.e., the maximization of the minimum probability of correct classification) is an illustration of the minimax principle. This principle, which seems to have been introduced into the theory of statistics by Neyman and Pearson in 1933, has been the subject of a great deal of modern development primarily by Abraham Wald. Chapter IX is devoted to the application of Wald's ideas to the classification problem.

The main practical disadvantage of the work of Welch and von Mises lies in the assumption made by these writers that the parameters of the normal distributions are all known. The Welch test statistic

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij}(\theta_{j1} - \theta_{j2})\, x_i$$

involves all of the population parameters. In the great majority of applicational problems we do not know the values of σ^{ij}, θ_{j1}, and θ_{j2}, but must rely on estimates of these quantities obtained from samples.

The problem of the estimation of the normal parameters from sample values arises in two main forms:

(1) there are available samples of known origin from π_1 and π_2,

(2) the samples are intermingled, so that we do not know for any individual in the sample the true population of origin.

The second form of the problem is of course much harder than the first. An approach to its solution, in the case of univariate normal samples, was made by Karl Pearson in 1894 by means of his method of moments. This technique has not worked well in practice (see Martin 1936) and is not theoretically efficient. Fisher's maximum likelihood method provides a theoretically better solution. However, the fact is that it is extremely difficult to decompose a mixture of two normal populations unless the populations are very well separated, so that the sample has two clear modes.

Rao (1948) has considered a problem of this kind. He considers the observed frequency distribution of heights of 454 plants, supposed to be of two different types but botanically indistinguishable. Assuming equal variances for the two types, Rao estimates that the sample is drawn from a compound population obtained by mixing in proportions 57% and

43% two normal populations whose means differ by about 1 1/2 standard deviations. He decides, "these estimates can be safely used for interpreting differences in heights," and checks his goodness of fit with a chi-square of 1.30 on 6 degrees of freedom.

The trickiness of problems of this kind is made clear by the following observation, due to Dr. Fix: if we fit a single normal distribution to the same data, we may obtain a fit whose chi-square is 0.68 with 8 degrees of freedom! Thus we obtain a better fit with the simpler model. This fact makes one doubt that there is much safety in Rao's interpretation of height differences, and points out that there is little hope of reliable results in resolving mixtures of normal populations unless the samples are extremely large (in which case departure from exact normality would cause trouble) or unless the population means are separated by a good deal more than 1 1/2 standard deviations.

Fortunately, samples of known origin are usually available so that the problem of estimating the population parameters arises in its simpler form. The obvious modification of the Welch test statistic

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij} (\theta_{j1} - \theta_{j2}) x_i$$

is to replace the unknown parameters by their estimates. This is in fact what the LDF does; it corresponds to the extension of the likelihood ratio principle to the composite hypothesis

case, in which one considers the ratio of maximized likelihoods.

In 1944, Wald considered the problem of finding the distribution of a statistic obtained in the manner just suggested. If s^{ij} is the estimate for σ^{ij}, so that s_{ij} is the usual unbiased estimate for σ_{ij} obtained from the pooled sample data, if $\bar{x}_i$ and $\bar{y}_i$ represent the arithmetic means of the sample measurements on the ith trait in the two samples respectively, and if $(z_1, z_2, \cdots, z_p)$ represent the measurements on the individual I to be classified, then Wald's statistic is

$$U = \sum_{i=1}^{p} \sum_{j=1}^{p} s^{ij} z_i (\bar{y}_j - \bar{x}_j).$$

The relation of U to the LDF is clear. Wald gives the large sample distribution of U (this being essentially the approach of Fisher in 1936) and investigates the exact distribution of U. His results are not simple, and are not in a form available for applicational use. Further work on the distribution needs to be done to make Wald's results more readily available for applications. In this connection, see Harter (1950).

A lengthy paper on the classification problem was published by Rao in 1948. The paper consists of three parts, the second and third of which are concerned with the problem of arranging a system of populations into a hierarchial order, and are hence not directly pertinent to discriminatory analysis. In the first part, Rao reobtains the 1945 results of

von Mises, and extends them in several ways. He develops further his suggestion, introduced in 1947, that the classification problem be modified to permit classification not to be made in certain cases. Thus, the sample space is partitioned into $k + 1$ parts, the usual classification regions $R_1, R_2, \cdots, R_k$, and a remaining part R_0 with the rule that if the sample point falls into R_0 no decision will be reached. It is of course true that in many applicational situations circumstances compel a decision to be reached; but there are problems in which the contrary is true, and for these cases the Rao method permits the construction of a classification rule with preassigned limits on all of the probabilities of misclassification.

Rao extends to several populations the Welch solution of the classification problem with known a priori probabilities. He adopts the idea of Heincke that if nothing is known about the a priori probabilities, they may be assumed to exist and all to be equal. Rao gives explicit statements of the likelihood principle in a variety of special cases.

Another recent work of interest is the 1949 paper of Hoel and Peterson. These authors presume the existence of a priori probabilities, and first obtain the same extension of Welch's work to k populations which was obtained by Rao. They then suppose that the a priori probabilities, while still existing, are not known but may be estimated from a sample. There may also be unknown parameters in the densities $f_i(x_1, x_2, \cdots, x_p)$. A set of estimators will be called opti-

mum if it maximizes the probability of correct classification. The authors then consider conditions under which the maximum likelihood estimates will be asymptotically optimum in this sense.

The Hoel-Peterson paper suggests the following question, which seems to be interesting. A more general formulation of the definition of optimum would be as follows: that classification procedure is optimum which maximizes the probability of correct classification. We may then ask, does this definition coincide with that of Hoel and Peterson--that is, can best use of the sample information be made by first estimating the a priori probabilities and parameter values, and then proceeding to classify as if these estimates were known to be correct? An answer to this question should be possible, using the methods of the general theory of statistical decision functions.

Problems which are essentially classificatory arise constantly in the field of medical diagnosis: the physician must assign the patient to one of several categories, which may be taken to correspond to the state of health and to the various diseases under consideration, or to various classifications of severity of a disease. Not much work seems to have been done toward the construction of a probabilistic theory for diagnosis, perhaps through reluctance to treat diagnosis as a chance phenomenon. A beginning was made recently by Neyman (1947), who proposed a simple probabilistic model which will account for observed variation in X-ray diag-

nosis for tuberculosis. Chiang and Hodges (1948) have continued this line of work. An interesting possibility is that sequential diagnostic schemes might be considered probabilistically. Sobel has initiated an attack on sequential solutions of the classification problem in his doctoral dissertation at Columbia University.

