DISCRIMINANT FUNCTIONS WITH COVARIANCE

By W. G. Cochran and C. I. Bliss

North Carolina State College; Connecticut Agricultural Experiment Station and Yale University

1. Summary. This paper discusses the extension of the discriminant function to the case where certain variates (called the covariance variates) are known to have the same means in all populations. Although such variates have no discriminating power by themselves, they may still be utilized in the discriminant function.

The first step is to adjust the discriminators by means of their 'within-sample' regressions on the covariance variates. The discriminant function is then calculated in the usual way from these adjusted variates. The standard tests of significance for the discriminant function (e.g. Hotelling's T^2 test) can be extended to this case without difficulty. A measure is suggested of the gain in information due to covariance and the computations are illustrated by a numerical example. The discussion is confined to the case where only a single function of the population means is being investigated.

2. Introduction. Discriminant function analysis is now fairly well advanced for the case where there are only two populations. The data consist of a number of measurements, called the *discriminators*, that have been made on each member of a random sample from each population. The technique has various uses. Fisher [1] used it in seeking a linear function of the measurements that could be employed to classify new observations into one or other of the two populations. He pointed out [2] that a test of significance of the difference between the two samples, developed from his discriminant, was identical with Hotelling's generalization of Student's t test, discovered some years earlier [3]. Mahalanobis' concept of the generalized distance between two populations [4] was also found to be closely related to the discriminant function. In any of these applications—to classification, testing significance, or estimating distance—we may also be interested in considering whether certain of the measurements really contribute anything to the purpose at hand, and helpful tests of significance are available for this purpose.

Recently the authors encountered a problem in which it seemed advisable to combine discriminant function analysis with the analysis of covariance. This case occurs whenever, in addition to the discriminators, there is a measurement whose mean is known to be the same in both populations. Suppose, for example, that the I.Q.'s of each of a sample of students are measured. The sample is then divided *at random* into two groups, each of which subsequently receives a different type of training. Measurements made at the end of the period of training would be potential discriminators, but in the case of the initial I.Q.'s we can clearly assume that there is no difference in the means of the populations corresponding to the two groups.

The initial I.Q. measurements are of course of no use in themselves in studying differences introduced by the training. Nevertheless, if they are correlated with the discriminators, they may serve in some way to 'improve' the discriminant: e.g. to increase the power of Hotelling's T^2 test, or to reduce the number of errors in classification. This paper discusses the problem of utilizing such measurements, which will be called *covariance variates.* The problem is analogous to that which is solved by the analysis of covariance. In covariance, as applied for instance in a controlled experiment, variates that are unaffected by the experimental treatments can be used to provide more accurate estimates of the effects of the treatments or to increase the power of the F test of the differences among the treatment means.

The procedure suggested is as follows. First, the multiple regression is obtained of each discriminator on all the covariance variates. These regressions are calculated from the 'within-sample' sums of squares and products: that is, from the sums of squares and products of deviations of the individual measurements from their sample means. Each discriminator is then replaced by its deviations from the multiple regression, and a new discriminant function is calculated in the usual way from these deviations. The extensions of Hotelling's T^2 and Mahalanobis' distance are both obtained from this discriminant, though a further adjustment factor is needed for tests of significance.

This paper is arranged in three parts. Part I presents a numerical example. The decision to place the example first was taken because most of the actual applications of the discriminant function in the literature appear to have been made by persons relatively unfamiliar with the theory of multivariate analysis. It is hoped that with the aid of the example readers in this class may be able to utilize covariance variates. For the same reason, the calculations have been presented as far as possible in terms of the operations of ordinary multiple regression, rather than in the form in which they first emerge from the theory. Actually, various equivalent methods of calculation are available, and it is not claimed that our method is necessarily the best. A mathematical statistician may prefer to follow the computing methods which come directly from theory (Part II, section 13).

The example is more complex in structure than the two-sample case. The data constitute a two-way classification, in which the row means are nuisance parameters, being of no interest, while only a single linear function of the column means is of interest. It is well known that the ordinary t test can be applied not only to the difference between two sample means, but to any linear function of a number of sample means in data that are quite complex. Discriminant function technique can be extended in the same way, and readers familiar with the analysis of variance should find no great difficulty in making the appropriate extension to such data.

Part II presents the theory. The reader who is primarily interested in theory should read Part II before Part I. Since the approaches used by Mahalanobis, Hotelling and Fisher all converge, we have chosen that of Mahalanobis, mainly because the extension of his techniques to include covariance variates seems straightforward. Maximum likelihood estimation of the generalized distance is presented in full for the two-population case. The frequency distribution of the estimated distance and the extension of the T^2 test are worked out. An attempt is also made to obtain a quantity that will measure what has been gained by the use of covariance.

In order to illustrate how the theory applies with other types of data, the mathematical model is given for the row by column classification that occurs in the example. The major results for this model are indicated, though without proof.

In Part III it is shown that the computational methods used in the example are equivalent to those developed by theory. While this can easily be verified in a particular case, it is not intuitively obvious.

PART I Numerical Example

3. Description. The data form part of an experiment on the assay of insulin of which other parts have been published [5]. Twelve rabbits were used. Each rabbit received in succession four doses of insulin, equally spaced on a log. scale. An interval of eight days or more elapsed between successive doses, and the order in which the doses were given to any rabbit was determined by randomization. Thus the experiment is of the 'randomized blocks' type, where each rabbit constitutes a block and there are 12 blocks with 4 treatments each.

The effect of insulin is usually measured by some function of the blood sugar of the rabbit in periodic bleedings after injection of the insulin. The blood sugar was measured for each rabbit at 1, 2, 3, 4, and 5 hours after injection, and also before injection. In order to simplify the arithmetic, only the initial blood sugar and the blood sugars at 3 and 4 hours after injection will be considered here. These data are shown for the first three rabbits (with totals for all 12 rabbits) in Table I.

Let x_{iwz} be a typical observation of blood sugar, where $i = 3, 4$ stands for the hour after injection, w for the rabbit and d for the dose. The mathematical model to be used is as follows.

$$(1) \qquad x_{iwz} = \mu_i + \rho_{iw} + \gamma_{iz} + \beta_{i0}(x_{0wz} - x_{0\cdot\cdot}) + e_{iwz} .$$

The parameters μ_i, ρ_{iw} and γ_{iz} represent the true mean and the effects of rabbit and log dose respectively. The quantity x_{0wz} is the initial blood sugar for the rabbit w before the test at dose z, while $x_{0\cdot\cdot}$ is the average initial blood sugar over the whole experiment. The blood sugar at i hours has been found experimentally to be correlated with the corresponding initial blood sugar, and the relationship is represented here as a linear regression, with β_{i0} as the regression coefficient. The residuals e_{iwz} are assumed to follow a multivariate (in this case bivariate) normal distribution, with zero means. The covariance between e_{iwz} and e_{jwz}

is taken as $\sigma_{ij\cdot 0}$. The model is the standard one for the ordinary analysis of covariance, except that we have *two* measures of the effect of insulin, x_3 and x_4.

One additional assumption was made. For all post-injection readings, the blood sugar seemed linearly related to the log dose t_z. Since this result has been found in other experiments, we assumed that

$$\gamma_{iz} = \delta_i t_z$$

where δ_i is the regression coefficient of blood sugar on log dose.

4. Object of the analysis. Our object was to find the linear combination of the three blood sugar readings that would measure best the effect of the insulin. Because of the linearity of the regression on log dose, the effect of insulin on each x_i is known completely if the slope δ_i is known. It seems reasonable to choose the linear compound of the x_i's which will give the maximum ratio when its estimated regression on log dose is divided by the estimated standard error of this regression. We now consider how to obtain this maximum. The argument given below is not intended to prove the validity of the method, for which reference should be made to Part II.

TABLE 1

Sample of original data on blood sugar levels in insulin experiment

Rabbit No.	Log dose											
	Initial blood sugar x_0				Three hours x_3				Four hours x_4			
	.32	.47	.62	.77	.32	.47	.62	.77	.32	.47	.62	.77
1	75	94	107	94	95	76	67	56	96	95	115	91
2	91	86	83	93	98	90	77	69	104	87	90	89
3	97	99	90	91	84	76	59	48	93	102	85	90
Total*.........	1065	1074	1121	1070	932	872	731	591	1098	1026	970	847

*12 rabbits.

