DISCRIMINANT FUNCTIONS

By George W. Brown

Iowa State College

1. Introduction: In the following sections the development of discriminant function techniques is approached from an elementary point of view, considering first an essentially trivial problem, then working up to the more complex situations which may be handled by discriminant function methods. No attempt has been made to follow the pattern of the historical development in this process, and no consistent attempt has been made to allocate proper credit, in the text, to those individuals responsible for the introduction and exploitation of these methods. A more or less exhaustive bibliography of discriminant function applications and related theory is given at the end of this paper.

Some historical perspective may be gained, however, from a very sketchy consideration of the early background of the subject. The first published application of the discriminant function seems to have been the work of Barnard (1935 [1]) on craniometry, following the suggestion of R. A. Fisher. Meanwhile P. C. Mahalanobis (1927, [30]; 1930, [31]) and, in this country, Hotelling (1931, [25]) had been concerned with a closely related problem, the construction of measures of the "distance" between two sets of multiple measurements, for which Karl Pearson's (1926, [34]) coefficient of racial likeness was not wholly adequate. Fisher (1936, [18]) gave a further example of the method and showed (1938, [19]) the relation between his work and that of Hotelling (1931, [25]; 1936, [27]). Thus the theory of discriminant function analysis proper is about ten years old, but is intimately related to researches which go back a few more years.

A simple problem: Consider the very simple case of a single measurement, say ξ, which may be made in each of two populations, and let us suppose, for the sake of discussion, that ξ is normally distributed, with unit variance, in each population, but with possibly different means in the two populations.

Let

$$E_1(\xi) = \alpha - \beta$$

$$E_2(\xi) = \alpha + \beta$$

be the mean values of ξ over the two populations, with $\beta > 0$. As an example, we may consider the pH measurements of Iowa soil samples (Cox and Martin, [12]), for two soil populations, distinguished by the presence or absence of Azotobacter. From 100 samples containing Azotobacter and 186 samples containing no Azotobacter, we have the estimated averages of pH equal to 7.423 and 6.015 respectively, with an estimated standard error of .625 within populations (see Fig. 1).

$$\hat{\alpha} = 6.719$$
$$\hat{\beta} = .704$$
$$\hat{\sigma} = .625$$
$$\hat{\beta}/\hat{\sigma} = 1.13.$$

Let us suppose further that ξ is the only measurement available on a single individual, not knowing to which of populations 1 and 2 the individual belongs.

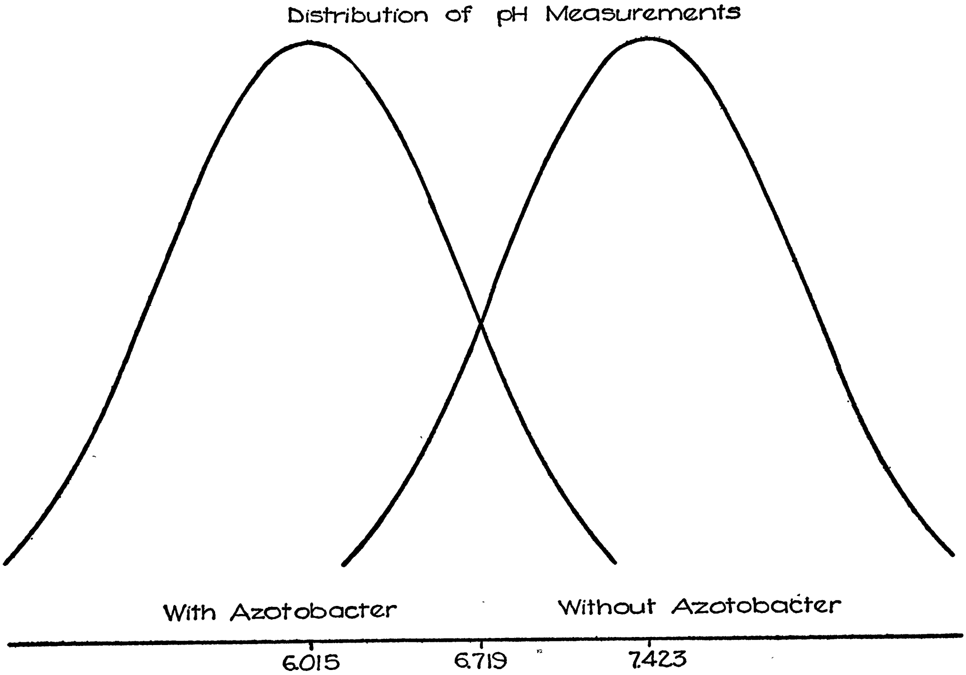

FIG. 1

The problem is to classify this individual as a member of population 1 or population 2. It is clear that ξ furnishes the only information on which to base a decision, and that essentially the only procedure available is to choose a number, say ξ_0, such that we choose population 1 when $\xi < \xi_0$ and population 2 when $\xi > \xi_0$. Furthermore, it is evident that the expected accuracy of classification depends on the size of β. If we wish to have equal risks of misclassification for members of the two populations we choose $\xi_0 = \alpha$. Then the probability of misclassification is given by $P\{\epsilon > \beta\}$, where ϵ is a normal deviate with unit variance. As one would expect, the probability of misclassification tends to 0 as $\beta \to \infty$ and tends to $\frac{1}{2}$ as $\beta \to 0$. In the Azotobacter example, if we assume that the estimates given are the population values, we choose $\xi_0 = 6.719$. The ratio $\hat{\beta}/\hat{\sigma} = 1.13$ is exceeded approximately 13% of the time in sampling from the normal distribution, leading to .13 as the probability of misclassification.