Recently Birnbaum and Chapman have considered a problem which is essentially discriminatory. Suppose we wish to select individuals who have a high value of a quantity Y which is not directly observable, but which is correlated with observable quantities $X_1, X_2, \cdots, X_p$. Birnbaum and Chapman show that if $X_1, X_2, \cdots, X_p, Y$ have a $(p+1)$-variate normal distribution, selection by means of an appropriate linear combination of the X's is optimum in various senses. For example, such a "linear truncation" will maximize the conditional expectation of Y among those selected, the frequency of selection being fixed.

It is disturbing to the theoretical statistician that the classification of an individual into a category may be preceded by other statistical inferences, often carried out with the same data. It seems clear that these preliminary inferences will alter in a serious way the theoretical performance of the discrimination itself. There may even be a whole chain of consecutive inferences. To illustrate, suppose that a statistician is given a set of data consisting of readings on a new serological test. He may first test the homogeneity of the data--is there evidence that the data come from more than

one population? If he decides that more than one population is present, he must then decide how many populations there are. At the same time he tries to formulate a probabilistic model for the observations, consisting of a form of probability distribution for each population. These distributions may contain parameters, which must then be estimated. And finally the sampled individuals may be classified. If it is desirable that theory correspond to reality, then there is need for an inclusive theory which will allow for these multi-stage decision procedures.

A beginning has been made by the Hoel-Peterson paper discussed earlier, where the estimation and classification stages are analyzed together. In another interesting paper Paulson (1949) considers the problem of grouping individuals into a "superior" group and an "inferior" group, or else of deciding that all of the individuals are "neutral." This amounts to a two-stage procedure: first we decide whether there are one or two populations represented; and if we decide there are two populations, we proceed to classify the individuals into them. Paulson proposes an intuitively reasonable procedure and considers its probabilistic behavior in the case of normal observations of known variance. His work opens up many interesting and important problems.

CHAPTER IX

Risk and Minimax Ideas.

We have seen in Chapter VIII that von Mises (1945) defined the optimum procedure for classification into one of several populations as that procedure for which the minimum probability of correct classification is maximized. This formulation marks the introduction of minimax ideas into discriminatory analysis. It is the purpose of the present chapter to describe some recent work in this direction.

The risk and minimax notions seem to have been introduced into statistical literature by Neyman and Pearson (1933b). These authors were concerned with testing hypotheses, but as we have seen, hypothesis testing is analogous to the two-population classification problem, and the generalization to k populations presents no difficulty. The specific extension of the risk and minimax notions to the k-population classification problem has been carried out by Rao (1947c, 1948c), Brown (1948, 1949), and Girshick (1949). We shall here present the notions directly in the extended form.

As was mentioned earlier, in classifying an individual into one of k populations, there are $k(k-1)$ distinct possible errors of classification. The complexity of analysis required for dealing with a large number of different kinds

of error is greatly reduced if we can in some way gauge the seriousness of all of these errors on a common scale. For example, we may be able to attach an economic value to the loss, say w_{ij}, which is incurred when an individual who in fact belongs to π_i is assigned to π_j. Presumably $w_{ii} = 0$, since no error is committed when an individual belonging to π_i is assigned to π_i, but the theory is flexible enough to permit $w_{ii} \neq 0$ and to allow the w_{ij} to be either positive or negative if this is desirable. Here, a negative "loss" would correspond to a gain. There will be k^2 of the quantities w_{ij}, which may be conveniently presented as a $k \times k$ matrix:

$$W = \begin{vmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & \cdots & w_{2k} \\ \cdots & & & \\ w_{k1} & w_{k2} & \cdots & w_{kk} \end{vmatrix}$$

This matrix is known as the "loss matrix," and its specification is not the task of the statistician but depends on the use to be made of the classification after it has been effected. (We may remark that W corresponds to the "pay-off matrix" of the theory of games.)

Certain special cases of W are of interest. If we equate the diagonal terms $w_{11}, w_{22}, \cdots, w_{kk}$ to zero, and give the remaining terms a common (positive) value which we

may take to be 1, the formulation reduces to that considered in Chapter VIII: no attention is paid to correct classifications, and all misclassifications are treated alike, (von Mises, 1945). If $k = 2$, and the diagonal terms are 0, we obtain the matrix

$$\begin{vmatrix} 0 & w_{12} \\ w_{21} & 0 \end{vmatrix}$$

We may think of w_{12} and w_{21} as giving the relative importance of the two types of error in a test of a statistical hypothesis. An interesting illustration of this situation, applied to an Air Force problem, has been given by Berkson (1947).

It should be emphasized that, in spite of the great flexibility of the present approach, it cannot be applied to all problems. There are situations in which the different errors are qualitatively so different that a common scale cannot be constructed for them, or an asymmetry of approach may be compelled by the conditions of the problem. We may need instead to adopt the typical method of hypothesis testing, and set preassigned bounds to the probabilities of certain of the errors. A combination of the loss and error-bound methods may be needed for some problems.

The simplification inherent in the loss approach is attained by the introduction of the idea of <u>risk</u>. The risk is simply the expected loss; that is, the average loss which may

be expected in long-run use of the classification procedure being considered. Recalling that any rule for classifying an individual into one of k populations on the basis of certain observations $X_1, X_2, \cdots, X_p$ corresponds to a partitioning of the p-dimensional sample space of the X's into k regions $R_1, R_2, \cdots, R_k$, let us denote this partitioning by R. The risk which results from using R if I in fact belongs to π_i is then

$$r_i(R) = \sum_{j=1}^{k} w_{ij} P(R_j | \pi_i). \tag{1}$$

If <u>a priori</u> probabilities $p_1, p_2, \cdots, p_k$ exist, there will be an unconditional risk

$$r(R) = \sum_{i=1}^{k} p_i r_i(R).$$

We may reasonably take our objective to be the finding of that classification rule R which minimizes the risk. In the case of <u>a priori</u> probabilities, this objective assumes a very simple form. We seek that partition R of the sample space for which

$$r(R) = \sum_{i=1}^{k} \sum_{j=1}^{k} p_i w_{ij} P(R_j | \pi_i)$$

$$= \sum_{j=1}^{k} \left[\sum_{i=1}^{k} p_i w_{ij} P(R_j | \pi_i) \right]$$

is minimum. The solution of this problem is not very different mathematically from that dealt with by Welch (1939). If there exist known probability density functions $f_i(x_1, \cdots, x_p)$ of the observable variables, for each of the populations π_i, we simply compute the k quantities

$$(2) \qquad c_j = \sum_{i=1}^{k} w_{ij}\, p_i\, f_i(x_1, \cdots, x_p), \qquad j = 1,2,\cdots,k$$

and assign I to that population π_j for which the corresponding quantity c_j is least. Intuitively, c_j is proportional to the a posteriori risk sustained when I is assigned to π_j, and we assign I to that population for which the a posteriori risk is least.

