Appendix A

A Note on the Theory of Discriminant Functions

VIEWED in the abstract, the present problem of statistical analysis is one of differentiating two species by means of a set of measurements; it is analogous to some of the problems of biology in which two varieties of plants or other organisms are differentiated on the basis of length of leaf, breadth of stem, etc. The two species under consideration in this study are the good and bad loans of consumer instalment lending, or rather the borrowers who repay their loans and those who fail to repay. This twofold classification, as we have pointed out, is somewhat artificial because loans or borrowers vary considerably in quality; but the distinction is useful and, roughly speaking, reasonably valid. The set of measurements includes information concerning borrower's income, occupation, sex, stability of residence, and the like. Again, to speak of measuring characteristics such as occupation, which is classified qualitatively and not quantitatively, may not be strictly correct, but in a broad sense the concept is satisfactory.

Statistical theorists have given considerable attention to the problem of differentiating two species by a set of measurements, and they have advanced the method of discriminant functions to solve it. This method permits an investigator to weight several credit factors according to their relative importance, and to allow for interrelationships between factors, which are extremely hard to account for by other approaches. A brief discussion of the theory underlying the method will be useful background for the study of good- and bad-loan samples.

Unfortunately, discriminant functions are usually determined on the rather restrictive assumptions that each species considered has the multivariate normal distribution, and that the two species differ only in the average values of the measurements or variates—in other words, that the standard deviation of the variates and the coefficients of correlation between them are the same for each species. These conditions are not met in the good- and bad-loan samples; hence the method in question is not strictly applicable. Nevertheless, for illustrative purposes, its value is sufficiently great to warrant detailed attention.

The problem of differentiating two species by a set of measurements may be introduced by a discussion of the one-variate case. Assume the two species are normally distributed with respect to the distinguishing criterion. Each distribution has variance σ^2; but the means are different—say $+\frac{\alpha}{2}$ and $-\frac{\alpha}{2}$, so that the difference between them is α. The two species then have the probability distributions

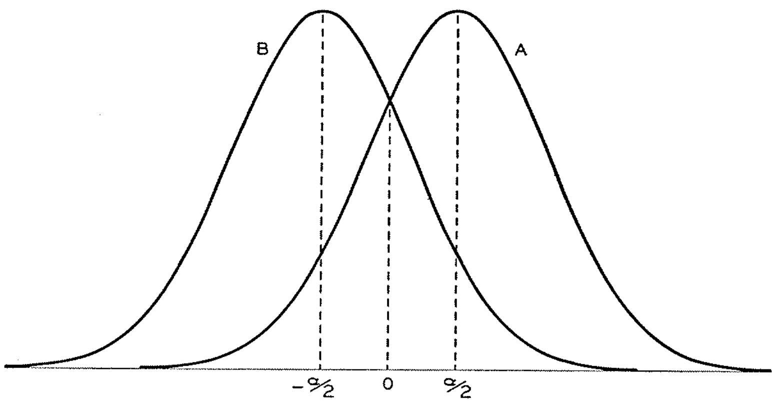
$$P(A) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{\left(x-\frac{\alpha}{2}\right)^2}{2\sigma^2}}\, dx, \qquad P(B) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{\left(x+\frac{\alpha}{2}\right)^2}{2\sigma^2}}\, dx.$$

If species A and species B are equally numerous, the distributions may be represented by two congruent curves, as in Figure 1. To make the example concrete, imagine that A represents good loans, that B represents bad loans, and that the distinguishing criterion is number of years at present occupation.

The ratio $\frac{P(A)}{P(B)} = e^{\frac{\alpha x}{\sigma^2}}$ is the ratio of the relative frequency of A's to B's in a small region around x. The ratio is an increasing function, approaching 0 as x approaches negative infinity, and approaching positive infinity as x approaches positive infinity. When x equals 0, the ratio equals one, indicating that in this region A's and B's are equally numerous. Because the ratio is an increasing function, all regions to the right of 0 contain more A's than B's, and conversely for all regions to the left of 0.