Consider now the slightly more general situation in which we consider a fixed variate, say w with measurements ξ distributed, for fixed w, with a mean of the form $\alpha + \beta w$. This is the standard regression situation. As before assume that ξ is normally distributed about this mean with unit variance, that is

$$\xi = \alpha + \beta w + \epsilon$$

where α and β are constants, w may take on any or all real values, and ϵ is a normal deviate. Note that if w is restricted to take on only two values the structure reduces to the first structure considered. An example of the continuous type might be constructed by considering w as genotypic yield of grain and ξ a phenotypic measure of yield (Smith, [36]).

The simple problem formulated for the two-population case may be reformulated here as follows: Given the relationship $\xi = \alpha + \beta w + \epsilon$, and given ξ for an individual for which no other information is known, how shall we estimate w? For selective breeding the problem may be to select individuals for which w is at one end of the scale, rather than to estimate w itself. Whatever decision is to be made, it is still clear that ξ furnishes the only available information, and that the certainty of the decision is a function of β. Since $(\xi - \alpha)/\beta = w + \epsilon/\beta$, the variance of this estimate of w is $1/\beta^2$. Note that confidence intervals for w, given ξ, may be constructed from the normally distributed quantity $\xi - \alpha - \beta w$.

It should be pointed out that in the usual regression case we are interested in predicting ξ for given w, with the hypothesis as stated above, whereas in this case ξ will be observed, and the problem is that of estimating, as a parameter of the distribution of ξ, the fixed variate w.

Obviously β must not vanish if ξ is to perform any discrimination among w values. In practice, of course, α and β will not be given as known values and the variance of ϵ will not be known, but a finite set of observations may be available, for which w values are known and ξ has been observed. The usual analysis of variance provides a significance test for the non-vanishing of β, which is equivalent to testing for the significance of the regression of ξ on w.

It is to be noted that this analysis reduces to the conventional between-within analysis (F or t-test) when we have the special case of two populations. Moreover, if we had treated ξ as the fixed variate instead of w, and considered the regression of w on ξ, the Analysis of Variance would have differed only in replacing $\Sigma(\xi - \bar{\xi})^2$ throughout by $\Sigma(w - \bar{w})^2$ and the relevant F-test would have been unchanged.

When probabilities of misclassification are estimated from finite samples, as in the soil classification example, there are three sources of error, sampling error in the estimate of the separation value ξ_0, sampling error in the estimate of the distance between the population means, and sampling error in the estimated standard deviation of ξ within populations. It does not appear difficult to set up confidence intervals for the probability of misclassification, assuming repeated classification of individuals given fixed initial samples.

2. The one-dimensional discriminant function. We have been dealing so far with the simple situation in which only one measurement per individual is available for purposes of discrimination. Suppose we still have this measurement, call it ξ_1, now, but we have other measurements as well, say $\xi_2, \cdots, \xi_p$. As before $\xi_1 = \alpha_1 + \beta w + \epsilon_1$. For the moment suppose that the remaining measurements have mean values independent of w, so that

$$\xi_m = \alpha_m + \epsilon_m, \qquad (m = 2, \cdots, p),$$

and let us assume also that the $\{\epsilon_m\}$ are mutually independent, $(m = 1, 2, \cdots, p)$ and are normal deviates with unit variance. It is safe to assume that nobody would ever argue, in this case, that the measurements $\xi_2, \cdots, \xi_p$, provide information about the w value for an individual. If, then, we were so fortunate that we were in this situation, and knew so, we could say that ξ_1 is our discriminant function, since, if any discriminating is to be done, ξ_1 has to do it.

TABLE 1

Analysis of Variance for Regression

	d.f.	Sums of Squares
Regression	1	$r^2\Sigma(\xi - \bar{\xi})^2$
Error	$N - 2$	$(1 - r^2)\Sigma(\xi - \bar{\xi})^2$
Total	$N - 1$	$\Sigma(\xi - \bar{\xi})^2$

$$r = \frac{\Sigma(\xi - \bar{\xi})(w - \bar{w})}{\sqrt{\Sigma(\xi - \bar{\xi})^2\Sigma(w - \bar{w})^2}}$$

Suppose, now that the measurements $\xi_1, \xi_2, \cdots, \xi_p$ are not explicitly available, but that we are able to observe a linearly equivalent set $x_1, x_2, \cdots, x_p$, related to the $\{\xi_m\}$ by the transformation

$$x_m = \sum_{n=1}^{p} l_{mn}\xi_n$$

where the l_{mn} are unknown. For fixed w, x_m has expected value

$$\sum_{n=1}^{p} l_{mn}\alpha_n + l_{m1}\beta w = a_m + b_m w,$$

so that in general each x_m observation provides information about w. Moreover, the x_m are not in general mutually independent; it is evident that the population matrix of variances and covariances for fixed w is given by $\sigma_{mn} = \sum_{k=1}^{p} l_{mk}l_{nk}$.