If no a priori probabilities are assumed, or if nothing is known about them, the problem is more complicated. The individual I may belong to any one of the k populations, and we need to consider all k of the conditional risks (1). A natural extension of the approach of von Mises (1945) would be the following: find that partition R for which the maximum of the conditional risks is minimum. Such a partition is termed a minimax partition. The adoption of this definition of optimum corresponds to a pessimistic viewpoint: we don't know anything about the true population of I, and should guard ourselves against the worst possibility--the performance of a classification rule being judged by the risk under the least favorable contingency.

A simplification of the problem, which does not lead to the specification of a unique procedure, but which clarifies the possibilities, is effected if we introduce the notions of admissibility and the complete class. A partition R is said to be inadmissible if there exists some other partition S for which none of the conditional risks are greater than they are for R:

$$r_i(S) \leq r_i(R), \quad i = 1,2,\cdots,k;$$

and such that at least one of the conditional risks is less for S than for R:

$$r_j(S) < r_j(R) \quad \text{for some} \quad j = 1,2,\cdots,k.$$

It is clear that we should not want to use an inadmissible classification rule, since there is available an alternative rule which cannot give higher risk and may give lower risk. If a rule is not inadmissible, it is called admissible, and a collection of rules which contains all admissible rules is called a complete class. From the risk point of view, we need never consider procedures which do not belong to a complete class. The notion of complete class was introduced in connection with hypothesis testing by Lehmann (1947), and was extended by Wald (1947). The concepts of loss, risk, minimax procedure, admissibility, and complete class play a fundemental role in the modern theory of statistical decision functions developed by Abraham Wald (1950). Various theorems relating

to these concepts, for the special case of the k-population classification problem, may be deduced from general theorems of Wald (1950), or may be obtained more simply for the special case. We shall merely state some of the main results.

Even if there are no a priori probabilities, we can introduce them artificially, and consider the class of all classification rules obtainable from (2) when we permit $p_1, p_2, \cdots, p_k$ to assume all possible sets of values. The class of rules so obtained is known as the class of Bayes solutions, and these constitute a complete class. Under certain restrictions one can show that all of the Bayes rules are admissible. The minimax rule turns out to be the Bayes rule for which the risks are all equal (the so-called "constant risk Bayes solution.")

The result of these theorems is to give a theoretical solution of the optimum classification problem, provided (i) the loss matrix W can be specified in a satisfactory way, and (ii) the distribution of the observable variables is completely known within each population. The same comments could be made here that were made about the von Mises results in Chapter VIII. In fact, the present result specializes to the von Mises result when W is appropriately chosen.

Even if provisos (i) and (ii) hold, there remains the practical problem of the explicit determination of the regions R_i. One may proceed by trial and error, chosing values for $p_1, p_2, \cdots, p_k$ arbitrarily, evaluating the corresponding risks, and then correcting the p's to bring the risks

closer to equality. If $k = 2$, it is usually not hard to obtain the explicit minimax procedure, but with $k = 3$ there may already be practical difficulties. There is need for more work on useful approximations and shortcuts in finding the minimax regions when $k \geq 3$. A start has been made by Rao (1948c). The problems which arise are rather different, according as the distributions f_i are discrete or continuous, and both cases deserve investigation.

BIBLIOGRAPHY

AEBLY, J., 1926. Über ein quantitatives Mass für die Ähnlichkeit von Individuen. Zeits. für Konst., 12, 712-715.

ANDERSON, T. W., 1946. The non-central Wishart distribution and certain problems of multivariate statistics. Ann. Math. Stat., 17, 409-431.

1947. Theory of multivariate statistical analysis (lecture notes). Columbia University, 169 pp.

1948. The asymptotic distributions of the roots of certain determinantal equations. J.R.S.S., B, 10, 132-139.

and Girshick, M. A., 1944. Some extensions of the Wishart distribution. Ann. Math. Stat., 15, 345-357.

and Rubin, H., 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. Ann. Math. Stat., 20, 46-63.

BAAS-BECKING, L. G. M., van de Sande Bakhuyzen, H., and Hotelling, H., 1928. The physical state of protoplasm. Verh. K. Akad. Wetens. Amst., Afd. Nat., Sec. 2, 25, No. 5, 53 pp.

BAGGALEY, A. R., 1947. The relation between scores obtained by Harvard freshmen on the Kuder preference record and their fields of concentration. J. Ed. Psych., 38, 421-427.

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

BARTLETT, M. S., 1934. The vector representation of a sample. Proc. Camb. Phil. Soc., 30, 327-340.

1938. Further aspects of the theory of multiple regression. Proc. Camb. Phil. Soc., 34, 33-40.

1939. The standard errors of discriminant function coefficients. Supp. J.R.S.S., 6, 169-173.

1941. The statistical significance of canonical correlations. Biom., 32, 29-37.

1947a. The general canonical correlation distribution. Ann. Math. Stat., 18, 1-17.

1947b. Multivariate analysis. Supp. J.R.S.S., 9, 176-197.

BATEN, W. D., 1943. The discriminant function applied to spore measurements. Mich. Acad. Sci., 29, 3-7.

1945. The use of discriminant functions in comparing judges' scores concerning potatoes. J.A.S.A., 40, 223-228.

and DeWitt, C. C., 1944. Use of the discriminant function in the comparison of proximate coal analyses. Indus. Eng. Chem., Anal. Ed., 16, 32-34.

and Hatcher, H. M., 1944. Distinguishing method differences by use of discriminant functions. J. Exp. Ed., 12, 184-186.