If species A and species B are to be differentiated on the basis of the value of x, several schemes are possible. One common scheme is to use the point 0, the midpoint between the means, as a criterion; values greater than 0 are classified as probably belonging to group A, and vice versa. Under this scheme the probability of misclassifying either an A or a B, P(Mis), is the ratio of the area of the portion of the A curve left of 0 to the total A area, which is the same as the area of the B portion right of 0 to the total B area. P(Mis) is therefore equal to one-half the probability that the absolute value of a normal variate will exceed the absolute value of the ratio $\alpha/2\sigma$. The ratio α/σ, or v, will be used in the future as a measure of the effectiveness of a criterion as a means of differentiating the two species. P(Mis) = $\frac{1}{2}$ when v is 0; it decreases as v becomes larger, approaching 0 as v becomes infinite. The quantity 1 − P(Mis), the probability of classifying correctly, varies from $\frac{1}{2}$ to 1 as v varies from 0 to infinity. Earlier in this study we have used the quantity 1 − 2P(Mis), which we have called the

Figure 1



efficiency index, to measure the effectiveness of the variate x as a means of distinction. This index, which varies from 0 to 1, can be expressed in terms of the ratio v by the following integral:

$$\text{Index} = \frac{1}{\sqrt{2\pi}} \int_{-v/2}^{v/2} e^{-\frac{t^2}{2}} \, dt$$

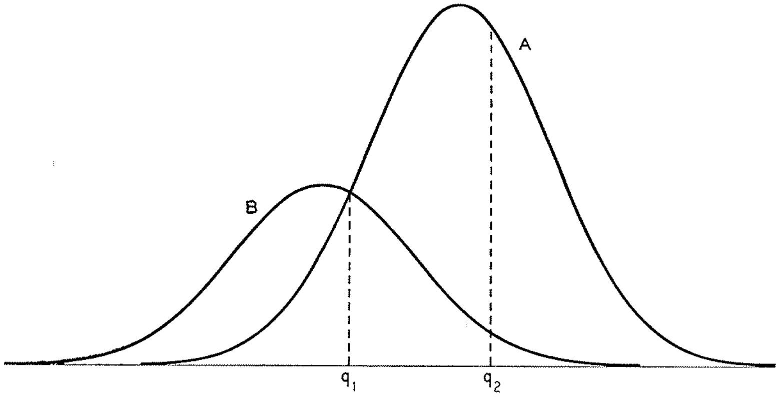
Equally numerous species differentiated by the midpoint between the means are a special case of a much more general situation. In credit analysis the generalization is desirable, for the special case is far from realistic. Good loans and bad loans are not equally numerous. If the ratio of good to bad—i.e., A to B—is k, then the relative frequency ratio

$$\frac{P(A)}{P(B)} = ke^{\frac{\alpha x}{\sigma^2}}$$

is no longer equal to unity when x is zero; it is equal to unity at some other point q_1, which depends on k, α, and σ. But the point q_1, where $\frac{P(A)}{P(B)}$ is unity, is not a satisfactory point of demarcation because the net loss on a bad loan is likely to be considerably greater than the net profit on a good loan; the suitable point, q_2, is determined by equating $\frac{P(A)}{P(B)}$ to the ratio of the average profit on good loans to the average loss on bad loans. In risk selection, two points of demarcation, $q_2'' > q_2'$, may be required in place of only one. For example, applicants to the right of q_2'' could be accepted unconditionally; applicants to the left of q_2' could be rejected unconditionally; and those between q_2'' and q_2' could be given a more rigorous investigation and be required to furnish additional collateral.

For the general case, the concept of the probability of misclassification is substantially altered. Instead of one simple quantity, there are now four as follows: (I) the probability that species A will be misclassified; (II) the probability that species B will be misclassified; (III) the probability that an observation with a value of x greater than the critical value (q_2) will be misclassified; (IV) the probability that an observation with a value less than q_2 will be misclassified. In Figure 2, (I) is represented by the fraction of curve A to the left of the critical value q_2; (II) by the fraction of B to the right of q_2; (III) by the ratio of the tail of B (to the right of q_2) to the sum of the tails of A and B; and (IV) by the ratio of the main portion of A (to the left of q_2) to the sum of the main portion of A and the main portion of B.

Figure 2



In practice, all these values can be determined from tables of the normal curve. These four quantities are not entirely independent; they can be reduced to two quantities. For example,

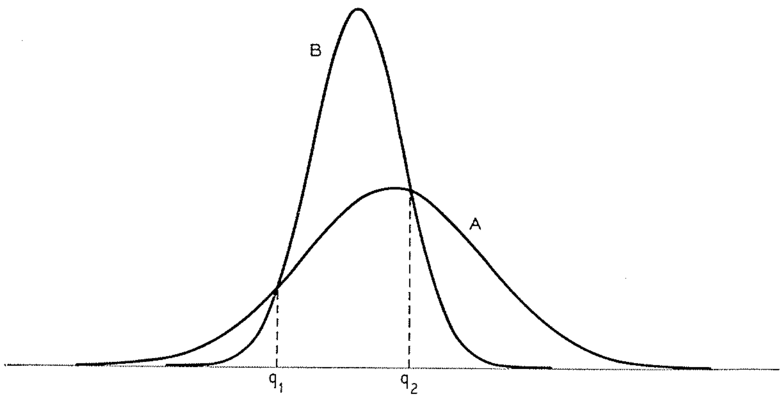
$$(\text{III}) = \frac{(\text{II})}{K[1 - (\text{I})] + (\text{II})}$$

$$(\text{IV}) = \frac{K(\text{I})}{1 - (\text{II}) + K(\text{I})},$$

where K is the ratio of A's to B's. In the special case, where the two species are equally numerous and where 0 is the point of demarcation, P(Mis) = (I) = (II) = (III) = (IV).