The true regression of the original blood sugar x_0 on log dose is known to be zero. Hence, it is clear that the variate x_0 is useful only in so far as it enables us to obtain more accurate estimates of δ_3 and δ_4. For this purpose we need to estimate the effect of x_0 upon x_3 and x_4, the blood sugar readings at 3 and 4 hours, independently of dose of insulin or of differences between rabbits. From the standard theory of covariance the best estimate is the regression coefficient $b_{i0} = E_{i0}/E_{00}$, where E denotes a sum of squares or products calculated from the *error* line in the analysis of covariance; that is from the sums of squares and products of deviations of the x_i from the fitted regression on row and column parameters.

The regression of the blood sugar at each hour on the log dose of insulin is calculated from totals adjusted for the regression on x_0. Since the 4 successive log doses ($z = 1, 2, 3, 4$) are spaced equally, they may be replaced in the computation by the coded doses -3, -1, $+1$, and $+3$. If we let T_{iz} be the total blood sugar, summed over 12 rabbits, at the ith hour with dose z, the following result is well known for the analysis of covariance. The best estimate of $\delta_i (i = 3, 4)$ is

$$[(-3T_{i1} - T_{i2} + T_{i3} + 3T_{i4}) - b_{i0}(-3T_{01} - T_{02} + T_{03} + 3T_{04})]/240.$$

The divisor, 240, is $12(3^2 + 1^2 + 1^2 + 3^2)$. The expression may be written

$$\frac{d_i'}{240} = \frac{(d_i - b_{i0} d_0)}{240},$$

where

$$d_i = -3T_{i1} - T_{i2} + T_{i3} + 3T_{i4}.$$

A linear combination is formed from d_3' and d_4', the numerators in the best estimates of δ_3 and δ_4, by means of the coefficients L_3 and L_4. L_3 and L_4 are computed so as to maximize the ratio of

$$d_I = L_3 d_3' + L_4 d_4'$$

to its estimated standard error.

From the definition of d_i', this requires a discriminant of the form

$$I = L_3(x_{3wz} - b_{30}x_{0wz}) + L_4(x_{4wz} - b_{40}x_{0wz}),$$

where each x_{0wz} is measured from its mean.

We require next the estimated standard error of d_I. This depends, in turn upon the variances of d_3' and d_4' and their covariance. As usual in the analysis, of variance we have

$$(5) \qquad V(d_3') = V(d_3) + d_0^2 V(b_{30}) = \sigma_{33\cdot0}\left(240 + \frac{d_0^2}{E_{00}}\right).$$

The residual variance $\sigma_{33.0}$ is estimated from the sums of squares and products in the error row of the analysis of covariance as

$$s_{33.0} = E_{33.0}/n = (E_{33} - E_{30}^2/E_{00})/n,$$

where n is the degrees of freedom in each E_{ii} diminished by one. Similar methods lead to the variance of d_4' and to the covariance of d_3' and d_4'. It follows that the *true* variance of d_i may be written

$$(6) \qquad V(d_I) \; \alpha \; L_3^2\, \sigma_{33.0} + 2L_3L_4\, \sigma_{34.0} + L_4^2\, \sigma_{44.0},$$

where the factor $\left(240 + \frac{d_0^2}{E_{00}}\right)$ in equation (5) is omitted since it does not involve the L's. Similarly, the *estimated* variance of d_I, apart from constant factors, may be written as

$$(7) \qquad L_3^2E_{33.0} + 2L_3L_4E_{34.0} + L_4^2E_{44.0} .$$

The quantity to be maximized is therefore

$$\frac{(L_3 d_3' + L_4 d_4')}{\sqrt{L_3^2 E_{33.0} + 2L_3 L_4 E_{34.0} + L_4^2 E_{44.0}}} .$$

Formally, this is the same type of quantity that is maximized in ordinary analysis with the discriminant function. Differentiation with respect to the L's leads to the equations (after omission of another constant factor)

$$(8) \qquad E_{33.0}L_3 + E_{34.0}L_4 = d_3' , \qquad E_{34.0}L_3 + E_{44.0}L_4 = d_4' .$$

The objective of the computation, therefore, is to obtain discriminant coefficients having the same ratio to each other as L_3 and L_4 in equations (8). As will be shown in the next section, this can be accomplished in practice more conveniently by substituting an alternative set of three simultaneous equations for the two in equations (8).

5. Calculations. The first step is to form the sums of squares and products in the analysis of covariance. With 12 rabbits and 4 doses, the conventional breakdown of each total sum is into components for rabbits (11 d.f.), doses (3 d.f.) and rabbits × doses (33 d.f.). Because of the assumed linear regression on log dose, the sum of squares for doses was further divided into two components. The first (1 d.f.) is the contribution due to this regression. For x_i, the sum of squares due to regression is $d_i^2/240$, or in the case of x_3, $(1164)^2/240$, or 5645. The remaining component, (2 d.f.) is called the *curvature*, since it measures the effect of deviations from the linear regression. The sum of squares for curvature is found by subtraction.

The following points may be noted. (i) For both x_3 and x_1, the F ratio of the curvature mean square to the rabbits × doses mean square will be found to be less than 1, so that the data do not suggest rejection of the hypothesis of a linear regression on log dose. (ii) The F ratios of the regression mean squares to the rabbits × doses mean squares are highly significant, being 57.8 for x_3 and 28.7 for x_4. This indicates, incidentally, that the three-hour reading may be a more responsive measure of the effect of insulin than the four-hour reading. (iii) With x_0, the F ratio does not approach significance for either the regression or the curvature, as is to be expected.

A consequence of the assumption of linear regression on log dose is that the curvature mean squares and products are estimates of the same quantities as the rabbits × doses mean squares and products. Consequently, the lines for curvature and rabbits × doses in Table 2 could be added to give 35 d.f. for the 'error' sums of squares or products, E_{33}, etc. We decided, however, to estimate the error only from the 33 d.f. for rabbits × doses. This was done because it seemed to facilitate a test of the curvature of the final discriminant I. (This test will not be reported here.)

The L's could now be obtained from equations (8). In this case the first equation would contain the terms

$$E_{33.0} = 3223 - (1259)^2/2351; \qquad E_{34.0} = 1200 - (1259)\ (1340)/2351;$$

$$d_3' = d_3 - b_{30}d_0 = -1164 - \left(\frac{1259}{2351}\right)62,$$

leading to the simultaneous equations

$$2548.8\ L_3 + \ 482.4\ L_4 = -1197.2$$

$$482.4\ L_3 + 2373.2\ L_4 = -\ \ 844.3,$$

which give $L_3/L_4 = -.41848/-.27070 = 1.5459$.

TABLE 2

Sums of squares and products

Component	d. f.	x_0^2	x_3^2	x_4^2	x_0x_3	x_0x_4	x_3x_4
Between rabbits	11	886	9376	11165	1952	2477	9206
Between doses	3	168	5806	2810	−247	−98	3981
{Reg. on log dose	1	16	5645	2727	−301	−209	3924
{Curvature	2	152	161	83	54	111	57
Rabbits × doses	33	2351	3223	3137	1259	1340	1200
Total	47	3405	18405	17112	2964	3719	14387

Instead of using these equations, we propose to solve alternatively the set of three equations

$$S_{00}L_0 + S_{03}L_3 + S_{04}L_4 = d_0$$

$$(9) \qquad S_{30}L_0 + S_{33}L_3 + S_{34}L_4 = d_3$$

$$S_{40}L_0 + S_{43}L_3 + S_{44}L_4 = d_4\ ,$$

where each S_{ij} ($i = 0, 3, 4$) is the sum of squares or products formed by adding the *error* line in the analysis of variance to the line for *regression on log dose.* Thus S_{ij} has 34 d.f. The ratio of L_3 to L_4, as found from equations (9), is exactly the same as that found from the original equations (8), as is proved in section 18. Further, the solution of the new equations seems to be more useful for performing tests of significance, as will appear in following sections.

Accordingly, the first step after forming the analysis of variance is to set up the three equations (9).