As an example of a set of correlated measurements, consider the Azotobacter example referred to above. In addition to pH values, determinations of available phosphate content and total nitrogen content were made on soil samples in each of the two populations. Means were as follows:

	pH	*Phosphate*	*Nitrogen*
Mean of 100 samples with Azotobacter	7.423	133.120	29.400
Mean of 186 samples without ”	6.015	51.113	21.140
Mean difference	1.408	82.007	8.260

Clearly the differences are proportional to the hypothetical b_m's. The variance-covariance matrix, estimated from the 284 degrees of freedom within populations, is given by Table 2.

TABLE 2

		pH	Phosphate	Nitrogen
	pH	111.0879	2,292.7192	198.4026
$284(\sigma_{mn})$ =	Phosphate		1,042,799.1890	5,066.2645
	Nitrogen			29,422.3655

Estimated correlation coefficients within populations are not large, .213 for *p*H and Phosphate, .110 for *p*H and Nitrogen, and .029 for Phosphate and Nitrogen.

Another example is furnished by Fisher's Iris measurements [8], providing sepal length, sepal width, petal length, and petal width for each of 50 individuals of Iris setosa and 50 individuals of Iris versicolor. This example is an unfortunate one in that either petal length or petal width alone is sufficient to discriminate the two populations as completely as anybody has a right to expect anytime. The petal lengths, for example, vary between 1.0 and 1.9 cm. for the 50 setosa, and between 3.0 and 5.1 cm. for the 50 versicolor.

Let us proceed, under the assumption that available measurements, x_m, are distributed normally about mean values $a_m + b_m w$, with variance covariance matrix σ_{mn} for fixed w, keeping in mind the underlying model of $\xi_1, \xi_2, \cdots, \xi_p$, with

$$x_m = \sum_{n=1}^{p} l_{mn}\xi_n, \qquad \xi_1 = \alpha_1 + \beta w + \epsilon_1; \qquad \xi_2 = \alpha_2 + \epsilon_2; \cdots; \xi_p = \alpha_p + \epsilon_p.$$

The skeptic may wish to grant the first part of our assumptions without granting the hypothetical structure of ξ's underlying the x's. Hotelling's work [27] shows that such an underlying structure of ξ's may always be provided, given the distribution of x's for fixed w. In other words, a distribution of x's for fixed w leads essentially uniquely to an underlying ξ model.

The discriminant function, given σ_{mn}, a_m and b_m, for $m, n, = 1, 2, \cdots, p$, is

$$X = \sum_{m,n=1}^{p} \sigma^{mn} b_m x_n = \sum_{n=1}^{p} t_n x_n$$

where

$$t_n = \sum_{m=1}^{p} \sigma^{mn} b_m, \quad \text{and} \quad \sigma^{mn}$$

is the reciprocal matrix to σ_{mn}. That is σ^{mn} are the solutions of the linear systems [17]

$$\sum_{x=1}^{p} \sigma^{ms}\sigma_{sn} = 0 \quad \text{if} \quad m = n; \qquad m, n, = 1, 2, \cdots, p$$

$$\sum_{s=1}^{p} \sigma^{ms}\sigma_{sm} = 1; \qquad m = 1, \cdots, p.$$

That X, as defined above, is properly called the discriminant function will become evident immediately. Putting $b_m = l_{m1}\beta$, $x_n = \sum_{k=1}^{p} l_{nk}\xi_k$, we have

$$X = \beta \sum_{m,n,k,} \sigma^{mn} l_{m1} l_{nk} \xi_k .$$

Recalling that the σ^{mn} are reciprocal to $\sigma_{mn} = \sum_{k} l_{mk} l_{nk}$, it can be seen that $\sum_{mn} \sigma^{mn} l_{m1} l_{nk} = 1$ if $k = 1$, and vanishes for $k = 1$. It follows that

$$X = \beta\, \xi_1 ,$$

in other words, X calculated as $\sum_{mn} \sigma^{mn} b_m x_n$ from known population quantities is proportional to the hypothetical ξ_1, the only one of the underlying measurements which is related to w, thus justifying the term discriminant function for X. It is clear that any other linear function of the x's is also a linear function of the ξ's, and can discriminate, at best, only as well as X itself, since all the ξ's are independent of w, with the exception of ξ_1. X itself discriminates w to the same extent that ξ_1, were it available, would discriminate.

The degree of discrimination of w's depends, as indicated in the previous section, on the ratio of the mean square of ξ_1, among w's (mean square for regression), to the mean square of ξ_1 for fixed w (mean square for error). Since X is proportional to ξ_1, the same is true when X is substituted for ξ_1. It turns out, of course, that X is that linear combination of x's for which the ratio of the mean square for regression to the mean square for error is a maximum, or, what is the same thing, X is that linear combination of X's which has the maximum correlation with w. From any point of view X appears to be the logical function of x's to compute. It is clear that λX is precisely as good as X, if λ is any constant.

