175

THE PRECISION OF DISCRIMINANT FUNCTIONS *

* See Author's Note, Paper 155.

1. Introductory

In a paper (1938a) on "The statistical utilization of multiple measurements" the author considered the general procedure of the establishment of discriminant functions, or sets of scores, based on an analysis of covariance, for a battery of different experimental determinations. In general, these functions are those giving stationary values to the ratio of apportionment of sums of squares between n_1 chosen and n_2 rèsidual degrees of freedom. In the simplest application n_1 is unity, and, as was first shown by Hotelling, the primary test of significance as to whether the set of measurements available are effective in making any significant discrimination, is exactly reducible to a simple analysis of variance in which $p-1$ degrees of freedom (for p variates) are transferred from the residual. In the general case the underlying problem of distribution has also now been solved (Fisher, 1939).

In both the simpler and in more general cases the question of the precision of the scores so ascertained is of immediate importance. Obviously, this presents certain peculiar features. If all coefficients of a discriminant function are increased or decreased in proportion, the function is effectively unchanged. No standard error can therefore be assigned to such a coefficient considered singly. Partial standard errors, on the other hand, in which all other coefficients are given fixed values, will certainly exist, although it is not at first sight obvious how they should be calculated.

In the case of the coefficients of a multiple regression equation, the author has often felt that the total standard errors ordinarily calculated were somewhat artificial, and certainly they are frequently misinterpreted. Thus, in the prediction of capacity to resist high altitude from data on individuals obtainable at sea level, it appeared in a recent study (Fisher, 1938 b), that when seven sea-level characteristics were employed in the prediction, not one of the coefficients was significant, although an apparently good prediction was obtained from the multiple regression formula. All that the non-significance meant, however, was that if any one of the coefficients were given the value zero *and the other coefficients readjusted*, the prediction formula was not significantly impaired. The sea-level characteristics showed, in fact, sufficiently close mutual correlation for any one of them to be capable of replacement by an appropriate linear function of the others, so as to compensate nearly completely for its absence from the prediction formula. Actually a prediction based on only four sea-level values was preferable to one based on all seven. Similar situations often arise in economics.

It is clear from this example that all questions relevant to the precision of the coefficients of a multiple regression formula may be expressed comprehensively in terms of a rule or test of significance as to whether any alternative formula proposed is significantly contradicted by the data. For multiple regression such a test is immediately available by multiplying the rows and columns of the c-matrix by the deviations between the coefficients arbitrarily chosen and those evaluated empirically.

In the paper referred to (1938*a*), I applied this concept of a generalized test of significance applicable to any function arbitrarily proposed, and showed that the sum of squares corresponding with Hotelling's T^2 became $T^2(1-r^2)$, where r is the correlation between the discriminant function proposed and that indicated by the data, *within* the samples which it is proposed to discriminate. This I carelessly interpreted to mean that Hotelling's T^2 was simply reduced to $T^2(1-r^2)$, forgetting that, in Hotelling's notation, T^2 also appears in the total sum of squares, which is unaffected. The numerical example, p. 386, is therefore incorrect, and the inconsistency of my formula has been pointed out by Bartlett (1939). The form in which Bartlett expresses the true relationship is, however, that my formula is correct if for the correlation within samples is substituted the correlation obtained when both samples are thrown together. This is true, but confusing, for while the correlation within homogeneous groups is an appropriate and natural method for measuring the similarity of two linear functions of the observations, the correlation when heterogeneous material is thrown together is of no intrinsic interest, its application being limited to the particular pair of samples under test.

The following section gives a simple demonstration of the correct formula from two complementary standpoints, with a view to exhibiting how the two correlations in question are related in any particular batch of data.

2. The test of significance of a proposed discriminant

In testing the significance of a discriminant function built of a number of different variates $x_1, \ldots x_p$ the analysis of variance appears in two different guises. We may consider the analysis of variance of a dummy variate y distinguishing the two contrasted samples, dividing the portion expressible in terms of $x_1, \ldots x_p$ as independent variates, from a residue not so expressible. Alternately, regarding our discriminant itself as a variate, we may analyse its variation between and within the samples.