The equations are solved by means of the inverse matrix. The values of d_i on the right side of the equations are replaced successively by 1, 0, 0 by 0, 1, 0 and by 0, 0, 1 to obtain the three sets of values for L_0, L_3 and L_4. These results are given in the first three columns of Table 4 and are designated as c_{ij}.

The L's follow from the c_{ij} by the usual rule for regressions. For example,

$$L_3 = \{(.003209)(62) + (.227781)(-1164) + (-.199655)(-809)\}\cdot 10^{-3} = -.103417$$

TABLE 3

Equations for determining L_3 and L_4

$$\begin{array}{rcl} 2367L_0 + 958L_3 + 1131L_4 &=& 62 \\ 958L_0 + 8868L_3 + 5124L_4 &=& -1164 \\ 1131L_0 + 5124L_3 + 5864L_4 &=& -809 \end{array}$$

The composite response or discriminant, adjusted for the covariance variate, is now taken as

$$I = L_3\left(x_3 - \frac{E_{30}}{E_{00}}\,x_0\right) + L_4\left(x_4 - \frac{E_{40}}{E_{00}}\,x_0\right)$$

or

$$-.103417\left(x_3 - \frac{1259}{2351}\,x_0\right) - .066883\left(x_4 - \frac{1340}{2351}\,x_0\right)$$

$$= .093503x_0 - .103417x_3 - .066883x_4 .$$

Note that the value of L_0 is not used at this stage and that $L_3/L_4 = 1.546$ agrees with the value found from equations (8).

TABLE 4

Inverse matrix ($\times$ 10^3) and L's

(10^3 c$_{ij}$)			d$_i$	L$_i$
.465408	.003209	−.092568	62	.100008
.003209	.227781	−.199655	−1164	−.103417
−.092568	−.199655	.362846	−809	−.066883

A similar method may be followed when there are more discriminators or more covariance variates. With two covariance variates, x_0 and x_0^1, for instance, the adjusted discriminant would be

$$L_3(x_3 - b_{30}x_0 - b_{30}^1x_0^1) + L_4(x_4 - b_{40}x_0 - b_{40}^1x_0^1)$$

where b_{30}, b_{30}^1 are the partial regression coefficients of x_3 on x_0, x_0^1 respectively, determined from the error line, and similarly for x_4. Further, since any linear function of the column (dose) means may be represented as a regression on some variate t_z, this method may be applied to any linear function of the column means in which we are interested, provided that the mathematical model is appropriate.

6. Test of the regression of the adjusted discriminant on log dose. The numerator of the regression of I on the coded doses is

$$d_I = L_3(d_3 - b_{30}d_0) + L_4(d_4 - b_{40}d_0).$$

Since the regressions of x_3 and x_4 on the coded doses were both significant, it may be confidently expected that the regression of I will also be significant. The test of significance will, however, be given in case it may be useful in other applications. For those who are familiar with multiple regression, the test is perhaps most easily made by means of a device due to Fisher [2].

Construct a dummy variate y_{wz} such that y_{wz} is always equal to t_z, or in our case to the coded doses. That is, y takes the value -3 for all observations at the lowest dose level, and -1, $+1$, and $+3$ respectively for all observations at the successive higher dosage levels. We shall show that equations (9) solved in finding the L's are formally the same as a set of normal equations for the linear regression of y on x_0, x_3, and x_4.

TABLE 5

Analysis of y^2 and yx_i

	d. f.	y^2	yx_i
Rabbits	11	0	0
Doses	3	240	d_i
Regression on log dose	1	240	d_i
Curvature	2	0	0
Rabbits × doses = error	33	0	0
Sum = Error plus reg. on log dose	34	240	d_i
Total	47	240	d_i

The following analysis for y^2 and yx_i may easily be verified.

It will be noted that the sum of products of y and x_i in the sum line is d_i. Further, S_{ij} is the sum of products of x_i and x_j for this line. It follows that the normal equations for the regression of y on the x's, as calculated from the "sum" line, are

$$S_{i0}L_0 + S_{i3}L_3 + S_{i4}L_4 = d_i \qquad (i = 0, 3, 4).$$

These are just the equations solved in obtaining the L's. Consequently, L_3 and L_4 are the partial regression coefficients of y on x_3 and x_4. A test of the null hypothesis that the true values of L_3 and L_4 are both zero can be made by the standard method for multiple regression, as will be shown later from theory. This test is equivalent to a test of the hypothesis that the true value of d_l is zero.

To apply the test, we require three items in the analysis of variance of y. First, the total sum of squares for the Sum line, already seen to be 240 (Table 5). Second, the reduction due to a regression on all variates (covariance variates plus discriminators). By the usual rules for regression, this is (from Table 4)

$$L_0 d_0 + L_3 d_3 + L_4 d_4 = (.100008)(62) + (-.103417)(-1164) + (-.066883)(-809) = 180.69.$$

Finally, we need the reduction due to a regression on the variates that are not being tested, i.e. on the covariance variates alone. From Table 4, the reduction in this case is simply d_0^2/S_{00} or $(62)^2/2367$, or 1.62. The difference, 180.69 − 1.62, represents the reduction due to the regression of y on L_3 and L_4, after fitting x_0. The resulting analysis is given below, the degrees of freedom being apportioned by the usual rules.

TABLE 6

Analysis of variance of dummy variate y

	d. f.	S. S.	M. S.
Reduction to regression on covariance variates	1	1.62	
Additional reduction to regression on discriminators	2	179.07	89.54
Deviations	31	59.31	1.913
Total (from Sum line)	34	240.00	

The F ratio, 89.54/1.913, or 46.80, with 2 and 31 d.f., is used to test the null hypothesis that the adjusted discriminant has no real regression on log dose.

7. Test of particular discriminators. Another useful test is that of the null hypothesis that a particular discriminator, or group of discriminators, contribute nothing to the adjusted discriminant. In other words, this is a test of the null hypothesis that the true values of a subset of the L's are all zero. The test is of interest in the present investigation, since it would be useful to know whether all five hourly readings of the blood sugar are really helpful. As might be expected by analogy with the previous section, the test is made by calculating the additional reduction due to the regression of y on the particular subset of the L's in question.

The test will be illustrated with respect to L_4. One method of making the test is to re-solve the normal equations with L_4 omitted. From this solution the reduction due to a regression of y on x_0 and x_3 alone is obtained. The additional reduction due to a regression on x_4 is found by subtraction from 180.69.

However, the additional reduction can be found directly from the well-known regression theorem that it is equal to L_4^2/c_{44}. The c's have already been found in Table 4. The result is $(.066883)^2/(.000362846)$, or 12.33. This value is tested against the residual error mean square of 1.913, F having 1 and 31 d.f. The contribution is found to be significant.

In fact, by this process a kind of estimated standard error can be attached to each of the L's for the discriminators, using the formula $s\sqrt{c_{ii}}$, where s is the residual root mean square. Thus for L_3, $(-.103417)$, the 'standard error' is $\sqrt{(1.913)(.000227781)}$, or .0209. It should be stressed that at this point the analogy with regression is rather thin. The L's are not normally distributed, nor do the estimated standard errors follow their usual distribution. It is, however, correct that if the true value of L_4 is zero, $L_4/s\sqrt{c_{44}}$ follows the t distribution with 31 d.f. Thus, if omission of some discriminators seems warranted, these t ratios are relevant in deciding which variate to eliminate first. Strictly speaking, the c's should be re-calculated after each elimination before deciding which other discriminators might also be discarded.

TABLE 7

Analysis of variance for regression of y on the discriminators

	d. f.	S. S.	M. S.
Regression	2	159.20	79.60
Deviations	32	80.80	2.525
Total	34	240.00	

8. Estimation of the gain due to covariance. The tests given above enable us to state whether the discriminators contribute significantly, in the statistical sense. It is also of interest to investigate what has been gained by the use of the covariance variates. From the practical point of view, the question: "What is the gain from covariance?" might be re-phrased as: "If x_0 is ignored, how many rabbits must be tested in order to estimate the regression on log dose as accurately as it was estimated with the adjusted discriminant for 12 rabbits?"