A new set of complications is introduced when the two species have different variances as well as different means. The situation is illustrated in Figure 3, where the A variance is larger than the B variance. For the case of equal variances, the logarithm of the ratio $\frac{P(A)}{P(B)}$ $\left(\text{equal to } \frac{\alpha x}{\sigma^2}\right)$ is the equation of an upward sloping straight line through the origin; all values are possible from negative infinity to positive infinity. This means that the ratio of A's to B's can be increased indefinitely by taking a region to the right of a sufficiently large value of x, and conversely. With unequal variances, however, the situation is entirely changed. The logarithm of the probability ratio represents a second degree parabola. In general, the relative frequency ratio is unity at two points, q_1 and q_2. In all regions between these two points, B's are preponderant, but the ratio of B's to A's is everywhere bounded. In the two external regions, the A's are preponderant, and the ratio of A's to B's can be increased indefinitely by taking sufficiently large or sufficiently small values of x.

Figure 3



When several variates or criteria are available for differentiating the two species, the one dimensional case, already discussed, can be generalized. The appropriate method is by means of discriminant functions, which have been developed by R. A. Fisher and a few other writers.[1] Fisher's discriminant function is a linear function of n-variables,

$$Z = l_1x_1 + l_2x_2 + \ldots\ldots + l_px_p$$

where the x's represent the p criteria available for differentiation. This function has a mean for the A species of $\overline{Z}_A = \Sigma l_i\overline{x}_i$ where $\overline{x}_i$ is the mean of the i^{th} variate for the A species; the function has a similar mean $\overline{Z}_B$ for the B species, and a pooled variance (based on both species) of $s_z{}^2 = \Sigma\Sigma l_il_js_{ij}$ where the s_{ij}'s are the pooled variances and covariances of the x's. Here the means, the variances, and the covariances refer to some specific sample. The problem is to determine the coefficients l_i so that the ratio $U^2 = \frac{(\overline{Z}_A - \overline{Z}_B)^2}{s_z{}^2}$ will be maximized. This is accomplished by solving the following set of equations for the l's:[2]

$$\begin{aligned} s_{11}l_1 + s_{12}l_2 + \ldots\ldots + s_{1p}l_p &= a_1 \\ s_{21}l_1 + s_{22}l_2 + \ldots\ldots + s_{2p}l_p &= a_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ s_{p1}l_1 + s_{p2}l_2 + \ldots\ldots + s_{pp}l_p &= a_p \end{aligned} \tag{1}$$

In these equations a_i is the mean difference $\overline{x}_i - \overline{x}_i{}'$, and $s_{ij} = \frac{1}{n + n'} [\Sigma(x_i - \overline{x}_i)\ (x_j - \overline{x}_j) + \Sigma(x_{i'} - \overline{x}_{i'})\ (x_{i'} - \overline{x}_{j'})]$, where n is the number of degrees of freedom in one sample and n′ is the number in the other sample. The solution is

$$l_i = \sum_j \frac{a_js^{ij}}{|s_{ij}|},$$

[1] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, part 2 (September 1936) pp. 179–88; and "The Statistical Utilization of Multiple Measurements," *ibid.*, vol. 8, part 4 (August 1938) pp. 376–86.

[2] Fisher presents these equations in terms of the actual sums S_{ij} instead of the covariances s_{ij}; the result is to multiply the l's by a constant.

where $|s_{ij}|$ is the determinant of the s_{ij}'s and s^{ij} is the cofactor of s_{ij} in that determinant.