BATRAWI, A., and Morant, G. M., 1947. A study of a First

Dynasty series of Egyptian skulls from Sakkara and of an Eleventh Dynasty series from Thebes. Biom., 34, 18-27.

BEALL, G., 1945. Approximate methods in calculating discriminant functions. Psych., 10, 205-217.

BERKSON, J., 1947. "Cost-utility" as a measure of the efficiency of a test. J.A.S.A., 42, 246-255.

BETTS, G. L., 1949. Test calibration for categorical classification. Ed. Psych. Meas., 9, 269-279.

BHATTACHARYYA, A., 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calc. Math. Soc., 35, 99-109.

1946. On a measure of divergence between two multinomial populations. Sankhya, 7, 401-406.

BHATTACHARYYA, D. P., and Narayan, R. D., 1941. Moments of the D^2-statistic for populations with unequal dispersions. Sankhya, 5, 401-412.

BIRNBAUM, Z. W. Effect of linear truncation on a multinormal population (unpublished).

and Chapman, D. G. On optimum selections from multinormal populations (unpublished).

BISHOP, D. J., 1939. On a comprehensive test for the homogeneity of variances and covariances in multivariate problems. Biom., 31, 31-55.

and Nair, U. S., 1939. A note on certain methods of testing for the homogeneity of a set of estimated variances.

Supp. J.R.S.S., 6, 89-99.

BLISS, C. I. (see Cochran, W. G.).

von BONIN, G., 1931a. Beitrag zur Kraniologie von Ost-Asien. Biom., 23, 52-113.

1931b. A contribution to the craniology of the Easter Islanders. Biom., 23, 249-270.

1936. On the craniology of Oceania. Crania from New Britain. Biom., 28, 123-148.

and Morant, G. M., 1938. Indian Races in the United States. A survey of previously published cranial measurements. Biom., 30, 94-129.

BOSE, P. K., 1947a. On recursion formulae, tables and Bessel function populations associated with the distribution of classical D^2-statistic. Sankhya, 8, 235-248.

1947b. Parametric relations in multivariate distributions. Sankhya, 8, 167-171.

1949. Incomplete probability integral tables connected with Studentised D^2-statistic. Bull. Calc. Stat. Ass., 2, 131-137.

(see Roy, S. N.).

BOSE, R. C., 1935. On the exact distribution and moment-coefficients of the D^2-statistics. Sci. Cult., 1, 205-206.

1936a. A note on the distribution of differences in mean values of two samples drawn from two multivariate normally distributed populations, and the definition of the D^2-statistic. Sankhya, 2, 379-384.

BOSE, R. C., 1936b. On the exact distribution and moment-coefficients of the D^2-statistic. Sankhya, 2, 143-154.

1938. On the distribution of the means of samples drawn from a Bessel function population (abstract). Sankhya, 4, 92.

and Roy, S. N., 1938. The distribution of the Student-ised D^2-statistic. Sankhya, 4, 19-38.

(see Roy, S. N.).

BOSE, S. N., 1936. On the complete moment-coefficients of the D^2-statistic. Sankhya, 2, 385-396.

1937. On the moment-coefficients of the D^2-statistic and certain integral and differential equations connected with the multivariate normal population. Sankhya, 3, 105-124.

BRIER, G. W., Schott, R. G., and Simmons, V. L., 1940. The discriminant function applied to quality rating in sheep. Proc. Amer. Soc. An. Prod., 1, 153-160.

BRODGEN, H. E., 1946. An approach to the problem of differential prediction. Psych., 11, 139-154.

BROOKNER, R. J., 1945. Choice of one among several statistical hypotheses. Ann. Math. Stat., 16, 221-242.

(see Wald, A.).

BROWN, G. W., 1947. Discriminant functions. Ann. Math. Stat., 18, 514-528.

1948. Basic principles for construction and application of discriminators (hectograph). The Rand Corporation,

Paper P-45, 8 pp.

1949. Basic principles for construction and application of discriminators (abstract). Biometrics, 5, 80.

CAVALLI, L. L., 1945. Alcuni problemi della analisi biometrica di popolazioni naturali. Mem. Ist. Ital. Idrobiol., 2, 200-323.

CHAPMAN, D. G. (see Birnbaum, Z. W.).

CHIANG, C. L., and Hodges, J. L., Jr., 1948. Power of certain tests relation to medical diagnosis (abstract). Ann. Math. Stat., 19, 433.

CLEAVER, F. H., 1937. A contribution to the biometric study of the human mandible. Biom., 29, 80-112.

COCHRAN, W. G., 1938. The omission or addition of an independent variate in multiple linear regression. Supp. J.R.S.S., 5, 171-176.

1943. The comparison of different scales of measurement for experimental results. Ann. Math. Stat., 14, 205-216.

1949. The present status of biometry. Proc. Sec. Int. Biom. Conf., 19 pp.

and Bliss, C. I., 1948. Discriminant functions with covariance. Ann. Math. Stat., 19, 151-176.

COLLETT, M., 1933. A study of Twelfth and Thirteenth Dynasty skulls from Kerma (Nubia). Biom., 25, 254-284.

COX, G. M., and Martin, W. P., 1937. Use of a discriminant function for differentiating soils with different Azotobacter populations. Iowa St. Col. J. Sci.,

11, 323-332.

CRAIG, C. C., 1942. Recent advances in mathematical statistics, II. Ann. Math. Stat., 13, 74-85.

DAS , A. C., 1948. A note on the D^2-statistic when the variances and co-variances are known. Sankhya, 8, 372-374.

DAY, B. B., 1937. A suggested method for allocating logging costs to log sizes. J. For., 35, 69-71.

and Sandomire, M. M., 1942. Use of the discriminant function for more than two groups. J.A.S.A., 37, 461-472.

(see Park, B. C.).

DEWITT, C. C. (see Baten, W. D.).

DINGWALL, D., and Young, M., 1933. The skulls from excavations at Dunstable, Bedfordshire. Biom., 25, 147-157.

DURAND, D., 1941. Risk elements in consumer installment financing, technical edition. Stud. Cons. Inst. Fin., No. 8, 163 pp.