Thus, if, for samples of N_1 and N_2, we take

$$y = N_2/(N_1+N_2)$$

for objects of the first sample, and

$$y = -N_1/(N_1+N_2)$$

for objects of the second sample, then the expected value of y for given values of $x_1, \ldots x_p$ will be

$$Y = X = \Sigma b^i x_i,$$

in which the coefficients b are given by the equations

$$\Sigma s'_{ij} b^j = S(x_i y) = \lambda^2 d_i,$$

where

$$\lambda^2 = N_1 N_2/(N_1+N_2), \quad s'_{ij} = S(x_i - \bar{x}_i)(x_j - \bar{x}_j),$$

and d_i is the difference between the means of the samples for variate i.

The analysis of variance for y is then

	Degrees of freedom	Sum of squares
Regression Remainder	p N_1+N_2-p-1	$S(Y^2)=\lambda^2\Sigma(bd)=\lambda^2R^2$ $S(y-Y)^2=\lambda^2\{1-\Sigma(bd)\}=\lambda^2(1-R^2)$
Total	N_1+N_2-1	$S(y^2)=\lambda^2$

On the other hand, considering X as a variate, we have

$$S(X^2)=\lambda^2R^2,$$

and if $\bar{X}_1$, $\bar{X}_2$ are the means of X in the two samples,

$$\lambda^2(\bar{X}_1-\bar{X}_2)=S(Xy)=\lambda^2R^2.$$

So that for the analysis of X we have

	Sum of squares
Between samples Within samples	λ^2R^4 $\lambda^2R^2(1-R^2)$
Total	λ^2R^2

an analysis equivalent, apart from a constant factor R^2, to the first.

Consider now any proposed form $\xi=\Sigma\beta^i x_i$

for the true discriminant of the population from which our sample is drawn. We shall be interested in its correlation with X *within* samples, denoted by r, rather than the total correlation r' when both samples are thrown together. If, however,

$$S(\xi^2)=A^2,$$

it follows that $S(\xi X)=A\lambda Rr'$.

Since, moreover, Y is the multiple regression prediction formula,

$$S(y-Y)\xi=0,$$

whence $S(\xi y)=A\lambda Rr'$.

From this it follows that $\lambda^2(\xi_1-\xi_2)=A\lambda Rr'$,

and that the sum of squares between samples

$$\lambda^2(\xi_1-\xi_2)^2=A^2R^2r'^2.$$

Hence the analysis for ξ may be completed, with the corresponding values for covariance with X, as follows:

	Degrees of freedom	Sum of squares (ξ^2)	Sum of products ($\chi\xi$)
Between samples Within samples	1 N_1+N_2-2	$A^2R^2r'^2$ $A^2(1-R^2r'^2)$	$A\lambda R^3r'$ $A\lambda Rr'(1-R^2)$
Total	N_1+N_2-1	A^2	$A\lambda Rr'$

The correlation coefficient r within samples is therefore given by

$$r^2 = \frac{r'^2(1-R^2)}{1-R^2r'^2},$$

or

$$r'^2 = \frac{r^2}{1-R^2(1-r^2)}.$$

Thus, the class of formulae specified by a fixed value of the correlation within samples has also a fixed value for the correlation when the samples are thrown together. Whereas, however, for any chosen formula, r is an intrinsic property of homogeneous populations, both R and r' will depend on the relative, and absolute, sizes of the samples.

If now ξ were used to predict y, we have for the analysis of y

	Degrees of freedom	Sum of squares
Prediction Remainder	1 N_1+N_2-2	$\lambda^2R^2r'^2 \quad = \lambda^2R^2r^2/\{1-R^2(1-r^2)\}$ $\lambda^2(1-R^2r'^2) = \lambda^2(1-R^2)/\{1-R^2(1-r^2)\}$
Total	N_1+N_2-1	$\lambda^2 \qquad \lambda^2$

and comparing this with prediction based on all p independent variates, we have, for testing the significance of the contribution of the others, after ξ has been taken account of,

	Degrees of freedom	Sum of squares
Additional information Remainder	$p-1$ N_1+N_2-p-1	$\lambda^2R^2(1-R^2)(1-r^2)/\{1-R^2(1-r^2)\}$ $\lambda^2(1-R^2)$
Total	N_1+N_2-2	$\lambda^2(1-R^2)/\{1-R^2(1-r^2)\}$

A similar analysis is found for X, if we eliminate covariance with ξ.

The modification of Hotelling's test needed when we wish to examine whether the discriminant indicated by the data differs significantly from any proposed form consists then in (i) reducing the number of degrees of freedom by unity, and (ii) substituting

$$R'^2 = R^2(1-r^2)$$

for R^2 as the ratio of the part to the whole in the sums of squares.