A somewhat different approach, which yields the same results with the proper assumptions, is to investigate the relative frequency of species A to species B in various regions of the p-dimensional variate space. Assume two multivariate normal distributions

$$P(A) = Ce^{-1/2\,\Sigma\Sigma Q_{ij}\left(x_i-\frac{\alpha_i}{2}\right)\left(x_j-\frac{\alpha_j}{2}\right)}\,dx_1\;.\;.\;.\;dx_p \qquad (2)$$

$$P(B) = Ce^{-1/2\,\Sigma\Sigma Q_{ij}\left(x_i+\frac{\alpha_i}{2}\right)\left(x_j+\frac{\alpha_j}{2}\right)}\,dx_1\;.\;.\;.\;dx_p,$$

which are identical except for the mean values of the variates. The Q_{ij}'s and the α_i's are supposed to be true population parameters and not sample estimates. In this particular form, which entails no loss of generality, α_i is the difference between the i-mean of the A's and the i-mean of the B's, and 0 is the midpoint between those means; but other forms in which the midpoints are not 0 are sometimes convenient. The ratio $\frac{P(A)}{P(B)}$ has the form $e^{\Sigma\Sigma Q_{ij}x_i\alpha_j}$, which may also be written $e^{\Sigma x_i\lambda_i}$ where $\lambda_i = \sum_j Q_{ij}\alpha_j$.

The equation $\frac{P(A)}{P(B)} = e^{\Sigma x_i\lambda_i} = K$ is the locus of all points in the vicinity of which the ratio of A's to B's is K. This can be transformed into

$$\Sigma x_i\lambda_i = \log_e K,$$

which is the equation of a hyperplane. In particular $\Sigma x_i\lambda_i = 0$ is the equation of a hyperplane through the origin, which is the locus of all points in whose vicinity A's and B's are equally numerous. Since the matrix of Q_{ij} is the inverse of that of σ_{ij}, the covariances of the x's,

$$\lambda_i = \sum_j \alpha_j \frac{\sigma^{ij}}{|\sigma_{ij}|}$$

This is the same as the solution of (1) if $s_{ij} = \sigma_{ij}$ and $a_j = \alpha_j$.

The function $Z = \Sigma x_i\lambda_i$ provides a unique means of differentiating the two species. According to (2), the function Z is normally distributed with variance $\sigma_z{}^2 = \Sigma\Sigma\lambda_i\lambda_j\sigma_{ij}$; it has a mean for the A species of $\bar{Z}_A = \sum_{i=1}^{p}\frac{\lambda_i\alpha_i}{2}$ and for the B species o $\bar{Z}_B = -\sum_{i=1}^{p}\frac{\lambda_i\alpha_i}{2}$, where $\frac{\alpha_i}{2}$ is the A-mean of x_i and $-\frac{\alpha_i}{2}$ is the B-mean. The function Z therefore transforms the multivariate problem into a one-variate problem exactly analogous to that considered earlier.

If A and B are equally numerous, all regions for which Z is greater than 0, which is the midpoint between $\bar{Z}_A$ and $\bar{Z}_B$, contain a preponderance of A's, and conversely. If A's are K times as numerous as B's, and if some adjustment must be made to equate the average loss on bad loans to the average profit on good loans, then an alternative point of demarcation Z_q can be determined.

In the one-variate case with normal distributions and equal variances, the ratio υ was advanced as a measure of the effectiveness of the variate as a differentiator. Two other measures, the probability of misclassification and the efficiency index, were also introduced, but for the case in point these measures depend only on υ and are merely supplementary to it. For the multivariate case, the ratio Υ is exactly analogous to υ in the one variate case; it serves as a measure of the effectiveness of the discriminant function as a differentiator. The probability of misclassification and the efficiency index for a discriminant function are determined by Υ just as they were determined by υ for one variate. It is interesting to note that U, the sample estimate of Υ, is related to Hotelling's generalized T^2 and to the D^2-statistic of Bose and Roy by the following:

$$U = T\sqrt{\frac{n+n'+2}{(n+1)(n'+1)}} = \sqrt{pD^2}$$ [3]

[3] By definition $U = \frac{\Sigma l_i a_i}{\sqrt{\Sigma\Sigma l_i l_j s_{ij}}}$. The numerator of this fraction can be rewritten $\frac{\Sigma\Sigma a_i a_j s^{ij}}{|s_{ij}|}$ since $a_i l_i = a_i \sum_j a_j \frac{s^{ij}}{|s_{ij}|}$; moreover, the quadratic form in the denomi-

(where n + 1 is the number of cases in one sample; n′ + 1 is the number in the other samples; and p is the number of variates).

The ratio Υ cannot be smaller than any of the individual ratios v, and in general it will be larger. It may be considerably or only slightly larger; and if it is only slightly larger, the necessary labor of computing the discriminant function may be hardly worthwhile. Consideration of the conditions that make for a larger ratio and those that make for a small one is therefore pertinent.