The theoretical aspects of the question are discussed in section 16; the calculations are described here. The only new quantity needed is the F ratio for the regression of y on the discriminators alone. This can be obtained by a new solution of the normal equations, this time with the covariance variates omitted. With just one covariance variate, it is quicker to use the fact that the additional reduction to the regression of y on x_0, after fitting x_3 and x_4, is L_0^2/c_{00}, or $(.100008)^2/(.000465408)$ or 21.49. Consequently, the reduction due to a regression of y on x_3 and x_4 alone is 180.69 − 21.49, or 159.20. The F ratio, 79.60/2.525, is 31.52, whereas the F ratio with covariance is 46.80 (from Table 6). The quantity suggested from theory for comparing the two techniques is

$$\frac{(n_2 - 2)F}{n_2} - 1$$

where n_2 is the number of d.f. in the denominator of F. These values are $\{(30 \times 31.52/32) - 1\}$ or 28.55 with no covariance and $\{(29 \times 46.80/31) - 1\}$ or 42.78 with covariance. The relative information is estimated as 42.78/28.55, or 1.50, so that the use of covariance gives 50 per cent more information. In other words about 18 rabbits would be needed if the initial blood sugars were ignored. To a slight extent this estimate favors the covariance analysis, since it ignores the increased accuracy that would accrue from the extra error d.f. if 18 rabbits were used without covariance.

PART II THEORY

9. Notation. The theory will be given first for the two-population case· We suppose that a random sample of size N has been drawn from each population. A typical discriminator is written $x_{iw\alpha}$ and a typical covariance variate $x_{\xi w\alpha}$, where

$$\begin{aligned} i, j &= 1, 2, \cdots p \text{ denote discriminators,} \\ \xi, \eta &= 1, 2, \cdots k \text{ denote covariance variates,} \\ w &= 1, 2 \text{ denotes the population, and} \\ \alpha &= 1, 2, \cdots N \text{ denotes the order within the sample.} \end{aligned}$$

The population mean of $x_{iw\alpha}$ is μ_{iw}, and the corresponding sample mean is $x_{iw\cdot}$. The difference $(\mu_{i2} - \mu_{i1})$ is denoted by δ_i and the corresponding estimated difference $(x_{i2\cdot} - x_{i1\cdot})$ by d_i.

10. Discriminant functions and generalized distance. Since we propose to approach the theory by means of the generalized distance, it may be well to review briefly the relation between the discriminant and the generalized distance. In the ordinary theory (with no covariance variates) it is assumed that the variates $x_{iw\alpha}$ follow a multivariate normal distribution, and that the covariance matrix σ_{ij} between $x_{iw\alpha}$ and $x_{jw\alpha}$ is the same in both populations. The generalized distance of Mahalanobis is defined by

$$p\Delta^2 = \sum_{i,j=1}^{p} \sigma^{ij}\delta_i\delta_j, \qquad \text{where} \qquad (\sigma^{ij}) = (\sigma_{ij})^{-1}. \tag{10}$$

In order to estimate this quantity from the sample, we first calculate the mean within-sample covariance s_{ij}, where

$$s_{ij} = \sum_{w=1}^{2} \sum_{\alpha=1}^{N} (x_{iw\alpha} - x_{iw\cdot})(x_{jw\alpha} - x_{jw\cdot})/2(N - 1), \tag{11}$$

The estimated distance is then taken as

$$pD^2 = \sum_{i,j=1}^{p} s^{ij}\, d_i d_j . \tag{12}$$

Apart from a factor $N/(N - 1)$, this is the maximum likelihood estimate.

In the discriminant function used by Fisher (1), the object is to find a linear function $I_{w\alpha} = \Sigma M_i x_{iw\alpha}$, where the M_i are chosen to maximize the ratio of the sum of squares between samples to that within samples in the analysis of variance of I. This is equivalent to maximizing the ratio of the difference between the two sample means of I to the estimated standard error of this difference. As Fisher showed (2), the M_i (apart from an arbitrary multiplier) are given by

$$M_i = \sum_{j=i}^{p} s^{ij} d_j$$

Consequently, the difference between the two sample means of I, the discriminant function, is

$$\sum_{i=1}^{p} M_i\, d_i = \sum_{i,j=1}^{p} s^{ij}\, d_i d_j .$$

This is exactly the same as pD^2 in equation (12). Thus the discriminant function leads to the estimated distance, and *vice versa.*

11. Extension to the present problem. In our case there are $(p + k)$ variates (p discriminators, k covariance variates) from which to estimate the distance. All variates, $x_{iw\alpha}$ and $x_{\xi w\alpha}$, are assumed to follow a multivariate normal distribution. The covariance matrix, assumed the same in both populations, now has $(p + k)$ rows and columns, and may be denoted by

$$\Lambda = \begin{pmatrix} \sigma_{ij} & \sigma_{i\eta} \\ \sigma_{\xi j} & \sigma_{\xi\eta} \end{pmatrix}. \tag{13}$$

For each of the covariance variates, it is known that the population means $\mu_{\xi 1}$, $\mu_{\xi 2}$ are equal, so that the difference δ_ξ is zero. This is the fact that distinguishes the problem from ordinary discriminant function analysis.

Hence, the generalized distance, as defined from all $(p + k)$ variates contains no contribution from terms in δ_ξ and is given by

$$(p + k)\Delta^2 = \sum_{i,j=1}^{p} \sigma^{ij}_{(p+k)}\, \delta_i\, \delta_j . \tag{14}$$

The matrix $\sigma^{ij}_{(p+k)}$ is that formed by the first p rows and columns of the inverse of Λ. Note that in general this will not be the same as the matrix σ^{ij}, which is the inverse of σ_{ij}.

In the next section we consider the estimation of this quantity from the sample data. By analogy with the previous section, it might be guessed that the estimate would be of the form $\Sigma s^{ij}_{(p+k)}\, d_i d_j$. The maximum likelihood estimate turns out to be of this form, except that instead of d_i we have d'_i, the difference between the two sample means of the deviations of x_i from its 'within-sample' linear regression on the x_ξ.

12. Estimation of the distance. It is known that the generalized distance is invariant under non-singular linear transformations of the variates. For convenience, we replace the $x_{iw\alpha}$ by variates $x'_{iw\alpha}$, where

$$x'_{iw\alpha} = x_{iw\alpha} - \sum_{\xi=1}^{k} \beta_{i\xi}(x_{\xi w\alpha} - \mu_{\xi w}).$$

Thus $x'_{iw\alpha}$ is the deviation of $x_{iw\alpha}$ from its population linear regression on the $x_{\xi w\alpha}$. The population mean of $x'_{iw\alpha}$ is clearly μ_{iw}, and the difference between the two population means is therefore δ_i.

The covariance matrix of the $x'_{iw\alpha}$, $x_{\xi w\alpha}$ may be written

$$\Lambda' = \begin{pmatrix} \sigma_{ij\cdot\xi} & 0 \\ 0 & \sigma_{\xi\eta} \end{pmatrix}, \tag{15}$$

where $\sigma_{ij\cdot\xi}$ denotes the covariance matrix of the deviations of the $x_{iw\alpha}$ from their regressions on the $x_{\xi w\alpha}$. It follows that in terms of the transformed variates the generalized distance is given by

$$(p + k)\Delta^2 = \sum_{i,j=1}^{p} \sigma^{ij\cdot\xi} \delta_i \delta_j \tag{16}$$

where $\sigma^{ij\cdot\xi}$ is the inverse of the $p \times p$ matrix $\sigma_{ij\cdot\xi}$.

The joint distribution of the $2N$ observations on each of the $x'_{iw\alpha}$ and $x_{\xi w\alpha}$ is as follows:

$$(2\pi)^{-N(p+k)} \mid \sigma^{ij\cdot\xi} \mid^{+N} \mid \sigma^{\xi\eta} \mid^{+N} \Pi dx'_{iw\alpha} dx_{\xi w\alpha} \cdot$$
$$\exp\left\{-\frac{1}{2}\left[\sum_{w=1}^{2}\sum_{\alpha=1}^{N}\sum_{i,j=1}^{p} \sigma^{ij\cdot\xi}(x'_{iw\alpha} - \mu_{iw})(x'_{jw\alpha} - \mu_{jw}) + \sum_{w=1}^{2}\sum_{\alpha=1}^{N}\sum_{\xi,\eta=1}^{k} \sigma^{\xi\eta}(x_{\xi w\alpha} - \mu_{\xi w})(x_{\eta w\alpha} - \mu_{\eta w})\right]\right\},$$

where $\sigma^{\xi\eta}$ is the inverse of the $k \times k$ matrix $\sigma_{\xi\eta}$.