In the two population case, where w takes on only two values, X is evidently proportional to $\Sigma\sigma^{mn}(\mu_{m1} - \mu_{m2})x_n$, where μ_{m1} and μ_{m2} are the mean values of x_m in the two populations. X is here the particular linear combination of x's for which the ratio of the mean square between populations to the mean square within populations is a maximum. The value of this ratio, which measures the degree of discrimination possible, depends on the spread of the means of X between the populations, or in general, on the spread of the means of X over some given distribution of w's. Given σ_{mn} and b_m the larger the spread of w values the better overall discrimination will be obtainable. On the other hand, the coefficients for X depend only on σ_{mn} and b_m.

Since X is proportional to ξ_1, it follows that the discriminant function is invariant under non-singular linear transformation of the x's, that is, if some set of y's, linearly dependent on the x's, had been observed, together with their means, variances and covariances, the discriminant values would not have changed. This invariance is obviously a desirable property, and as such was one of the goals of Fisher, Hotelling, and Mahalanobis. One more property of the discriminant function is of interest; X is essentially equivalent to the maximum likelihood estimate of w.

In our statistical model w plays the role of a fixed variate or population parameter, and the x's have a joint distribution about linear functions of w as means. Suppose now that (σ_{mn}) and $\{b_m\}$ are estimated from an analysis of variance and covariance on data for which w as well as x values are known. The problem of estimating w for a single individual whose x measurements are given resolves into a two-stage estimation process, the first stage being the estimation of (σ_{mn}) and $\{b_m\}$ from the initial data, the second stage being the estimation of w by the discriminant function whose coefficients are computed from the estimated (σ_{mn}) and $\{b_m\}$. It has already been pointed out that X is the linear combination of x's which has greatest correlation with w. It turns out, then, that the coefficients of X are proportional to those which would have been obtained from a formal regression analysis of w on $x_1, x_2, \cdots, x_p$, considering the x's as independent variables and w as dependent variable, a direct interchange of roles as compared with the statistical model we have assumed. Of course two linear functions differing only by a factor of proportionality are equivalent in discrimination. If the formal analysis of variance is carried out for testing the significance of the regression of w on $x_1, x_2, \cdots, x_p$, the relevant F ratio remains a valid test for the non-vanishing of the b_m in spite of the inversion of dependent and independent variables. The analysis of variance is given in Table 3.

R is, of course, the conventional multiple correlation coefficient. An equivalent analysis can be carried out for X itself, allowing sufficient degrees of freedom for the estimation of the constants in X, as given in Table 4.

This analysis is proportional to the analysis given above. It might be noted that the mean square corresponding to error sum of squares in this analysis is $\Sigma\sigma^{mn}b_mb_n$, which is X evaluated for $x_n = b_n$, $(n = 1,2, \cdots ,p)$.

In the Azotobacter example, Cox and Martin arrive at a discriminant function which has the analysis given in Table 5.

It is evident that the difference between populations is highly significant. The choice of scale for X in this case forces the sum of squares within populations to be equal to the difference between the mean X values for the two populations. Thus the mean X differs by .021777 for the two populations, and has an estimated standard error, within populations, equal to $\sqrt{.00007722} = .008788$. Half the difference, divided by the standard error is the normal deviate corresponding to misclassification, if equal risks are taken. In this case the value of the normal deviate is 1.24, approximately, leading to an estimated probability of misclassification of about .11, which is not very much better than the .13 which one would have obtained if pH alone had been used.

TABLE 3

Analysis of Variance for Regression

	d.f.	Sums of Squares
Regression	p	$R^2\Sigma(w - \bar{w})^2$
Error	$N - p - 1$	$(1 - R^2)\Sigma(w - \bar{w})^2$
Total	$N - 1$	$\Sigma(w - \bar{w})^2$

TABLE 4

Analysis of Variance for X on w

	d.f.	Sums of Squares
Regression	p	$R^2\Sigma(X - \bar{X})^2$
Error	$N - p - 1$	$(1 - R^2)\Sigma(X - \bar{X})^2$
Total	$N - 1$	$\Sigma(X - \bar{X})^2$

TABLE 5

Analysis of Variance of Discriminant Function

	d.f.	Sums of Squares	Mean Square
Between populations	3	.030842	.01028
Within populations	282	.021777	.00007722
Total	285		

In this problem, as in conventional regression analysis, it is tempting, for various reasons, to consider the possibility of using smaller sets of classifying measurements. Moreover, a significance test for this situation is in general more interesting, as a practical matter, than the significance test for differences among populations, since the initial presumption is that we are interested in being able to discriminate, on the basis of $x_1, x_2, \cdots, x_p$. Suppose, for example, we wish to test whether the discriminant function $X_{(p)}$ based on $x_1, x_2, \cdots, x_p$ is significantly better than the discriminant function $X_{(r)}$ based on $x_1, \cdots, x_r$, with $r < p$. The relevant test is precisely the same as the test calculated formally from the regression of w on the sets $x_1, \cdots, x_r$ and $x_1, x_2, \cdots, x_p$, with the analysis of variance given in Table 6.