In general, the computation of the discriminant function and of the ratio Υ is a difficult task, which grows more difficult as the number of variates increases; but for the special case of complete independence of variates, the computation is almost simple. For the case of complete independence $\sigma_{ij} = 0$ except when $i = j$; therefore, $\lambda_i = \frac{\alpha_i}{\sigma_i^2}$. This means that the discriminant function can be computed as soon as the α's and σ's are known. The ratio Υ, equal to

$$\frac{\Sigma\lambda_i\alpha_i}{\sqrt{\Sigma\Sigma\lambda_i\lambda_j\sigma_{ij}}},$$

simplifies to

$$\frac{\Sigma\frac{\alpha_i^2}{\sigma_i^2}}{\sqrt{\Sigma\frac{\alpha_i^2}{\sigma_i^2}}}$$

and thence to

$$\sqrt{\Sigma\frac{\alpha_i^2}{\sigma_i^2}},$$

nator, $l_il_js_{ij}$, is equal to its inverse, $\Sigma\Sigma a_ia_j\frac{s^{ij}}{|s_{ij}|}$, for the same reason (Cf. Bôcher, *Introduction to Higher Algebra*, 1936, p. 160). Therefore, $U = \sqrt{\frac{\Sigma\Sigma a_ia_js^{ij}}{|s_{ij}|}}$. Since $T^2 = \frac{\Sigma\Sigma a_ia_js^{ij}}{|s_{ij}|} \cdot \frac{n' + n + 2}{(n' + 1)(n + 1)}$, and since $D^2 = \frac{1}{p}\Sigma\Sigma a_ia_j\frac{s^{ij}}{|s_{ij}|}$ (cf. Appendix C, pp. 146, 148, 150-51) the relation of U to T^2 and D^2 follows easily.

which will be written hereafter $\sqrt{\Sigma_i v_i^2}$. This also is extremely easy to compute when the ratios $\frac{\alpha_i}{\sigma_i} = v_i$ are known.

It would be very convenient if the expression $\sqrt{\Sigma v_i^2}$ could be used as a first approximation for the true value of Υ. One might be able to predict whether the actual computation of a discriminant function would be justified by the results obtainable. The following pertinent relation has been worked out for the case of two variates; but a simple generalization for more than two variates appears to be impossible.

The true ratio Υ is equal to $\sqrt{v_1^2 + v_2^2}$ at two points, $\rho = 0$ and $\frac{2}{\frac{v_1}{v_2} + \frac{v_2}{v_1}}$ (where ρ is the correlation coefficient between x_1 and x_2). The ratio reaches a minimum value of v_1 or v_2, whichever is larger, at the point $\rho = \frac{v_1}{v_2}$ or $\frac{v_2}{v_1}$, whichever is less than one in absolute value. Naturally the minimum point lies between 0 and $\frac{2}{\frac{v_1}{v_2} + \frac{v_2}{v_1}}$. On either side of the minimum point, the ratio increases steadily, approaching infinity as ρ approaches ±1.[4]

[4] For two variates $\Upsilon = \left[\frac{\alpha_1^2\sigma_{22} - 2\alpha_1\alpha_2\sigma_{12} + \alpha_2^2\sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\right]^{\frac{1}{2}}$ (see footnote 3). Dividing both numerator and denominator by $\sigma_{11}\sigma_{22}$, and writing $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, $v_1 = \alpha_1/\sqrt{\sigma_{11}}$, $v_2 = \alpha_2/\sqrt{\sigma_{22}}$, we get

$$\Upsilon = \left[\frac{v_1^2 - 2v_1v_2\rho + v_2^2}{1 - \rho^2}\right]^{\frac{1}{2}}$$

When $|\rho|$ approaches unity, Υ becomes infinite except in two special cases: when $v_1 = v_2$ and ρ approaches one, or when $v_1 = -v_2$ and ρ approaches minus one, then $|\Upsilon|$ approaches $|v_1| = |v_2|$. The derivative of Υ^2 with respect to ρ, which is

$$\frac{2\rho(v_1^2 + v_2^2) - 2v_1v_2(1 + \rho^2)}{(1 - \rho^2)^2}$$

is equal to zero at the point v_1/v_2 or v_2/v_1, whichever is less than one in absolute value. At this point Υ has a minimum value of v_1 or v_2, whichever is larger.

We now inquire: At what values of ρ will $\Upsilon = \sqrt{v_1^2 + v_2^2}$? We get $v_1^2 - 2v_1v_2\rho + v_2^2 = (1 - \rho^2)(v_1^2 + v_2^2)$, whence $\rho = 0$ or $\frac{2}{v_1/v_2 + v_2/v_1}$.