We now proceed to estimate Δ^2 in equation (16) by maximum likelihood. For this, we obviously need the sample estimates of the $\sigma^{ij\cdot\xi}$ and the δ_i, and it will appear presently that the sample estimates of the $\beta_{i\xi}$ are also required. However, it happens that the $\sigma^{\xi\eta}$ and the $\mu_{\xi w}$ are not needed. Hence the relevant part of the likelihood function is

$$L = N \log \mid \sigma^{ij\cdot\xi} \mid - \frac{1}{2}\sum_{w=1}^{2}\sum_{\alpha=1}^{N}\sum_{i,j=1}^{p} \sigma^{ij\cdot\xi}(x'_{iw\alpha} - \mu_{iw})(x'_{jw\alpha} - \mu_{jw}) \tag{17}$$

where

$$x'_{iw\alpha} = x_{iw\alpha} - \sum_{\xi=1}^{k} \beta_{i\xi}(x_{\xi w\alpha} - \mu_{\xi w}) \,.$$

Differentiating first with respect to μ_{iw} , we obtain

$$\sum_{\alpha=1}^{N} \sum_{j=1}^{p} \sigma^{ij\cdot\xi}(x'_{jw\alpha} - \hat{\mu}_{jw}) = 0. \tag{18}$$

Except in the case (with probability zero) where our estimate of $\sigma^{ij\cdot\xi}$ turns out to be singular, these equations have no solution except

$$\sum_{\alpha=1}^{N} (x'_{jw\alpha} - \hat{\mu}_{jw}) = 0 \tag{19}$$

for every j, w. Consequently

$$\hat{\mu}_{jw} = x'_{jw\cdot}$$

so that

$$\hat{\delta}_j = \hat{\mu}_{j2} - \hat{\mu}_{j1} = x'_{j2\cdot} - x'_{j1\cdot} = d_j - \sum_{\xi=1}^{k} \beta_{j\xi}\, d_\xi \,.$$

This shows that the $\beta_{i\xi}$ must also be estimated. Now

$$\frac{\partial L}{\partial \beta_{i\xi}} = \sum_{w=1}^{2} \sum_{\alpha=1}^{N} \frac{\partial L}{\partial x'_{iw\alpha}} \frac{\partial x'_{iw\alpha}}{\partial \beta_{i\xi}} = \sum_{w=1}^{2} \sum_{\alpha=1}^{N} \sum_{j=1}^{p} \sigma^{ij\cdot\xi}(x_{\xi w\alpha} - \mu_{\xi w})(x'_{jw\alpha} - \mu_{jw}).$$

Once again, unless the estimate of $\sigma^{ij\cdot\xi}$ is singular, the only solutions of the equations formed by equating this quantity to zero are

$$\sum_{w=1}^{2} \sum_{\alpha=1}^{N} (x_{\xi w\alpha} - \mu_{\xi w})(x'_{jw\alpha} - \hat{\mu}_{jw}) = 0 \tag{20}$$

for every ξ, j.

Since $\hat{\mu}_{jw} = x'_{jw\cdot}$, the term in $\mu_{\xi w}$ vanishes. Substituting for x' in terms of x from (17), we obtain

$$\sum_{w=1}^{2} \sum_{\alpha=1}^{N} x_{\xi w\alpha} \left\{ (x_{jw\alpha} - x_{jw\cdot}) - \sum_{\eta=1}^{k} b_{j\eta}(x_{\eta w\alpha} - x_{\eta w\cdot}) \right\} = 0$$

where $b_{j\eta}$ stands for the maximum likelihood estimate of $\beta_{j\eta}$. These equations may be written

$$\sum_{\eta=1}^{k} b_{j\eta} E_{\xi\eta} = E_{j\xi} \tag{21}$$

where E denotes a sum of squares or products of deviations from the sample means, containing $2(N - 1)$ degrees of freedom. The equations are therefore the ordinary normal equations for the 'within-sample' multiple regression of $x_{jw\alpha}$ on the $x_{\xi w\alpha}$.

Finally, differentiation of L with respect to the $\sigma^{ij\cdot\xi}$ leads to

$$2N\hat{\sigma}_{ij\cdot\xi} = \sum_{w=1}^{2}\sum_{\alpha=1}^{N}(x'_{iw\alpha} - x'_{iw\cdot})(x'_{jw\alpha} - x'_{jw\cdot}). \tag{22}$$

This is just the 'within-samples' sum of squares or products of the variates x'. On substituting for the x' in terms of the x and using equations (21), we obtain

$$2N\hat{\sigma}_{ij\cdot\xi} = E_{ij} - \sum_{\xi=1}^{k} b_{i\xi}E_{j\xi} = E_{ij\cdot\xi} \qquad \text{(say)}.$$

To summarize, the estimated distance is given by means of the equation

$$(p + k)D^2 = \sum_{i,j=1}^{p} \hat{\sigma}^{ij\cdot\xi}\hat{\delta}_i\hat{\delta}_j = 2N\sum_{i,j=1}^{p} E^{ij\cdot\xi}\, d'_i\, d'_j,$$

where $E^{ij\cdot\xi}$ is the inverse of $E_{ij\cdot\xi}$ and

$$d'_i = d_i - \sum_{\xi=1}^{k} b_{i\xi}\, d_\xi .$$

This estimate was obtained by assuming *all* variates jointly normally distributed. From the form of the likelihood function (17) it can be seen that the M.L. estimate of the distance remains the same under the less restrictive assumptions that the $x_{\xi w\alpha}$ are fixed, while the deviations of the $x_{iw\alpha}$ from their regressions on the $x_{\xi w\alpha}$ are jointly normal.

13. Computational procedure. An orderly procedure for calculating the generalized distance will now be given. From this, the method for computing the corresponding discriminant function will be shown. The computations also lead to the generalization of Hotelling's T^2. The steps are as follows.

(i). First form the 'within-sample' sums of squares and products of all variates, with $2(N - 1)$ degrees of freedom. These are the quantities denoted by E_{ij}, $E_{i\xi}$, $E_{\xi\eta}$.

(ii). Invert the matrix $E_{\xi\eta}$, giving $E^{\xi\eta}$.

(iii). The regression coefficients $b_{i\xi}$, estimates of the $\beta_{i\xi}$, are now obtainable by means of the relations

$$b_{i\xi} = \sum_{\eta=1}^{k} E_{i\eta}E^{\xi\eta},$$

as is clear from the usual matrix solution of equations (21).

(iv). The sums of squares and products of the deviations of the x_i from their 'within-sample' regressions on the x_ξ are now computed from equations (22)

$$2N\hat{\sigma}_{ij\cdot\xi} = E_{ij\cdot\xi} = E_{ij} - \sum_{\xi=1}^{k} b_{i\xi}E_{j\xi}.$$

(v). The final step is to invert the matrix $E_{ij\cdot\xi}$, giving $E^{ij\cdot\xi}$, and to form the product

$$(p + k)D^2 = 2N \sum_{i,j=1}^{p} E^{ij\cdot\xi} d'_i d'_j, \quad \text{where} \quad d'_i = d_i - \sum_{\xi=1}^{k} b_{i\xi} d_\xi .$$

When there were no covariance variates, the discriminant function I had the property that the difference between the two sample means of I was equal to the estimated distance (Section 10). This relationship can be preserved when covariance variates are present by defining I so that

$$I_{w\alpha} = \sum_{i=1}^{p} M_i \left(x_{iw\alpha} - \sum_{\xi=1}^{k} b_{i\xi} x_{\xi w\alpha} \right),$$

and calculating the weights M_i from the equations,

$$\sum_{j=1}^{p} E_{ij\cdot\xi} M_j = d'_i .$$

For in that case,

$$M_i = \sum_{j=1}^{p} E^{ij\cdot\xi} d'_j .$$

Consequently the difference between the two sample means of I is

$$\sum_{i=1}^{p} M_i d'_i = \sum_{i,j=1}^{p} E^{ij\cdot\xi} d'_i d'_j ,$$

which (apart from the constant $2N$) is equal to $(p + k)D^2$.