TABLE 6

Analysis of Variance for Rejecting $x_{r+1}, \cdots, x_p$

		Sums of Squares	d.f.
S_r^2	Regression on	$x_1, \cdots, x_r$	r
S_p^2	Regression on	$x_1, \cdots, x_r, x_{r+1}, \cdots, x_p$	p
$S_p^2 - S^2$	Difference		$p - r$
$S_T^2 - S_p^2$	Error		$N - p - 1$
S_T^2	Total		$N - 1$

TABLE 7

Analysis of Variance for $X = X_0$

	Sums of Squares	d.f.
S_p^2	Regression on X_0	1
S_p^2	Regression on $x_1, \cdots, x_p$	p
$S_p^2 - S_1^2$	Difference	$p - 1$
$S_T^2 - S_p^2$	Error	$N - p - 1$
S_T^2	Total	$N - 1$

Similarly, if we wish to test for the significance of a theoretical discriminant function, X_0, with preassigned coefficients, as compared with X_p, we have again the conventional test calculated from the formal analysis of the regression of w on $x_1, x_2, \cdots, x_p$, as given in Table 7.

As shown by Fisher [21] the relevant F-Test for this hypothesis is computable as

$$F_{p-1,n-p+1} = \frac{n - p + 1}{p - 1} \frac{R'^2}{1 - R'^2}$$

where $R'^2 = R^2(1 - r^2)$, r is the correlation between X and X_0 for fixed w, and R is the multiple correlation for w on $x_1, \cdots, x_p$, or, what is the same thing, the correlation of w and X.

The example of Smith [36] is an example in which the relationships of x's to w have to be estimated from analysis of variance and covariance of data in which the w's are not really known, being related to genotypes. The regression of x's on w is estimated by a generalization of the components-of-variance method, from variance-covariance analyses in which the usual null hypotheses are significantly contradicted. The net effect is that the usual significance tests now fail to hold, although the algebraic calculations are formally equivalent to those given above, once the population relations of x's to w are established. When work of this kind is based on small samples, there is some difficulty in estimating the reliability of the results.

3. Multi-dimensional discriminant functions. Instead of trying to discriminate between two populations or estimate a single parameter w, our problem may be to discriminate among several populations, not necessarily linearly related, or to estimate many independent parameters w_1, $w_2, \cdots, w_s$. Just as a single parameter w is sufficient to distinguish between means of measurements for two different populations, s parameters are sufficient to distinguish between means of $s + 1$ different populations, and exactly s parameters will be required, if no linear relation obtains among the $s + 1$ populations. For example, with three populations, any measurement mean may be given the three possible values α, $\alpha + \beta$, $\alpha + \gamma$, corresponding to $w_1 = w_2 = 0$ for population 1, $w_1 = 1$, $w_2 = 0$ for population 2, and $w_1 = 0$, $w_2 = 1$ for population 3. Geometrically we have to consider a set of parameter values as a point in an s-dimensional space.

The one-dimensional discriminant function admits two very different generalizations in higher dimensions. The practical solution to a particular problem for which s is moderately large may involve a mixture of both generalizations.

Let us generalize our statistical model before discussing the discrimination problem. To avoid complication of algebraic notation, let us for the moment assume $s = 2$. We will now postulate a set of hypothetical measurements $\xi_1, \xi_2, \cdots, \xi_p$, with

$$\begin{aligned}
\xi_1 &= \alpha_1 + \beta_1 u + \gamma_1 v + \epsilon_1 \\
\xi_2 &= \alpha_2 + \beta_2 u + \gamma_2 v + \epsilon_2 \\
\xi_3 &= \alpha_3 + \epsilon_3 \\
&\cdot \\
&\cdot \\
&\cdot \\
\xi_p &= \alpha_p + \epsilon_p,
\end{aligned}$$

where the ϵ_p are independent normal deviates with unit variance, u and v are fixed variates or parameters corresponding to the different populations, and $\alpha_1, \alpha_2, \cdots, \alpha_p, \beta_1, \beta_2, \gamma_1$, and γ_2 are constants. Evidently $\xi_3, \cdots, \xi_p$ can yield no information about u and v; ξ_1 and ξ_2 together contain all the information there is to get about u and v. As before, assume that our data will be in the form of linear combinations $x_m = \Sigma l_{mn}\xi_n$, with unknown coefficients l_{mn}. The variance-covariance matrix within populations, or for fixed u, v, is still given by $\sigma_{mn} = \Sigma l_{mk}l_{nk}$. The mean values of the x's for fixed u, are given by

$$\begin{aligned} E(x_m) &= \Sigma l_{mn}\alpha_n + (l_{m2}\beta_1 + l_{m2}\beta_2)u + (l_{m1}\gamma_1 + l_{m2}\gamma_2)v \\ &= A_m + b_m u + c_m v. \end{aligned}$$

This model is again justifiable on the basis of Hotelling's work.