There are, then, four different types of cases, which are illustrated in Figure 4. To make the example concrete, imagine that A represents good loans, that B represents bad loans, and the two correlated criteria for differentiation are number of years at present address and number of years at present occupation. In the first two of these (4a and 4b), the true ratio is higher than $\sqrt{\upsilon_1^2 + \upsilon_2^2}$; in the second two it may be higher or lower depending on the value of ρ.

A few concrete applications of this theory may be in order. Suppose that for stability of occupation $\upsilon = .5$, which corresponds to an efficiency index of about 20; and that for stability of residence $\upsilon = .4$, which corresponds to an efficiency index of 16. (These are approximately the efficiency indices actually obtained in the commercial bank samples.) If there is no correlation between stability of residence and stability of employment, the ratio Υ will be .64, which corresponds to an efficiency index of 25. But actually a positive correlation is to be expected. The situation is like that of Figure 4c below; if the correlation lies between 0 and $.976 = \dfrac{2}{\frac{.4}{.5} + \frac{.5}{.4}}$, the actual ratio will be less than .64. Furthermore, since the actual correlation is very likely to lie between 0 and .976, it is a fairly safe prediction that Υ will actually be less than .64. In the commercial bank samples an estimate of the correlation between stability of residence and stability of occupation was made from a small number of cases. The result, .15, was well within the limits of 0 and .976. (See Table B-3, p. 132.)

In the commercial bank samples no appreciable difference was found between the good- and bad-loan samples in connection with either borrower's income or amount borrowed. What then can be inferred about the ratio of amount borrowed to income? Under the assumptions of normality and of equal standard deviations and correlation coefficients, two definite conclusions are possible: (1) as a means of differentiating good and bad loans, the ratio of the amount of the loan to borrower's income, which is just one possible way of combining amount and income, will be inferior to a linear discriminant function; (2) the discriminant function will not show any appreciable difference between good and bad loans. Under the assumed conditions, an independent study of the amount/income ratio, or any other combination of income and amount, would not be warranted. Actually, the distribution of loans according to the amount/income ratio was determined, and the results were negative.

Conclusions such as the above rest on the assumption of normality and the equality of standard deviations and correlation coefficients. Since these assumed conditions do not exist in the loan samples, any of the foregoing conclusions may be invalid. Situations that will upset almost any conclusions based on the theory of this chapter are easily invented. No standardized procedure can be worked out for handling such cases, for each one presents its own problem. A few examples will be shown.

Although a linear discriminant function is entirely appropriate for multivariate normal distributions with equal variances and covariances, it is not so appropriate for most other forms of distributions. For example, when the logarithms of the variates are distributed normally with equal variances and covariances, the appropriate discriminant function has the form

$$Z = \lambda_1 \log x_1 + \lambda_2 \log x_2 + \ . \ . \ . \ ,$$

for which we may conveniently substitute

$$Z' = e^z = x_1^{\lambda_1} x_2^{\lambda_2} \ . \ . \ . \ .$$

A very interesting case occurs when there are only two variates. If $\lambda_1 = \pm\lambda_2$, as will be the case when $\alpha_1(\sigma_{22} \pm \sigma_{12}) = \alpha_2(\sigma_{12} \pm \sigma_{11})$, then the appropriate discriminant function will be x_1x_2 or $\dfrac{x_1}{x_2}$.

When the distributions are normal but with the variances and covariances of A unequal to those of B, the appropriate