DWYER, P. S., 1942. Recent developments in correlation technique. J.A.S.A., 37, 441-460.

EBEL, R. L., 1947. Frequency of errors in the classification of individuals on the basis of fallible test scores. Ed. Psych. Meas., 7, 725-734.

FISHER, R. A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biom., 10, 507-521.

FISHER, R. A., 1921. On the "probable error" of a coefficient

of correlation deduced from a small sample. Metron, 1, No. 4, 3-32.

1928. The general sampling distribution of the multiple correlation coefficient. Proc. Roy. Soc., A, 121, 654-673.

1936a. "The Coefficient of Racial Likeness" and the future of craniometry. J. Roy. Anthrop. Inst., 66, 57-63.

1936b. The use of multiple measurements in taxonomic problems. Ann. Eugen., 7, 179-188.

1938a. On the statistical treatment of the relation between sea-level characteristics and high-altitude acclimatization. Proc. Roy. Soc., A, 126, 25-29.

1938b. The statistical utilisation of multiple measurements. Ann. Eugen., 8, 376-386.

1939. The sampling distribution of some statistics obtained from non-linear equations. Ann. Eugen., 9, 238-249.

1940. The precision of discriminant functions. Ann. Eugen., 10, 422-429.

1946a. Statistical methods for research workers. Tenth Edition. Oliver and Boyd, London, 354 pp.

1946b. A system of scoring linkage data with special reference to the pied factors in mice. Amer. Nat., 80, 568-578.

GARRETT, H. E., 1943. The discriminant function and its use in psychology. Psych., 8, 65-79.

GINSBURG, I., 1938. Arithmetical definition of the species, sub species and race concept, with a proposal for a modified nomenclature. Zoologica, 23, 253-286.

GIRSHICK, M. A., 1936. Principal components. J.A.S.A., 31, 519-528.

1939. On the sampling theory of roots of determinantal equations. Ann. Math. Stat., 10, 203-224.

1949. Bayes, minimax, and other approaches to multiple classification problems (unpublished).

(see Anderson, T. W.).

GOODMAN, C. N., and Morant, G. M., 1940. The human remains of the Iron Age and other periods from Maiden Castle, Dorset. Biom., 31, 295-312.

GUTTMAN, L., 1946. An approach for quantifying paired comparisons and rank order. Ann. Math. Stat., 17, 144-163.

HARROWER, G., 1928. A study of the crania of the Hylam Chinese. Biom., 20B, 245-293.

HARTER, H. L., 1950. On the distribution of Wald's classification statistic (abstract). Ann. Math. Stat., 21, 145.

HARTLEY, H. O., 1946. The application of some commercial calculating machines to certain statistical calculations. Supp. J.R.S.S., 8, 154-183.

HASLUCK, M. M., and Morant, G. M., 1929. Measurements of Macedonian men. Biom., 21, 322-336.

HATCHER, H. M. (see Baten, W. D.).

HAZEL, L. N., 1943. The genetic basis for constructing selection indexes. Genetics, 28, 476-490.

HEINCKE, F., 1898. Naturgeschichte des Herings. Abh. d. Deut. Seef., 2, two volumes, 128 pp. and 223 pp.

HODGES, J. L., Jr. (see Chiang, C. L.).

HOEL, P. G., and Peterson, R. P., 1949. A solution to the problem of optimum classification. Ann. Math. Stat., 20, 433-438.

HOOKE, B. G. E., 1926. A third study of the English skull with special reference to the Farrington Street crania. Biom., 18, 1-55.

and Morant, G. M., 1926. The present state of our knowledge of British craniology in late prehistoric and historic times. Biom., 18, 99-104.

HOTELLING, H., 1931. The generalization of Student's ratio. Ann. Math. Stat., 2, 360-378.

1933. Analysis of a complex of statistical variables into principal components. J. Ed. Psych., 24, 417-441 and 498-520.

1935. The most predictable criterion. J. Ed. Psych., 26, 139-142.

1936a. Relations between two sets of variates. Biom., 28, 321-377.

1936b. Simplified calculation of principal components. Psych., 1, 27-35.

1940. The selection of variates for use in prediction with some comments on the general problem of nuisance

parameters. Ann. Math. Stat., 11, 271-283.

1947. A generalized T measure of multivariate dispersion (abstract). Ann. Math. Stat., 18, 298.

(see Baas-Becking, L. G. M.).

HSU, P. L., 1938. Notes on Hotelling's generalized T. Ann. Math. Stat., 9, 231-243.

1939. On the distribution of roots of certain determinantal equations. Ann. Eugen., 9, 250-258.

1940. On generalized analysis of variance (I). Biom., 31, 221-237.

1941a. Analysis of variance from the power function standpoint. Biom., 32, 62-69.

1941b. Canonical reduction of the general regression problem. Ann. Eugen., 11, 42-46.

1941c. On the limiting distribution of roots of a determinantal equation. J. Lond. Math. Soc., 16, 183-194.

1941d. On the limiting distribution of the canonical correlations. Biom., 32, 38-45.

1941e. On the problem of rank and the limiting distribution of Fisher's test function. Ann. Eugen., 11, 39-41.

1945. On the power functions of the E^2-test and the T^2-test. Ann. Math. Stat., 16, 278-286.

HUNT, G., and Stein, C. Most stringent tests of statistical hypotheses (unpublished).

JACKSON, R. W. B., 1943. Approximate multiple regression weights. J. Exp. Ed., 11, 221-225.

JOHNSON, H. M., 1944. Multiple contingency versus multiple correlation; an old time-saving way of handling multiple contingency. Amer. J. Psych., 57, 49-62.

JOHNSON, P. O., 1950. The quantification of qualitative data in discriminant analysis. J.A.S.A., 45, 65-76.

JOYCE, T. A., 1912. Notes on the physical anthropology of Chinese Turkestan and the Pamirs. J. Roy. Anthrop. Inst., 42, 450-484.

KATZELL, R. H., and Cureton, E. E., 1947. Biserial correlation and prediction. J. Psych., 24, 273-278.

KENDALL, M. G. The advanced theory of statistics. Charles Griffin and Co., London, two volumes, 457 pp and 521 pp.

KITSON, E., 1931. A study of the Negro skull with special reference to the crania from Kenya colony. Biom., 23, 271-314.

and Morant, G. M., 1933. A study of the Naga skull. Biom., 25, 1-20.