The first question to ask is whether we can now form two linear combinations of the x's and get rid of $\xi_3, \cdots, \xi_p$ in both, thus providing a two dimensional description of an individual on the basis of $x_1, x_2, \cdots, x_p$. The answer here is in the affirmative, as a result of a direct generalization of the method discussed earlier. If we calculate $X_1 = \Sigma\sigma^{mn}b_m x_n$ and $X_2 = \Sigma\sigma^{mn}c_m x_n$, we are fortunate enough to get

$$\begin{aligned} X_1 &= \beta_1\xi_1 + \beta_2\xi_2 \\ X_2 &= \gamma_1\xi_1 + \gamma_2\xi_2 \end{aligned}$$

with no disturbing elements from $\xi_3, \cdots, \xi_p$. Assuming for now that X_1 and X_2 are not merely proportional, i.e. $\beta_1\gamma_2 - \beta_2\gamma_1 \neq 0$, what do we do with X_1 and X_2?

For fixed u, v, we have

$$\begin{aligned} E(X_1) &= \Sigma\sigma^{mn}b_m a_n + u\Sigma\sigma^{mn}b_m b_n + v\Sigma\sigma^{mn}b_m c_n \\ &= A_1 + B_1 u + C_1 v \\ E(X_2) &= \Sigma\sigma^{mn}C_m a_n + u\Sigma\sigma^{mn}c_m b_n + v\Sigma\sigma^{mn}c_m c_n \\ &= A_2 + B_2 u + C_2 v \end{aligned}$$

and variances and covariance

$$\begin{aligned} \tau_{11} &= \Sigma\sigma^{mn}b_m b_n = B_1 \\ T_{12} &= \Sigma\sigma^{mn}b_m c_n = C_1 = B_2 \\ \tau_{22} &= \Sigma\sigma^{mn}c_m c_n = C_2. \end{aligned}$$

We may for example, estimate u and v by solving the equations

$$\begin{aligned} B_1 u + c_1 v &= X_1 - A_1 \\ B_2 u + C_2 v &= X_2 - A_2, \end{aligned}$$

or we may set up regions in the X_1, X_2 plane for which certain decisions are made. For example, when classifying an individual into one of three populations, we might delineate regions, as in Fig. 2.

Then the particular individual would be classified as coming from population I, II, or III, according to which region X_1, X_2 falls in. The individual points shown in the figure represent the expected values of X_1, X_2 for each of the three populations. No exhaustive investigation has been made for this situation, but some fairly obvious methods are available for constructing such regions.

With respect to significance tests when the σ_{mn}, a_m, b_m, c_m are estimated from samples, the whole gamut of multivariate analysis has to be run. Tests analogous to (but more complicated than) F tests exist for testing the significance of the discrimination, the significance of a subset of the x's, and the significance of a theoretical pair $X_{1,0}$, $X_{2,0}$ (Wilks [41], [42], [43]).

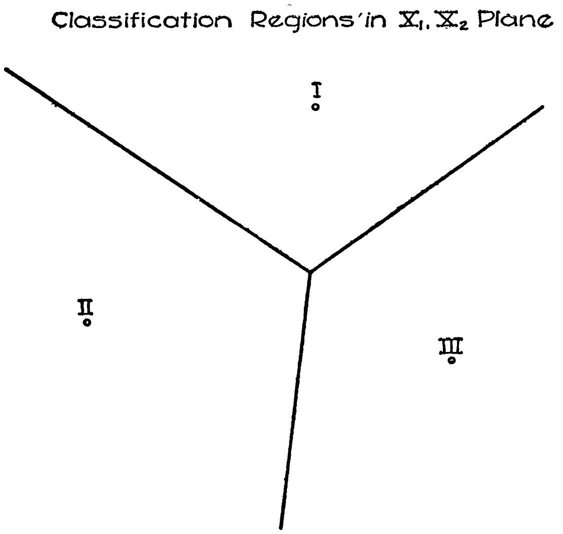

FIG. 2

For some purposes a two-dimensional discrimant function X_1, X_2 may be unsatisfactory. For example, we might suspect that $\beta_1\gamma_2 = \beta_2\gamma_1$ (or that the relationship is nearly satisfied). Under these circumstances X_1 is (nearly) proportional to X_2, and we would like to compute the best one-dimensional discriminant function, even though we have started with two linear parameters u and v. Even if $\beta_1\gamma_2 \neq \beta_2\gamma_1$ we might still ask for the best one-dimensional discriminant function, in order to rank our populations on the "best" linear scale. If we define Y as that linear combination of $x_1, x_2, \cdots, x_p$ which has the largest multiple correlation with u and v, we have generalized the simple one-dimensional discriminant function in a second direction.

Before proceeding, it is useful to recognize that Y, as defined above, must be a function of X_1, X_2, since X_1 and X_2 together contain all the information about u and v that can be obtained from the x's.