Figure 4a

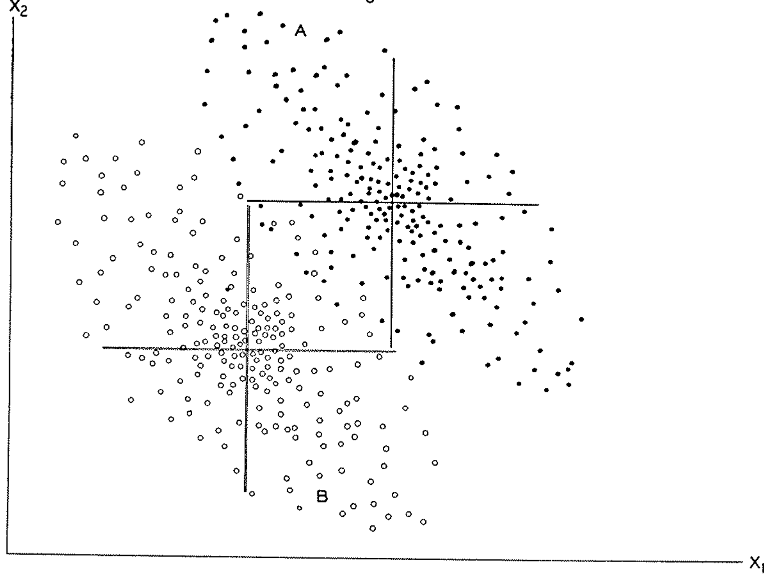


Figure 4b

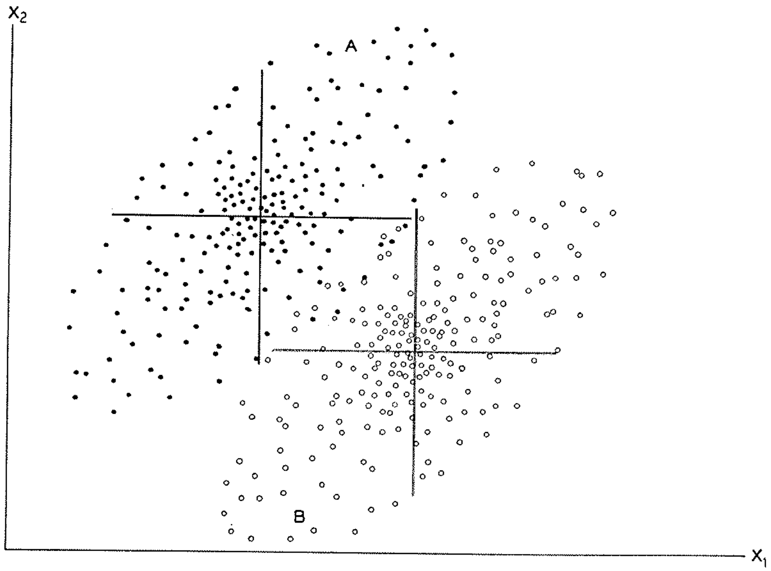


Figure 4c

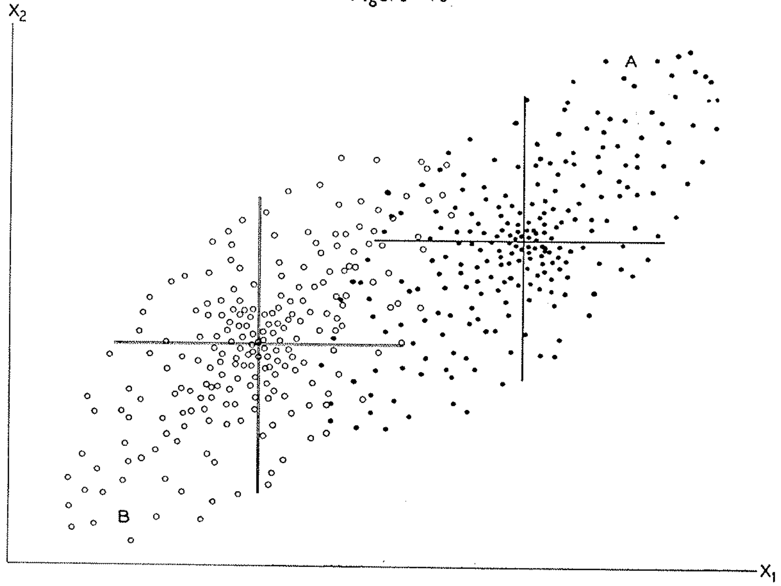
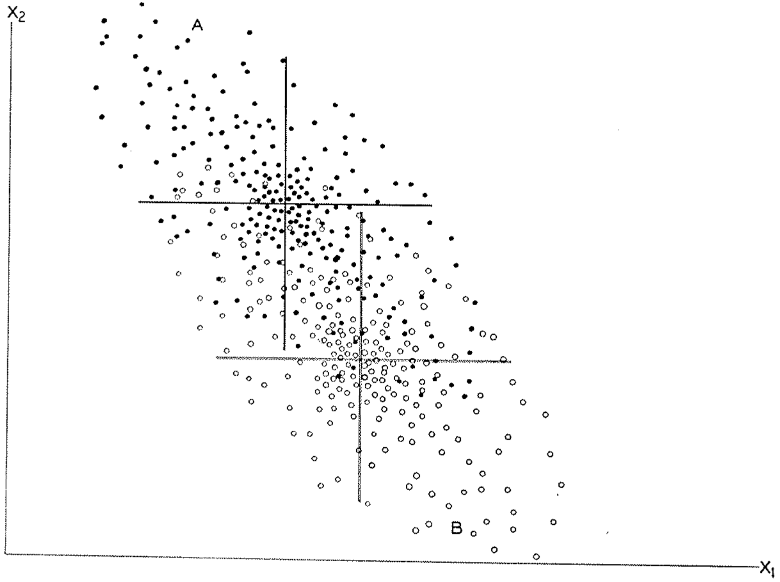