14. Distribution of the estimated distance. In the ordinary case, with no covariance variates, the frequency distribution of the estimated distance has been given by several authors, e.g. Hsu [6]. It will be found that in our problem the distribution is essentially the same, except that the quantity D^2 must be multiplied by a new factor and that one set of degrees of freedom entering into the result must be changed from $(n - p + 1)$ to $(n - p - k + 1)$.

Thus far we have assumed that all variates jointly follow a multivariate normal distribution. It is convenient at this stage to regard the covariance variates $x_{\xi w\alpha}$ as fixed from sample to sample, and to use the conditional distribution of the $x_{iw\alpha}$, subject to this restriction. It is well known (e.g. Cramér [7, section 24.6]) that this conditional distribution is the multivariate normal

$$(2\pi)^{-Np} \,|\, \sigma^{ij\cdot\xi} \,|^{+N} \prod dx_{iw\alpha}$$
$$\exp\left\{ -\tfrac{1}{2} \left[\sum_{w=1}^{2} \sum_{\alpha=1}^{N} \sum_{i,j=1}^{p} \sigma^{ij\cdot\xi} (x_{iw\alpha} - \mu_{iw} - \gamma_{iw\alpha})(x_{jw\alpha} - \mu_{jw} - \gamma_{jw\alpha}) \right] \right\} \tag{23}$$

where

$$\gamma_{iw\alpha} = \sum_{\xi=1}^{k} \beta_{i\xi} (x_{\xi w\alpha} - \mu_{\xi w}).$$

Since the estimated distance is a function of the quantities $E_{ij\cdot\xi}$, d'_i, we now find the joint distribution of these variates. The joint distribution of the sums of squares and products $E_{ij\cdot\xi}$ is obtained by quoting a slight extension of a result due to Bartlett [8], which may be stated as follows.

Let the variates $x_{iw\alpha}$ *follow the distribution* (23) *and let*

$$(i) \qquad E_{ij} = \sum_{w=1}^{2}\sum_{\alpha=1}^{N} (x_{iw\alpha} - x_{iw\cdot})(x_{jw\alpha} - x_{jw\cdot})$$

be a typical 'within-samples' sum of squares or products,

$$(ii) \qquad b_{i\xi} = \sum_{\eta=1}^{k} E_{i\eta}E^{\xi\eta}$$

be the 'within-samples' partial regression coefficient of x_i *on* x_ξ, *and*

$$(iii) \qquad E_{ij\cdot\xi} = E_{ij} - \sum_{\xi=1}^{k} b_{i\xi}E_{j\xi}$$

be the sum of squares or products of deviations from these regressions. Then
(*a*) *the quantities* $E_{ij\cdot\xi}$ *follow the Wishart distribution*

$$c\,|\,E_{ij\cdot\xi}\,|^{\frac{1}{2}(n-k-p-1)} \exp\left\{-\tfrac{1}{2}\sum_{i,j=1}^{p} \sigma^{ij\cdot\xi}E_{ij\cdot\xi}\right\}\prod dE_{ij\cdot\xi}$$

with $(n - k)$ *d.f., where* $n = 2(N - 1)$,

(*b*) *this distribution is independent of that of the* $b_{i\xi}$, *and*

(*c*) *both distributions are independent of that of the means* $x_{iw\cdot}$ *and consequently of that of the difference* $d_i = (x_{i2\cdot} - x_{i1\cdot})$.

The result was proved by Bartlett for a sample from a single population. The extension to the case of two populations is straightforward and will not be given in detail.

From (b) and (c) it follows that the distribution of the $E_{ij\cdot\xi}$ is independent of that of the quantities

$$d'_i = d_i - \sum_{\xi=1}^{k} b_{i\xi}\,d_\xi\,.$$

Further, with the x_ξ variates fixed, the d'_i are linear functions of the $x_{iw\alpha}$ with constant coefficients and hence follow a multivariate normal distribution, Wilks [9]. We now find the means and the covariance matrix of this joint distribution.

From the joint distribution (23) of the $x_{iw\alpha}$, it is easily seen that

$$(24) \qquad E(d_i) = \delta_i + \sum_{\xi=1}^{k} \beta_{i\xi}\,d_\xi\,.$$

Also, since by standard regression theory the $b_{i\xi}$ are unbiased estimates of the $\beta_{i\xi}$,

$$E\left\{\sum_{\xi=1}^{k} b_{i\xi}\,d_\xi\right\} = \sum_{\xi=1}^{k} \beta_{i\xi}\,d_\xi\,.$$

Hence, by subtraction,

$$E(d'_i) = \delta_i . \tag{25}$$

Now

$$\text{Cov}\ (d'_i d'_j) = \text{Cov}\ (d_i - \sum_{\xi=1}^{k} b_{i\xi}\, d_\xi)(d_j - \sum_{\eta=1}^{k} b_{j\eta}\, d_\eta).$$

By (c) the distributions of the d_i, $b_{j\eta}$ are independent, so that there will be no contribution from products of the form $d_i b_{j\eta}$. Hence

$$\text{Cov}\ (d'_i\, d'_j) = \text{Cov}\ (d_i\, d_j) + \sum_{\xi,\eta=1}^{k} d_\xi\, d_\eta\ \text{Cov}\ (b_{i\xi} b_{j\eta}). \tag{26}$$

Since d_i is the difference between the means of two samples of size N, Cov $(d_i d_j)$ is $2\ \sigma_{ij\cdot\xi}/N$. The covariance of $b_{i\xi}$ and $b_{j\eta}$ is more troublesome. Writing the expressions for these regression coefficients in terms of the original data, we have

$$\text{Cov}\ (b_{i\xi} b_{j\eta}) = \sum_{\lambda,\nu=1}^{k} E^{\lambda\xi}E^{\nu\eta}\ \text{Cov}\ (E_{i\lambda} E_{j\nu}) =$$

$$\sum_{\lambda,\nu=1}^{k} E^{\lambda\xi}E^{\nu\eta} \sum_{w,z=1}^{2} \sum_{\alpha,\zeta=1}^{N} (x_{\lambda w\alpha} - x_{\lambda w\cdot})(x_{\nu z\zeta} - x_{\nu z\cdot})\ \text{Cov}\ (x_{iw\alpha}\, x_{jz\zeta}).$$

Since successive observations are assumed independent, the covariance term vanishes unless $w = z$ and $\alpha = \zeta$, in which case it equals $\sigma_{ij\cdot\xi}$. Thus

$$\text{Cov}\ (b_{i\xi} b_{j\eta}) = \sigma_{ij\cdot\xi} \sum_{\lambda,\nu=1}^{k} E^{\lambda\xi}E^{\nu\eta}E_{\lambda\nu} = \sigma_{ij\cdot\xi} E^{\xi\eta}.$$

Finally, from (26)

$$\text{Cov}\ (d'_i\, d'_j) = \sigma_{ij\cdot\xi}\left(\frac{2}{N} + \sum_{\xi,\eta=1}^{k} E^{\xi\eta}\, d_\xi\, d_\eta\right) = v\sigma_{ij\cdot\xi} \qquad \text{(say)}. \tag{27}$$

Having obtained the distributions of the $E_{ij\cdot\xi}$, d'_i, we may apply Hsu's result [6] for the general distribution of Hotelling's T^2. In our notation, this may be stated as follows.

If the variates $d'_i/\sqrt{v}$ follow the multivariate normal distribution with means $\delta_i/\sqrt{v}$ and covariance matrix $\sigma_{ij\cdot\xi}$, and if the variates $E_{ij\cdot\xi}$ follow the Wishart distribution with $(n - k)$ d.f. and covariance matrix $\sigma_{ij\cdot\xi}$, the two distributions being independent, then

$$y = \sum_{i,j=1}^{p} E^{ij\cdot\xi}\, d'_i\, d'_j/v,$$

follows the distribution

$$e^{-\tau} \sum_{h=0}^{\infty} \frac{\tau^h}{h!}\, \frac{1}{B\{\frac{1}{2}p + h, \frac{1}{2}(n - k - p + 1)\}} \cdot y^{\frac{1}{2}p+h-1}(1 + y)^{-\frac{1}{2}(n-k+1)-h}\, dy, \tag{28}$$

where

$$\tau = \tfrac{1}{2} \sum_{i,j=1}^{p} \sigma^{ij\cdot\xi} \delta_i \delta_j / v,$$

$$v = \frac{2}{N} + \sum_{\xi,\eta=1}^{k} E^{\xi\eta} \, d_\xi \, d_\eta \,, \qquad n = 2(N - 1).$$

This distribution is, of course, the distribution of the ratio of two independent values of χ^2, with p and $(n - k - p + 1)$ d.f. respectively, in the case where the numerator is non-central.