KOŁODZIEJCZYK, ST., 1935. On an important class of statistical hypotheses. Biom., 27, 161-190.

KOSSACK, C. F., 1945. On the mechanics of classification. Ann. Math. Stat., 16, 95-98.

KOZMINSKI, Z., 1936. Morphometrische und okologische Untersuchungen an Cyclopiden der strenuus-Gruppe. Int. Rev. d. ges. Hydr. u. Hydr., 33, 161-240.

LAWLEY, D. N., 1938. A generalization of Fisher's z-test. Biom., 30, 180-187.

1939. A correction to "A generalization of Fisher's z-test." Biom., 30, 467-469.

LAYARD, D., and Young, M., 1935. The Burwell skulls, including a comparison with those of certain other Anglo-Saxon series. Biom., 27, 388-408.

LEHMANN, E. L., 1947. On families of admissible tests. Ann. Math. Stat., 18, 97-104.

1950. Some principles of the theory of testing hypotheses. Ann. Math. Stat., 21, 1-26.

LETESTU, S., 1948. Note sur l'analyse discriminatoire. Experientia, 4, 22-23.

LITTLE, K. L., 1943. A study of a series of human skulls from Castle Hill, Scarborough. Biom., 33, 25-35.

LORGE, I., 1940. Two-group comparisons by multivariate analysis. Amer. Ed. Res. Ass., Off. Rep., 117-121.

MADOW, W. G., 1937. Contributions to the theory of comparative statistical analysis. I. Fundamental theorems of comparative analysis. Ann. Math. Stat., 8, 159-176.

1938. Contributions to the theory of multivariate statistical analysis. Trans. Amer. Math. Soc., 44, 454-495.

MAHALANOBIS, P. C., 1922. Anthropological observations on the Anglo-Indians of Calcutta. Part I. Analysis of male stature. Rec. Ind. Mus., 23, 1-96.

MAHALANOBIS, P. C., 1927. Analysis of race-mixture in Bengal. J. As. Soc. Beng., 23, 301-333.

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

1930a. On tests and measures of group divergence. J. As. Soc. Beng., 26, 541-588.

1930b. A statistical study of certain anthropometric measurements from Sweden. Biom., 22, 94-108.

1931. Anthropological observations on the Anglo-Indians of Calcutta. II. Analysis of Anglo-Indian head length. Rec. Ind. Mus., 23, 97-149.

1936. On the generalized distance in statistics. Proc. Nat. Inst. Sci. Ind., 2, 49-55.

1940a. The application of statistical methods in physical anthropometry. Sankhya, 4, 594-598.

1940b. Anthropological observations on the Anglo-Indians of Calcutta. III. Statistical analyses of measurements of seven characters. Rec. Ind. Mus., 23, 151-187.

1949. Historical note on the D^2-statistic. Sankhya, 9, 237-240.

Majumdar, D. N., and Rao, C. R., 1949. Anthropometric survey of the United Provinces, 1941: A statistical study. Sankhya, 9, 89-324.

and Rao, C. R., 1949. A note on the use of indices. Sankhya, 9, 240-246.

MAJUMDAR, D. N. (see Mahalanobis, P. C.).

MARTIN, E. S., 1936. A study of an Egyptian series of mandibles, with special reference to mathematical

methods of sexing. Biom., 28, 149-178.

MARTIN, W. P. (see Cox, G. M.).

MAUNG, K., 1941. Discriminant analysis of Tocher's eye colour data for Scottish school children. Ann. Eugen., 11, 64-76.

1942. Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. Ann. Eugen., 11, 189-223.

MERRINGTON, M. (see Thompson, C. M.).

MINER, J. R. (see Pearl, R.).

von MISES, R., 1945. On the classification of observation data into distinct groups. Ann. Math. Stat., 16, 68-73.

MORANT, G. M., 1923. A first study of the Tibetan skull. Biom., 14, 193-260.

1924. A study of certain Oriental series of crania including the Nepalese and Tibetan series in the British Museum (Natural History). Biom., 16, 1-105.

1925. A study of Egyptian craniology from prehistoric to Roman times. Biom., 17, 1-52.

1926a. A first study of the craniology of England and Scotland from Neolithic to early historic times, with special reference to the Anglo-Saxon skulls in London museums. Biom., 18, 56-98.

MORANT, G. M., 1926b. Studies of Palaeolithic man. I. The Chancelade skull and its relation to the modern

Eskimo skull. Ann. Eugen., 1, 257-276.

1926c. The use of biometric methods applied to craniology, being a critique of Professor Gordon Harrower's "A study of the Hokien and the Tamil skull." Biom., 18, 414-417.

1927a. Studies of Palaeolithic man. II. A biometric study of Neanderthaloid skulls and of their relationships to modern racial types. Ann. Eugen., 2, 318-381.

1927b. A study of the Australian and Tasmanian skulls, based on previously published measurements. Biom., 19, 417-440.

1928a. A preliminary classification of European races based on cranial measurements. Biom., 20B, 301-375.

1928b. Studies of Palaeolithic man. III. The Rhodesian skull and its relations to Neanderthaloid and modern types. Ann. Eugen., 3, 337-360.

1929a. A contribution to Basque craniometry. Biom., 21, 67-84.

1929b. Studies of Palaeolithic man. IV. A biometric study of the Upper Palaeolithic skulls of Europe and of their relationships to earlier and later types. Ann. Eugen., 4, 109-214.

1931. A study of the recently excavated Spitalfields crania. Biom., 23, 191-248.

1935. A study of predynastic Egyptian skulls from Badari based on measurements taken by Miss B. N. Stoessiger

and Professor D. E. Derry. Biom., 27, 293-309.

1936a. A biometric study of the human mandible. Biom., 28, 84-122.

1936b. A contribution to the physical anthropology of the Swat and Hunga Valleys based on records collected by Sir Aurel Stein. J. Roy. Anthrop. Inst., 66, 27-28.

1937. A contribution to Eskimo craniology based on previously published measurements. Biom., 29, 1-20.

1939a. Note on Dr. J. Wunderly's survey of Tasmanian crania. Biom., 30, 338-340.

1939b. The use of statistical methods in the investigation of problems of classification in anthropology. Biom., 31, 72-98.