Now suppose we consider an arbitrary linear combination $Y = \lambda_1 X_1 + \lambda_2 X_2$. Y correlates best with

$$\lambda_1(\tau_{11}u + \tau_{12}v) + \lambda_2(\tau_{12}u + \tau_{22}v) = (\lambda_1\tau_{11} + \lambda_2\tau_{12})u + (\lambda_2\tau_{12} + \lambda_2\tau_{22})v.$$

We now have to choose λ_1 and λ_2 to maximize this correlation. This correlation will be maximized if we maximize the ratio of the variance of

$$(\lambda_1\tau_{11} + \lambda_2\tau_{12})u + (\lambda_1\tau_{12} + \lambda_2\tau_{22})v$$

(over the distribution of u and v values) to the variance of Y for fixed u and v. Call the first quantity S_1, the second S_2. Then $S_2 = \lambda_1^2\tau_{11} + 2\lambda_1\lambda_2\tau_{12} + \lambda_2^2\tau_{22}$ and S_1 is of the form $\lambda_1^2\mu_{11} + 2\lambda_1\lambda_2\mu_{12} + \lambda_2^2\mu_{22}$ where

$$\begin{aligned} \mu_{11} &= \tau_{11}^2\sigma_{uu} + 2\tau_{11}\tau_{12}\sigma_{uv} + \tau_{12}^2\sigma_{vv} \\ \mu_{12} &= \tau_{11}\tau_{12}\sigma_{uu} + (\tau_{12}^2 + \tau_{11}\tau_{22})\sigma_{uv} + \tau_{12}\tau_{22}\sigma_{vv} \\ \mu_{22} &= \tau_{12}^2\sigma_{uu} + 2\tau_{12}\tau_{22}\sigma_{uv} + \tau_{22}^2\sigma_{vv}. \end{aligned}$$

Maximizing S_1/S_2 leads to the equations:

$$\lambda_1\tau_{11} + \lambda_2\tau_{12} = \frac{S_1}{S_2}(\lambda_1\mu_{11} + \lambda_2\mu_{12})$$

$$\lambda_1\tau_{12} + \lambda_2\tau_{22} = \frac{S_1}{S_2}(\lambda_1\mu_{12} + \lambda_2\mu_{22})$$

i.e.

$$\lambda_1(\tau_{11} - \theta\mu_{12}) + \lambda_2(\tau_{12} - \theta\mu_{12}) = 0$$

$$\lambda_1(\tau_{12} - \theta\mu_{12}) + \lambda_2(\tau_{22} - \theta\mu_{22}) = 0, \qquad \text{with } \theta = S_1/S_2.$$

It is thus seen that θ must satisfy the quadratic equation

$$(\tau_{11} - \theta\mu_{11})(\tau_{22} - \theta\mu_{22}) - (\tau_{12} - \theta\mu_{12})^2 = 0,$$

in order for solutions λ_1, λ_2 to exist. In general there will be two solutions, of which the greater corresponds to that linear combination $\lambda_1 X_1 + \lambda_2 X_2$ which has greatest multiple correlation with u and v, whereas the smaller corresponds to that linear combination which has least multiple correlation with u and v. θ itself corresponds to $R^2/(1 - R^2)$ for the regression of $\lambda_1 X_1 + \lambda_2 X_2$ on u, v.

In the general case with s degrees of freedom corresponding to $w_1, w_2, \cdots, w_s$, there is an s-dimensional discriminant function $(X_1, X_2, \cdots, X_s)$, and a set of s linear combinations for which $R^2/(1 - R^2)$ is stationary with respect to

$$\lambda_1, \cdots, \lambda_s.$$

The s roots (corresponding to an equation of degree s) arranged in decreasing order, permit construction of the best one-dimensional, two-dimensional, $\cdots$, $(s - 1)$-dimensional discriminant functions.

Discussion of the relevant significance tests for these reduced discriminant functions is beyond the scope of this paper. Reference may be made to the work of Hotelling and Fisher.

REFERENCES

[1] M. M. Barnard, "The secular variations of skull characters in four series of Egyptian skulls", *Ann. Eugen.*, Vol. 6 (1935), pp. 352–371.

[2] M. S. Bartlett, "The standard errors of discriminant function coefficients", *Jour. Roy. Stat. Soc.*, Suppl. 6 (1939), pp. 169–173.

[3] M. S. Bartlett, "Statistical significance of cannonical correlations", *Biometrika*, Vol. 32 (1942), pp. 29–37.

[4] W. D. Baten, "The discriminant function applied to spore measurements", *Mich. Acad. of Sci., Arts, and Letters*, Vol. 29 (1943), pp. 3–7.

[5] W. D. Baten and C. C. DeWitt, "Use of the discriminant function in the comparison of proximate coal analyses", *Indust. and Eng. Chem., Anal. Ed.*, Vol. 16 (1944), pp. 32–34.

[6] W. D. Baten and H. M. Hatcher, "Distinguishing method differences by use of discriminant functions", *Jour. of Exp. Ed.*, March, 1944.

[7] W. D. Baten, "The use of discriminant functions in comparing judges' scores concerning potatoes", *Jour. Amer. Stat. Assoc.*, Vol. 40 (1945), pp. 223–227.

[8] R. C. Bose, "On the exact distribution of D^2 statistic", *Sankhya*, Vol. 2 (1936), pp 143–154.