We should then reject any proposed discriminant formula, if its correlation r within samples with the best discriminant function obtainable is so low that

$$\begin{aligned} e^{2z} &= \frac{n-p+1}{p-1}\,\frac{R'^2}{1-R'^2} \\ &= \frac{n-p+1}{p-1}\,\frac{R^2(1-r^2)}{1-R^2(1-r^2)} = \frac{n-p+1}{p-1}\,\frac{T^2(1-r^2)}{n+T^2r^2} \end{aligned}$$

is significant for $\qquad n_1 = p-1, \quad n_2 = n-p+1.$

The corrected rule gives a much more reasonable basis for rejection. Thus the discriminant on four flower measurements for *Iris versicolor* and *I. setosa* (Fisher, 1936, p. 184) gives

$$R^2 = 0{\cdot}963416$$

for 4 against 95 degrees of freedom. For $n_1 = 3$, $n_2 = 95$, the 5 % value of z is 0·4968, the variance ratio is 2·7015. Multiplying by 3/95, the ratio of sums of squares is

$$R'^2/(1-R'^2) = 0\cdot085312;$$

hence

$$R'^2 = 0\cdot078606.$$

Dividing by R^2 it appears that the limiting value of r is given by

$$1-r^2 = 0\cdot081591$$

or

$$r = 0\cdot95834.$$

The precision with which the coefficients of the discriminant function have been determined is thus sufficient to reject at the 5 % level of significance any formula having a correlation with that found less than 0·95834, within the species. In this way we have a comprehensive and appropriate measure of the precision with which the discriminant function has been determined by the data.

3. Discriminant functions based on non-linear equations

The method of approach used in the present paper, in which the precision of the coefficients of a discriminant function is discussed through a test of significance of deviations from the hypothesis that the function has some other assigned form, brings clearly to view the complications that arise when more than a single degree of freedom is maximized.

For example, in a contingency table individuals are cross classified in two categories, such as eye colour and hair colour, as in the following example (Tocher's data for Caithness compiled by K. Maung of the Galton Laboratory).

Eye colour	Hair colour					
	Fair	Red	Medium	Dark	Black	Total
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Variation among the four eye colours may be regarded as due to variations in three variates defined conveniently in some such way as the following:

Eye colour	x_1	x_2	x_3
Blue	0	0	0
Light	1	0	0
Medium	0	1	0
Dark	0	0	1

We may then ask for what eye colour scores, i.e. for what linear function of x_1, x_2, x_3, are the five hair colour classes most distinct. The answer may be found in a variety of ways. For example, by starting with arbitrarily chosen scores for eye colour, determining from these average scores for hair colour, and using these latter to find new scores for eye colour.

Apart from a contraction of scale by a factor R^2 for each completed cycle, this form tends to a limit, and yields scores such as the following:

Eye colour	x	Hair colour	y
Light	−0·9873	Fair	−1·2187
Blue	−0·8968	Red	−0·5226
Medium	0·0753	Medium	−0·0941
Dark	1·5743	Dark	1·3189
		Black	2·4518

The particular values given above have been standardized so as to have mean values zero, and mean square deviations unity. In the sample from which they are derived each score has a linear regression on the other, the regression coefficient being 0·44627; this is, of course, equal to the correlation coefficient between the two scores regarded as variates. Hotelling has called pairs of functions of this kind canonical components. It may be noticed that no assumption is introduced as to the order of the classes of each category. In Tocher's schedule Light eyes come between Blue and Medium, but the discriminant function puts Blue between Medium and Light, though near the latter.

The precision of the scores assigned to different eye colours must be judged by the conformity of the data to various possible hypotheses concerning these scores. For example, we might test the hypothesis that the hair colour scores are correct, but that the apparent difference in score between Light and Blue eyes is illusory, their true scores being the same. The blue-eyed and the light-eyed children may here be compared directly, using the variate y:

Blue-eyed

	Frequency f	Score y	fy	fy^2
Fair	326	−1·2187	−397·30	484·2
Red	38	−0·5226	−19·86	10·4
Medium	241	−0·0941	−22·68	2·1
Dark	110	1·3189	145·08	191·3
Black	3	2·4518	7·36	18·0
	718		−287·40	706·0 115·0
			$\bar{y}$ −0·40028	$Sf(y-\bar{y})^2$ 591·0

Light-eyed

	Frequency	Score	fy	fy^2
Fair	688	−1·2187	−838·47	1021·8
Red	116	−0·5226	−60·61	31·7
Medium	584	−0·0941	−54·95	5·2
Dark	188	1·3189	247·95	327·0
Black	4	2·4518	9·81	24·0
	1580		−696·27	1409·7 306·8
			$\bar{y}$ −0·44068	$Sf(y-\bar{y})^2$ 1102·9

The sum of squares for error is 1693·9 for 2296 degrees of freedom, giving a mean square 0·73776: dividing this by 718 and 1580 we have 0·001028 and 0·000467, so that the variance of the difference between the scores is 0·001495 and the standard error 0·03867. The actual difference 0·04040 is therefore not significant.