(see Batrawi, A.; von Bonin, G.; Goodman, C. N.; Hasluck, M. M.; Hooke, B. G. E.; Kitson, E.; Reid, R. W.; Stoessiger, B. N.; Woo, T. L.).

MORROW, D. J., 1948. On the distribution of the sums of the characteristic roots of a determinantal equation (abstract). Bull. Amer. Math. Soc., 54, 75-76.

NAIR, U. S. (see Bishop, D. J.).

NANDA, D. N., 1948a. Distribution of a root of a determinantal equation. Ann. Math. Stat., 19, 47-57.

1948b. Limiting distribution of a root of a determinantal equation. Ann. Math. Stat., 19, 340-350.

1949. The standard errors of discriminant function coefficients in plant-breeding experiments.

J.R.S.S., B, 11, 283-290.

NANDI, H. K., 1946. On the power function of Studentised D^2-statistic. Bull. Calc. Math. Soc., 38, 79-84.

1947. Tests of significance on multivariate samples. Bull. Calc. Stat. Ass., 1, 42-43.

NARAYAN, R. D. (see Bhattacharyya, D. P.).

NEYMAN, J., 1947. Outline of statistical treatment of the problem of diagnosis. Publ. H. Rep., 62, 1449-1456.

and Pearson, E. S., 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Biom., 20A, 175-240.

and Pearson, E. S., 1931. On the problem of k samples. Bull. Int. Acad. Sci. Crac., A, (1931), 460-481.

and Pearson, E. S., 1933a. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans., A, 231, 289-337.

and Pearson, E. S., 1933b. The testing of statistical hypotheses in relation to probabilities a priori. Proc. Camb. Phil. Soc., 29, 492-510.

and Pearson, E. S., 1936. Contributions to the theory of testing statistical hypotheses. Stat. Res. Mem., 1, 1-37.

(see Pearson, E. S.).

PANSE, V. G., 1946. An application of the discriminant function for selection in poultry. J. Gen., 47, 242-248.

PARK, B. C., and Day, B. B., 1942. A simplified method for determining the condition of whitetail deer herds. U. S. D. A. Tech. Bull., No. 840.

PATNAIK, P. B., 1949. The non-central χ^2- and F-distributions and their applications. Biom., 36, 202-232.

PAULSON, E., 1949. A multiple decision procedure for certain problems in the analysis of variance. Ann. Math. Stat., 20, 95-98.

PEARL, R., and Miner, J. R., 1935. On the comparison of groups in respect of a number of measured characters. Hum. Biol., 7, 95-107.

PEARSON, E. S., and Neyman, J., 1930. On the problem of two samples. Bull. Int. Acad. Sci. Crac., A, 73-96.

and Wilks, S. S., 1933. Methods of statistical analysis appropriate for k samples of two variables. Biom., 25, 353-378.

(see Neyman, J.).

PEARSON, K., 1894. Contributions to the mathematical theory of evolution. Phil. Trans., A, 185, 71-110.

1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag., Ser. 5, 50, 157-175.

PEARSON, K., 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag., Ser. 6, 2, 559-572.

1915. On the problem of sexing osteometric material. Biom., 10, 479-487.

1926. On the coefficient of racial likeness. Biom., 18, 105-117.

1928a. The application of the coefficient of racial likeness to test the character of samples. Biom., 20B, 294-300.

1928b. Note on the standardisation of method of using the coefficient of racial likeness. Biom., 20B, 376-378.

PENROSE, L. S., 1945. Discrimination between normal and psychotic subjects by revised examination M. Bull. Can. Psych. Ass., 5, 37-40.

1947. Some notes on discrimination. Ann. Eugen., 13, 228-237.

PETERSON, R. P. (see Hoel, P. G.).

QUENOUILLE, M. H., 1949a. A further note on discriminatory analysis. Ann. Eugen., 15, 11-14.

1949b. Note on the elimination of insignificant variates in discriminatory analysis. Ann. Eugen., 14, 305-308.

RAO, C. R., 1946. Tests with discriminant functions in multivariate analysis. Sankhya, 7, 407-414.

1947a. On the significance of the additional information obtained by the inclusion of some extra variables in the discrimination of populations (letter). Cur. Sci., 16, 216-217.

1947b. The problem of classification and distance between two populations (letter). Nature, 159, 30-31.

1947c. A statistical criterion to determine the group to which an individual belongs (letter). Nature, 160, 835-836.

1948a. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proc. Camb. Phil. Soc., 44, 50-57.

1948b. Tests of significance in multivariate analysis. Biom., 35, 58-79.

1948c. The utilization of multiple measurements in problems of biological classification. J.R.S.S., B, 10, 159-203.

1949a. On a transformation useful in multivariate computations. Sankhya, 9, 251-253.

1949b. On the distance between two populations. Sankhya, 9, 246-248.

1949c. Representation of "p" dimensional data in lower dimensions. Sankhya, 9, 248-251.

and Slater, P., 1949. Multivariate analysis applied to differences between neurotic groups. B. J. Psych., Stat., 2, 17-29.

REID, R. W., and Morant, G. M., 1928. A study of the Scottish short cist crania. Biom., 20B, 379-388.

RIDER, P. R., 1936. Annual survey of statistical technique. Developments in the analysis of multivariate data--

Part 1. Econ., 4, 264-268.

RISDON, D. L., 1939. A study of the cranial and other human remains from Palestine excavated at Tell Duweir (Lachish) by the Wellcome-Marston archaeological research expedition. Biom., 31, 99-166.

ROMANOWSKY, V., 1928. On the criteria that two given samples belong to the same normal population (On the different coefficients of racial likeness). Metron, 7, No. 3, 3-46.

ROY, S. N., 1939a. A note on the distribution of the Studentised D^2-statistic. Sankhya, 4, 373-380.

1939b. On the K-statistic. Sci. Cult., 5, 131-132.

1939c. p-Statistics or some generalisations in analysis of variance appropriate to multivariate problems. Sankhya, 4, 381-396.

1940a. Distribution of certain symmetric functions of p-statistics on the null hypothesis. Sci. Cult., 5, 563.

1940b. Distribution of p-statistics on the non-null hypothesis. Sci. Cult., 5, 562-563.

1942a. Analysis of variance for multivariate normal populations: the sampling distribution of the requisite p-statistics on the null and non-null hypotheses. Sankhya, 6, 35-50.