15. Tests of significance. This result leads to the extension of Hotelling's T^2 test. For if $\delta_i = 0$, $(i = 1, 2, \cdots p)$, then τ is zero and

$$\sum_{i,j=1}^{p} E^{ij\cdot\xi} \, d'_i \, d'_j$$

is distributed as $vpF/(n - k - p + 1)$, with p and $(n - k - p + 1)$ d.f. The distribution (28) above gives the power function of this test.

We may also wish to apply a test of this type to a subgroup x_i of the discriminators $(i = 1, 2, \cdots q < p)$. Speaking popularly, this is a test of the null hypothesis that the above variates x_i contribute nothing to the discrimination between the two populations, given that the remaining discriminators and the covariance variates have already been included.[1] To see what is meant more precisely, consider the following transformation:

$$\begin{aligned} x'_i &= x_i - \Sigma\, \beta_{il} x_l - \Sigma\, \beta_{i\xi} x_\xi, & i &= 1, 2, \cdots q; \\ x'_l &= x_l - \Sigma\, \beta_{l\xi} x_\xi, & l &= q + 1, \cdots p; \\ x'_\xi &= x_\xi, & \xi &= 1, 2, \cdots k, \end{aligned}$$

where the β's are population regression coefficients. Then it is not difficult to see that the distance is now given by

$$(p + k)\Delta^2 = \sum_{i,j=1}^{q} \sigma^{ij\cdot l\xi} \delta'_i \delta'_j + \sum_{l,m=q+1}^{p} \sigma^{lm\cdot\xi} \delta_l \delta_m$$

where $\sigma^{ij\cdot l\xi}$ is the inverse of the covariance matrix of the deviations of the x_i from their regressions on the x_l plus the x_ξ, and

$$\delta'_i = \delta_i -' \Sigma\, \beta_{il} \delta_l.$$

Consequently if $\delta'_i = 0$, $(i = 1, 2, \cdots q)$ the distance is exactly the same as it would be if the variates x_i were omitted. The test in question is therefore a test of the null hypothesis that $\delta'_i = 0$, $(i = 1, 2, \cdots q)$.

If both the remaining discriminators x_l and the covariance variates x_ξ are regarded as fixed, the method of proof in the previous section provides an F test

[1] The test is illustrated in section 7.

for this hypothesis also. It is found that the sums of squares or products $E^{ij \cdot l\xi}$ follow a Wishart distribution with $(n - k - p + q)$ d.f., while the quantities

$$d_i' = d_i - \sum_{l=q+1}^{p} b_{il}\, d_l - \sum_{\xi=1}^{k} b_{i\xi}\, d_\xi$$

are normally distributed, with zero means when the null hypothesis is true. This leads to the result that

$$\sum_{i,j=1}^{q} E^{ij \cdot l\xi}\, d_i'\, d_j'$$

is distributed as $v'qF/(n - k - p + 1)$, with q and $(n - k - p + 1)$ d.f., and

$$v' = \frac{2}{N} + \Sigma\, E^{l\xi}\, d_l\, d_\xi\,,$$

the sum extending over both the covariance variates and the discriminators that are not being tested.

16. Discussion of the gain due to covariance. In this section we attempt to construct a measure of the amount that has been gained by the use of the covariance variates. Only a preliminary discussion will be given: a complete discussion would be rather lengthy, owing to the many different uses to which the discriminant function can be put. Perhaps the problem can most easily be seen by considering the effect on Hotelling's generalized T^2 test of significance.

The power function of this test, as obtained from equation (28) section 14, depends on four factors; the level of significance that is chosen, the degrees of freedom n_1 and n_2 in the numerator and denominator of F, and the parameter τ. If the covariance variates were ignored, the usual T^2 test could be applied to the discriminators alone. In this case we would have

$$n_1' = p, \qquad n_2' = n - p + 1, \qquad \tau' = \tfrac{1}{2}\Sigma\sigma^{ij}\delta_i\delta_j/v', \qquad \text{where } v' = 2/N.$$

With the covariance variates, we have

$$n_1 = p, \qquad n_2 = n - p - k + 1, \qquad \tau = \tfrac{1}{2}\Sigma\sigma^{ij \cdot \xi}\delta_i\delta_j/v,$$

where

$$v = \frac{2}{N} + \Sigma\, E^{\xi\eta}\, d_\xi\, d_\eta\,.$$

The first point to note is that

$$\Sigma\, \sigma^{ij \cdot \xi}\delta_i\delta_j \geq \Sigma\, \sigma^{ij}\delta_i\delta_j$$

This is an instance of the general result that the addition of new variates cannot decrease the value of $p\Delta^2$. To see this, replace the covariance variates by their deviations from their regressions on the discriminators. This transformation gives

$$(29) \qquad \sum_{i,j=1}^{p} \sigma^{ij \cdot \xi} \delta_i \delta_j = \sum_{i,j=1}^{p} \sigma^{ij} \delta_i \delta_j + \sum_{\xi,\eta=1}^{k} \sigma^{\xi\eta \cdot i} \delta'_\xi \delta'_\eta ,$$

where

$$\delta'_\xi = \delta_\xi - \sum_{i=1}^{p} \beta_{\xi i} \delta_i .$$

Since the term on the right of equation (29) is a positive definite quadratic form, the result follows.

Consequently, the first effect of the covariance variates is to make the numerator of τ greater than that of τ'. As a partial compensation, the denominator v is also greater than v', but it may be shown that the difference in the denominators will usually be trivial if k is small relative to n. We therefore expect τ to be greater than τ'. Now for fixed n_1, n_2 and significance level, it is well known that the power function (28) is monotone increasing with τ. Hence, other things being equal, the increase in τ due to the covariance variates leads to a more powerful test.

The two power functions, however, differ in another respect, in that with covariance the value of n_2 is reduced from $(n - p + 1)$ to $(n - p - k + 1)$. This decrease in the number of degrees of freedom in the denominator of F will to some extent offset the gain from an increased τ. Examination of Tang's tables [10] indicates, however, that if the degrees of freedom are substantial, this effect will not be important. Moreover, in most practical applications, k is likely to be only 1 or 2. Hence, as a first approximation the effect will be ignored, though to do so tends to overestimate the advantage of covariance.

Suppose now that $\tau = r\tau'$, where $r > 1$. Since τ' is proportional to N, the size of sample taken from each population, we could make $\tau' = \tau$ by increasing the size of sample (when covariance is not used) from N to rN. This suggests that the ratio τ/τ' can be used, as a first approximation, to measure the relative accuracy obtained with and without the use of covariance. This measure carries approximately the usual interpretation that the inferior method would become as good as the superior method if the sample size for the inferior method were increased by the factor r. A further refinement could be made to take account of the difference in the n_2 values. By trial and error applied to Tang's tables, one could determine r' so that the two power functions would be as nearly coincident as possible.

In practice, the ratio τ/τ' must be estimated from the data. From the power function in equation (28) it is found by integration that the mean value of y is

$$(2\tau + p)/(n_2 - 2),$$

so that an unbiased estimate of τ is

$$\tfrac{1}{2}\{(n_2 - 2)y - p\} = \tfrac{1}{2}p\left\{\frac{(n_2 - 2)}{n_2} F - 1\right\}.$$

This suggests that the quantity

$$\frac{(n_2 - 2)}{n_2} F - 1$$

should be calculated both with and without covariance. The ratio of the two values will probably not be an unbiased estimate of τ/τ', but may be used pending further information about its sampling distribution. This type of calculation is made for the numerical example in section 8.