In general, if we wish to compare the observed scores, derived from the data, with any proposed values ξ, we may test the linearity of the regression of y on ξ.

Thus, if ξ takes the values 0, 0, 1, 2 in the four classes for eye colour, we have

$$S(\xi) = 4404 \qquad \bar{\xi} = 0{\cdot}81752$$

$$S(\xi^2) = 7034 \qquad S(\xi-\bar{\xi})^2 = 3433{\cdot}63$$

$$S(y\xi) = RS(x\xi) = \qquad 1907{\cdot}83$$

$$S(y^2) = \qquad 5387{\cdot}00$$

$$S(y^2) - S^2(y\xi)/S(\xi^2) \qquad 4326{\cdot}95$$

Now the sum of squares for y within arrays is $5387(1-R^2) = 4313{\cdot}67$. So the analysis can be set out as follows:

	D.F.	S.S.	Mean square
Deviations from linear regression	2	13·28	6·64
Within arrays	5383	4313·67	0·80135
Total	5385	4326·95	

Thus the data show a decidedly significant departure from linearity. So that, if the scores for hair colour, y, be accepted, the data contradict significantly any set of values for ξ for which not only are Light and Blue eyes given equal scores, but Medium eyes are placed exactly half-way between these and Dark.

The consistency of these two methods may be illustrated by finding the contributions to the analysis above of two separate components. Of these one is the discrepancy between the means for Blue- and Light-eyed children, while the second is found by taking the means of Blue and Light together, adding the mean for the Dark-eyed, and comparing the sum with twice the mean of the Medium-eyed.

For the first comparison, we have the difference between Blue- and Light-eyed children, 0·04040; dividing the square of this by the sum of the reciprocals of 718 and 1580, we have 0·806 as the contribution of this component to the sum of squares.

For the second component we have

	Number	$S(y)$	Mean	Reciprocal
Blue and Light	2298	−983·73	−0·42808	0·00043516
Medium	1774	59·63	0·03361	0·00056370
Dark	1315	924·10	0·70274	0·00076046
Discrepancy		0·00	0·20744	0·00345042

The divisor now is $$\frac{1}{2298}+\frac{4}{1774}+\frac{1}{1315}=0\cdot00345042;$$

dividing the square of 0·20744 by this, we have 12·471 as the contribution of the second chosen component. The two components together give 13·28, checking with the value obtained for deviations from linear regression in the analysis of variance. The two discrepancies may thus be tested separately in succession. The significance of the two degrees of freedom is clearly due only to the second component.

It might seem that the problem which we have discussed for simple discriminant analysis was not analogous to that examined above, but to the wider question whether the data are compatible with the chosen values ξ for x, together with any set of scores for hair colour. In considering this problem, however, we must remember that there are three pairs of canonical components with corresponding correlations. If for the remaining two of them the correlation is insignificant, the corresponding components are presumably arbitrary, so that no significant deviation is to be expected from any ξ arbitrarily assigned. The practical question must involve the further stipulation that the correlation corresponding to our chosen component shall be the largest of the three possible values. Such a problem is not likely to have any easy solution.

REFERENCES

R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems." *Ann. Eugen., Lond.*, **7**, 179–88.

—— (1938*a*). "The statistical utilization of multiple measurements." *Ann. Eugen., Lond.*, **8**, 376–86.

—— (1938*b*). "On the statistical treatment of the relation between sea-level characteristics and high-altitude acclimatization." *Proc. Roy. Soc.* A, **126**, 25–9.

—— (1939). "The sampling distribution of some statistics obtained from non-linear equations." *Ann. Eugen., Lond.*, **9**, 238–49.

M. S. Bartlett (1939). "The standard errors of discriminant function coefficients." *J. Roy. Statist. Soc.* Suppl. **6**, 169–73.