Figure 4d



discriminant function is a general second degree function. We have

$$\frac{P(A)}{P(B)} = \frac{C_A}{C_B} \frac{e^{-\Sigma\Sigma A_{ij}(x_i-\alpha_1)\,(x_j-\alpha_j)}}{e^{-\Sigma\Sigma B_{ij}(x_i-\beta_i)\,(x_j-\beta_j)}} =$$

$$\frac{C_A}{C_B}\, e^{-\Sigma\Sigma[(A_{ij}-B_{ij})x_ix_j-2x_i(\alpha_jA_{ij}-\beta_jB_{ij})+A_{ij}\alpha_i\alpha_j-B_{ij}\beta_i\beta_j]},$$

which indicates a discriminant function of the form

$$\Sigma\Sigma\lambda_{ij}x_ix_j + \Sigma\lambda_ix_j.$$

Such a function will not be normally distributed.

It is even conceivable that the means of the sample may be equal and that the only differences may be in the variances or covariances. A single example is cited by way of illustration. Assume only two variables, and assume the distributions are given by

$$P(A) = Ce^{-\frac{1}{2(1-\rho^2)}\,(x_1{}^2-2\rho x_1x_2+x_2{}^2)}dx_1dx_2$$

$$P(B) = Ce^{-\frac{1}{2(1-\rho^2)}\,(x_1{}^2+2\rho x_1x_2+x_2{}^2)}dx_1dx_2;$$

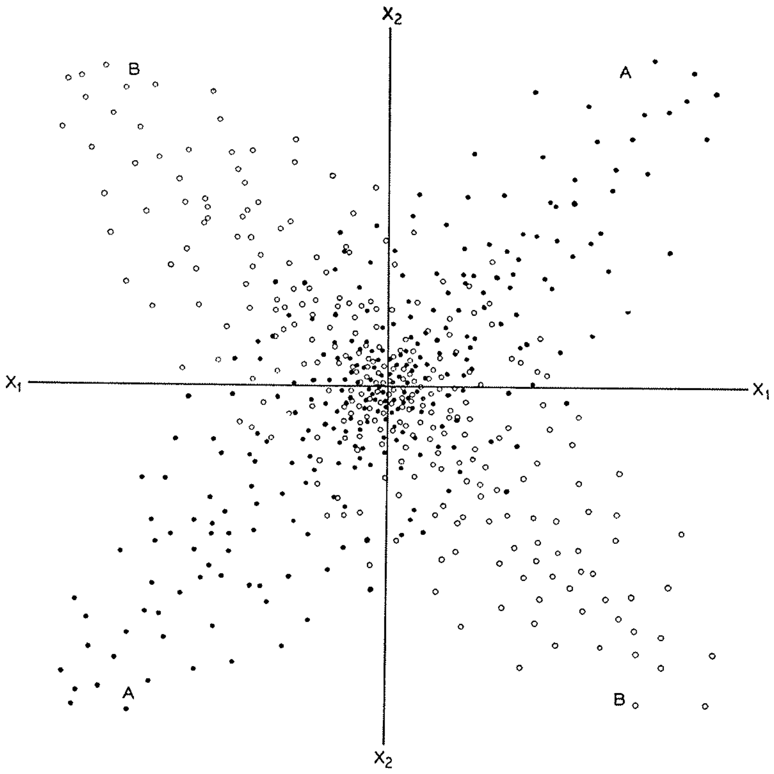
in other words, the means are equal; the variances are both unity; and the correlation coefficients are equal in absolute magnitude but opposite in sign, the A correlation being positive. (See Figure 5.) The probability ratio is

$$K = e^{\frac{2\rho x_1x_2}{(1-\rho^2)}}, \text{ whence } \log K = \frac{2\rho x_1x_2}{(1-\rho^2)}.$$

When K is greater (less) than one, the above equation represents a pair of hyperbolas lying in the lower right (left) and upper left (right) quadrants; and as K approaches ±1, the hyperbolas approximate the coordinate axis. Thus, when A's and B's are equally numerous, all regions in the upper right and lower left quadrants contain a preponderance of A's.

Enough examples have been presented to show that for departures from ideal conditions a linear discriminant function is less appropriate than some other form, the precise nature of which depends on the nature of the distribution. For special cases like the above, the task of determining the appropriate function would not be unduly onerous; but for more general cases the task would be next to impossible. Most practical investigators will probably prefer to determine a linear function, even when the ideal conditions do not exist; and in many instances the resulting approximations will probably be satisfactory.

Figure 5


X2
B
A
X1
X1
A
B
X2