17. The case of a row by column classification. Thus far the discussion has been confined to the case where there are only two populations. The technique may also be used when there are more than two populations. The difference δ_i between the two population means is replaced by some linear function of the population means. As an illustration we consider a row by column classification, the case that arises in the numerical example. No detailed proofs will be given, though it is hoped that the theory can be fairly easily developed from the mathematical model.

A typical variate is x_{iwz}, where $i = 1, 2, \cdots p$ denotes the variate, $w = 1, 2, \cdots r$ denotes the row and $z = 1, 2, \cdots c$ denotes the column, there being one observation in each cell. The variates x_{iwz} follow a multivariate normal distribution, with covariance matrix $\sigma_{ij \cdot \xi}$ and means

$$E(x_{iwz}) = \mu_i + \rho_{iw} + \gamma_{iz} + \sum_{\xi=1}^{k} \beta_{i\xi}(x_{\xi wz} - x_{\xi \cdot \cdot}),$$

where ρ_{iw} denotes the effect of the row and γ_{iz} that of the column. Without loss of generality we may assume that

$$\sum_{w} \rho_{iw} = \sum_{z} \gamma_{iz} = 0.$$

In addition, there exists a *known* set of variates t_z such that

$$\gamma_{iz} = \delta_i t_z, \qquad \sum_{z} t_z = 0.$$

That is, the column constants have a linear regression on a set of known numbers.

The following are the maximum likelihood estimates of the relevant constants.

$$b_{i\xi} = \sum_{\eta=1}^{k} E_{i\eta} E^{\eta\xi},$$

where

$$E_{i\eta} = \sum_{w,z} x_{iwz} \left\{ x_{\eta wz} - x_{\eta w \cdot} - t_z \frac{(\sum t_z x_{\eta \cdot z})}{\sum t_z^2} \right\}.$$

$$\hat{\delta}_\iota = \frac{\sum_{w,z} t_z (x_{iwz} - \sum_{\xi} b_{i\xi} x_{\xi \cdot z})}{\sum t_z^2}.$$

In the notation used for numerical calculation,

$$\hat{\delta}_i = \frac{(d_i - \sum b_{i\xi}\, d_\xi)}{r \sum t_z^2} = \frac{d'_i}{r \sum t_z^2}, \qquad \text{where} \quad d_i = \sum_z t_z X_{i \cdot z},$$

the quantity $X_{i \cdot z}$ being the column *total.* Finally

$$rc\hat{\sigma}_{ij \cdot \xi} = E_{ij \cdot \xi} = E_{ij} - \sum_{\xi} b_{i\xi} E^{\xi j}.$$

The distributional properties are similar to those in the two-population case. The quantities $E_{ij \cdot \xi}$ follow a Wishart distribution ith $(rc - r - 1 - k)$ d.f. and covariance matrix $\sigma_{ij \cdot \xi}$. The variates d'_i follow a multivariate normal distribution with means $r\delta_i \Sigma t_z^2$ and covariance

$$\sigma_{ij \cdot \xi}(r\Sigma t_z^2 + \Sigma E^{\xi\eta}\, d_\xi\, d_\eta) = v\sigma_{ij \cdot \xi} \qquad \text{(say)}.$$

Consequently,

$$y = \Sigma E^{ij \cdot \xi}\, d'_i\, d'_j$$

is distributed as $vpF/(rc - r - p - k)$ with p and $(rc - r - p - k)$ d.f. and parameter

$$\tau = \tfrac{1}{2}(r\Sigma t_z^2)^2 \Sigma \sigma^{ij \cdot \xi} \delta_i \delta_j / v.$$

Thus in the numerical example, with $r = 12, c = 4, p = 2, k = 1$, this procedure would have given an F test of the null hypothesis $\tau = 0$, where F has 2 and 33 d.f. However, the contribution from 2 degrees of freedom was deliberately omitted from the quantities E_{ij}, so that F actually had 2 and 31 d.f.

PART III

18. Justification of the 'dummy variate' approach. It remains to show that the method of calculation used in the example (sections 5 and 6) is equivalent to that derived from theory. There are two chief points to prove. First, that the M's found from the equations

$$\sum_j E_{ij \cdot \xi} M_j = d'_i \tag{30}$$

are proportional to the corresponding L's found from the equations

$$\sum_a S_{ij} L_j = d_i \tag{31}$$

where the suffix a denotes summation over both x_i and x_ξ variates.

Now, since $S_{ij} = E_{ij} + d_i\, d_j/240$, equations (31) are the same as

$$\sum_a E_{ij} L_j = d_i(1 - \sum_a L_j\, d_j/240). \tag{32}$$

Hence the L's in (31) are proportional to the values found from the equations

$$\sum_a E_{ij} L'_j = d_i\,. \tag{33}$$

But it is well known that if the L'_ξ are eliminated one by one from equations (33), we obtain

$$\sum_j E_{ij\cdot\xi} L'_j = d'_i ,$$

which is the same as (30). This proves the first point.

The second point to establish is that the F test in the example is the same as that obtained from theory. In section 15, it was shown that

$$\sum_{i,j} E^{ij\cdot\xi}\, d'_i\, d'_j/v \tag{34}$$

is distributed as $pF/(n - p - k + 1)$. In the analysis of variance of Table 6, section 6, the quantity following the same distribution was

$$\frac{(S_a - S_\xi)}{(240 - S_a)}\,, \tag{35}$$

where

$$S_a = \sum_a S^{ij}\, d_i\, d_j , \qquad S_\xi = \sum_{\xi,\eta} S^{\xi\eta}\, d_\xi\, d_\eta .$$

Since equations (31) and (32) have the same solution, we must have

$$S^{ij} = E^{ij}(1 - \sum_a L_j\, d_j/240) = E^{ij}(1 - S_a/240).$$

Multiplying both sides by $d_i\, d_j$ and summing over all i, j, we obtain

$$S_a = E_a(1 - S_a/240) = E_a(1 + E_a/240),$$

where E_a is defined analogously to S_a. Similarly

$$S_\xi = E_\xi/(1 + E_\xi/240).$$

Hence

$$\frac{S_a - S_\xi}{240 - S_a} = \frac{E_a - E_\xi}{240 + E_\xi} = \frac{E_a - E_\xi}{v}. \tag{36}$$

Transform the variates x_i , x_ξ into variates x'_i , x_ξ , where $x'_i = x_i - \Sigma b_{i\xi} x_\xi$. It is easy to see that this transforms

$$\sum_a E^{ij}\, d_i\, d_j \quad \text{into} \quad \sum E^{\xi\eta}\, d_\xi\, d_\eta + \sum_{i,j} E^{ij\cdot\xi}\, d'_i\, d'_j .$$

That is,

$$E_a = E_\xi + \sum_{i,j} E^{ij\cdot\xi}\, d'_i\, d'_j ,$$

since the quantity on the left is invariant under non-singular linear transformations. Hence from (36),

$$\frac{(S_a - S_\xi)}{(240 - S_a)} = \sum_{i,j} E^{ij\cdot\xi}\, d'_i\, d'_j/v.$$

From (34) and (35), this establishes the equivalence of the F tests. While the proof has been given only for the type of data encountered in the example, the same method will apply to other types of data.

In conclusion, we wish to thank the referees for many helpful suggestions in connection with the presentation of this paper.

REFERENCES

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

[2] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, Vol 8 (1938), pp. 376–386.

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

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

[5] C. I. Bliss and H. P. Marks, "The biological assay of insulin," *Quart. Jour. Pharm. and Pharmacol.*, Vol. 12 (1939), pp. 82–110; 182–205.

[6] P. L. Hsu, "Notes on Hotelling's generalized T," *Annals of Math. Stat.*, Vol. 9 (1938), pp. 231–243.

[7] H. Cramér, *Mathematical methods of statistics*, Princeton University Press, 1946.

[8] M. S. Bartlett, "On the theory of statistical regression," *Roy. Soc. Proc. Edin.*, Vol. 53 (1933), pp. 271–277.

[9] S. S. Wilks, *Mathematical Statistics*, Princeton University Press, 1943, p. 70.

[10] P. C. Tang, "The power function of the analysis of variance tests," *Stat. Res. Memoirs*, Vol. 2 (1938), pp. 126–157.