ROY, S. N., 1942b. The sampling distribution of p-statistics and certain allied statistics on the non-null hypothesis. Sankhya, 6, 15-34.

1945. The individual sampling distribution of the maximum, the minimum and any intermediate of the p-statistics on the null-hypothesis. Sankhya, 7, 133-158.

1946a. Multivariate analysis of variance: the sampling distribution of the numerically largest of the p-statistics on the non-null hypothesis. Sankhya, 8, 15-52.

1946b. A note on multivariate analysis of variance when the number of variates is greater than the number of linear hypotheses per character. Sankhya, 8, 53-66.

and Bose, P. K., 1940. On the reduction formulae for the incomplete probability integral of the Studentised D^2-statistic. Sci. Cult., 5, 773-775.

and Bose, R. C., 1940. The use and distribution of the Studentised D^2-statistic when the variances and covariances are based on k samples. Sankhya, 4, 535-542.

(see Bose, R. C.).

RUBIN, H. (see Anderson, T. W.).

van de SANDE BAKHUYZEN, H. (see Baas-Becking, L. G. M.).

SANDOMIRE, M. M. (see Day, B. B.).

SCHOTT, R. G. (see Brier, G. W.).

SELOVER, R. B., 1942. A study of the sophomore testing program at the University of Minnesota - Part II. J. Appl. Psych., 456-467.

SELTZER, C. C., 1937. A critique of the coefficient of racial likeness. Amer. J. Phys. Anthrop. 23, 101-109.

SIMAIKA, J. B., 1941. On an optimum property of two important statistical tests. Biom., 32, 70-80.

SIMMONS, V. L. (see Brier, G. W.).

SLATER, P. (see Rao, C. R.).

SMITH, C. A. B., 1947. Some examples of discrimination. Ann. Eugen., 13, 272-282.

SMITH, H. F., 1937. A discriminant function for plant selection. Ann. Eugen., 7, 240-250.

STEIN, C. (see Hunt, G.).

STEVENS, W. L., 1940. The standardization of rubber flexing tests. India Rubber World, 102, 36-41.

STOESSIGER, B. N., 1927. A study of the Badarian crania recently excavated by the British School of Archaeology in Egypt. Biom., 19, 110-150.

and Morant, G. M., 1932. A study of the crania in the vaulted ambulatory of St. Leonard's church, Hythe. Biom., 24, 135-202.

STUDENT, 1908. On the probable error of a mean. Biom., 6, 1-

TANG, P. C., 1938. The power function of the analysis of variance tests with tables and illustrations of their use. Stat. Res. Mem., 2, 126-149.

THOMPSON, C. M., and Merrington, M., 1943. Tables for testing the homogeneity of a set of estimated variances. Biom., 33, 296-304.

TILDESLEY, M. L., 1921. A first study of the Burmese skull. Biom., 13, 176-262.

TINTNER, G., 1946a. Multiple regression for systems of equations. Econ., 14, 5-36.

1946b. Some applications of multivariate analysis to economic data. J.A.S.A., 41, 472-500.

TRAVERS, R. M. W., 1939. The use of a discriminant function in the treatment of psychological group differences. Psych., 4, 25-32.

(see Wallace, N.).

TUKEY, J. W., 1948. Vector methods in the analysis of variance (hectograph). 90 pp.

WALD, A., 1944. On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Stat., 15, 145-162.

1947. An essentially complete class of admissible decision functions. Ann. Math. Stat., 18, 549-555.

1950. Statistical decision functions. John Wiley and Sons (in press).

and Brookner, R. J., 1941. On the distribution of Wilks' statistic for testing the independence of several groups of variates. Ann. Math. Stat., 12, 137-152.

WALLACE, N. and Travers, R. M. W., 1938. A psychometric sociological study of a group of specialty salesmen. Ann Eugen., 8, 266-302.

WAUGH, F. V., 1942. Regression between sets of variables. Econ., 10, 290-310.

WELCH, B. L., 1939. Note on discriminant functions. Biom.,

31, 218-220.

WHERRY, R. J., 1947. Multiple bi-serial and multiple point bi-serial correlation. Psych., 12, 189-195.

WILKS, S. S., 1932. Certain generalizations in the analysis of variance. Biom., 24, 471-494.

1935. On the independence of k sets of normally distributed statistical variables. Econ., 3, 309-326.

1938. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psych., 3, 23-40.

(see Pearson, E. S.).

WILLIAMS, O. L., 1929. A critical analysis of the specific characters of the genus Acuaria Nematodes of birds, with descriptions of new American species. Univ. Calif. Publ. Zool., 33, 69-107.

WISHART, J., 1928. The generalised product moment distribution in samples from a normal multivariate population. Biom., 20A, 32-52.

WISHART, J., 1932. A note on the distribution of the correlation ratio. Biom., 24, 441-456.

WOLFOWITZ, J., 1949. The power of the classical tests associated with the normal distribution. Ann. Math. Stat., 20, 540-551.

WOO, T. L., 1930. A study of seventy-one Ninth Dynasty Egyptian skulls from sediment. Biom., 22, 65-93.

and Morant, G. M., 1932. A preliminary classification of Asiatic races based on cranial measurements. Biom.,

24, 108-134.

YARDI, M. R., 1946. A statistical approach to the problem of chronology of Shakespeare's plays. Sankhya, 7, 263-268.

YOUNG, M., 1931. The West Scottish skull and its affinities. Biom., 23, 10-22.

(see Dingwall, D.; Layard, D.).

ZARAPKIN, S. R., 1934. Zur Phänoanalyse von geographischen Rassen und Arten. Arch. Naturgesch., 3, 161-186.

UNIVERSITY OF CALIFORNIA, STATISTICAL LAB., BERKELEY

DISCRIMINATORY ANALYSIS - I - SURVEY OF DISCRIMINATORY ANALYSIS

JOSEPH L. HODGES, JR. OCT 50 115PP. GRAPH

USAF SCHOOL OF AVIATION MEDICINE, RANDOLPH AIR FORCE BASE, TEX., USAF CONTR. NO. AF41-128-8 (REPORT NO. 1)

AVIATION MEDICINE (19)
GENERAL (0)

BIOMETRICS
STATICAL ANALYSIS