[9] R. C. Bose and S. N. Roy, "The exact distribution of the Studentized D^2statistic", *Sankhya*, Vol. 4 (1938) pp. 19–38.

[10] G. W. Brier, R. G. Schoot, and V. L. Simmons, "The discriminant function applied to quality rating in sheep", *Proc. Amer. Soc. An. Prod.*, Vol. 1 (1940), pp. 153–160.

[11] W. G. Cochran, "The comparison of different scales of measurement for experimental results", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 205–216.

[12] G. M. Cox and W. P. Martin, "Use of a discriminant function for differentiating soils with different Azotobacter populations", *Iowa State Col. Jour. of Sci.*, Vol. 11 (1937), pp. 323–331.

[13] B. B. Day and M. M. Sandomire, "Use of the discriminant function for more than two groups", *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942), pp. 461–472.

[14] W. E. Deming, "On the chi test and curve fitting", *Jour. Amer. Stat. Assoc.*, Vol. 29 (1934), pp. 372–382.

[15] D. Durand, "Risk elements in consumer installment financing", *Nat. Bur. Econ. Res., Inc.*, Studies in Consumer Installment Financing, No. 8, 1941.

[16] R. A. Fisher, "The general sampling distribution of the multiple correlation coefficient", *Proc. Roy. Soc. A.*, Vol. 121 (1928), pp. 654–673.

[17] R. A. Fisher, "*Statistical Methods for Research Workers*", Oliver and Boyd, Section 29.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Ann. Eugen.*, Vol. 7 (1936), pp. 179–188.

[19] R. A. Fisher, "Statistical utilization of multiple measurements", *Ann. Eugen.*, Vol. 8 (1938), pp. 376–386.

[20] R. A. Fisher, "The sampling distribution of some statistics obtained from non-linear equations", *Ann. Eugen.*, Vol. 9 (1939), pp. 238–249.

[21] R. A. Fisher, "The precision of discriminant functions", *Ann. Eugen.*, Vol. 10 (1940), pp. 422–429.

[22] H. E. Garrett, "The discriminant function and its use in psychology", *Psychometrika*, Vol. 8 (1943), pp. 65–79.

[23] M. A. Girshick, "On the sampling theory of the roots of determinantal equations", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 203–224.

[24] T. Haalvelmo, "Probability applications in econometrics", *Econometrica*, Suppl. Vol. 12.(1944).

[25] H. Hotelling, "Generalization of Student's ratio", *Annals of Math. Stat.* ,Vol. 2 (1931), pp. 360–378.

[26] H. Hotelling, "The most predictable onterion", *Jour. Educ. Psychol.*, Vol. 26 (1935), pp. 139–142.

[27] H. Hotelling, "Relations between two sets of variates", *Biometrika*. Vol. 28 (1936), pp. 321–377.

[28] P. L. Hsu, "On the distribution of roots of certain determinantal equations", *Ann. Eugen.*, Vol. 9 (1939), pp. 250–258.

[29] W. G. Madow, "Contributions to the theory of multivariate statistical analyses", *Trans. Am. Math. Soc.*, Vol. 44 (1938), pp. 454–495.

[30] P. C. Mahalanobis, "Analysis of race mixture in Bengal", *Jour. Asiat. Soc. Beng.*, Vol. 23 (1927), pp. 301–333.

[31] P. C. Mahalanobis, "On tests and measures of group divergence", *Jour. Asiat. Soc. Beng.*, Vol. 26 (1930), pp. 541–588.

[32] P. C. Mahalanobis, "On the generalized distance in statistics", *Proc. Mat. Inst. Sci. Ind.*, Vol. 12 (1936), pp. 49–55.

[33] E. A. Martin, "A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing", *Biometrika*, Vol. 28 (1936), pp. 149–178.

[34] K. Pearson, "On the coefficient of racial likeness", *Biometrika*, Vol. 18 (1926) ,pp. 105–117.

[35] C. R. Rao, "Tests with discriminant functions in multivariate analysis", *Sankhya*, Vol. 7 (1946), pp. 407–414.

[36] H. F. Smith, "A discriminant function for plant selection", *Ann. Eugen.*, Vol. 7 (1936), pp. 240–250.

[37] W. L. Stevens, "The standardization of rubber flexing tests", *India Rubber World*, August 1, 1940.

[38] A. Wald, "On a statisical problem arising in the classification of an individual into one of two groups", *Annals of Math. Stat.*, Vol. 14 (1944), pp. 145–162.

[39] N. Wallace and R. M. W. Travers, "A psychometric sociologia study of a group of specialty salesmen", *Ann. Eugen.*, Vol. 8 (1938), pp. 266–302.

[40] F. V. Waugh, "Regression between sets of variables", *Econometrica*, Vol. 10 (1942), pp. 290–310.

[41] S. S. Wilks, "Certain generalizations in the analysis of variance", *Biometrika*, Vol. 24 (1932), pp. 471–494.

[42] S. S. Wilks, "On the independence of k sets of normally distributed statistical variables", *Econometrica*, Vol. 3 (1935), pp. 309–326.

[43] S. S. Wilks, *Mathematical Statistics*. Princeton Univ. Press, 1943.