Discriminant Analysis

It is a useful tool for situations where the total sample is to be divided into two or more groups which are mutually exclusive and collectively exhaustive, on the basis of a set of a predictor variable.

For e.g.:- A problem involving classifying customers into owners and non owners of video tape recorder.

Objectives:-

1) Finding linear composites of the predictor variables that enable the analyst to separate the groups by maximizing among groups relative to with in-groups variation
2) Establishing procedures for assigning new individuals, whose profile is not group identified, to one of the two groups
3) Testing whether significant differences exist between the mean predictor variable profile of the two groups
4) Determining which variables account most for intergroup differences in mean profiles

When the criterion variable has two categories, the technique is known as two group discriminant analysis. Multiple discriminant analysis refers to the case three or more categories are involved.

Conducting discriminant analysis is a five step procedure:-

1) Formulating the discriminant problem requires identification of the objectives and the criterion and predictor variables
2) Developing the linear combination of the predictor called discriminant function, so that the group differ as much as possible on the predictor variable
3) Determination of the statistical significance. It involves testing the null hypothesis i.e. in the population the means of all the discriminant functions in all groups is equal. If the null hypothesis is rejected, it is meaningful to interpret the result
4) The interpretation of the discriminant weights or coefficients.
   Relative importance of the variables may be obtained by examining the absolute magnitude of the standardized discriminant function coefficient
5) Validation:- it involves the development of classification matrix. The discriminant weights estimated by using the analysis sample are multiplied by the value of the predictor variables in the holdout sample to generate discriminant scores for the cases in the holdout sample

The cases are then assigned to groups based on their discriminant scores and an approximate decision rule. The percentage of cases correctly classified is determined and compared to the rate that would be expected by chance classification.

Approaches for estimating the coefficients

1) Direct Method :- it involves estimating all discriminant functions so that all the predictor are included simultaneously
2) Stepwise Method :- In this the predictor variables are entered sequentially based on their ability to discriminant among groups

In multiple discriminant analysis if there are G groups and K predictors, it is possible to estimate up to the smaller of G-1 or K discriminant functions

Numerical Example:-

Ready to eat cereal that was presented, here die ten consumer raters are simply asked to classify the cereals into one of the two categories like versus dislike.

X_1 – The amount of protein (in grams) per standard serving

X_2 – The percentage of minimum daily requirements of vitamin D per standard serving

Person	**Evaluation**	**X_1**	**X_2**		**X_1^2**	**X_2^2**	**X_1X_2**
1	Dislike	2	4		4	16	8
2	Dislike	3	2		9	4	6
3	Dislike	4	5		16	25	20
4	Dislike	5	4		25	16	20
5	Dislike	6	7		36	49	42
Mean		**4**	**4.4**	**Sum**	**90**	**110**	**96**
6	Like	7	6		49	36	42
7	Like	8	4		64	16	32
8	Like	9	7		81	49	63
9	Like	10	6		100	36	60
10	Like	11	9		121	81	99
Mean		**9**	**6.4**	**Sum**	**415**	**218**	**296**
Grand Mean		**6.5**	**5.4**				
Standard Deviation		**3.028**	**2.011**				

If we were forced to choose just one of the variables, it would appear that X_1 is better than X_2 We wonder if some linear composite of X_1 and X_2 could do better than X_1 alone.

Accordingly we have the following linear function:-

$$Z = K_1 X_1 + K_2 X_2$$

Where K_1 and K_2 are the weights that we seek.

Let us define,

$x_1 = X_1 - \overline{X_1}$

$x_2 = X_2 - \overline{X_2}$

(i.e. each observation measured from its mean)

Mean corrected sums of squares and cross products.

	Dislikers	**Likers**	**Total**
$\sum x_1^2 = \sum (X_1 - \overline{X_1})^2$	10	10	20
$\sum x_2^2 = \sum (X_2 - \overline{X_2})^2$	13.2	13.2	26.4
$\sum x_1 x_2 = \sum (X_1 - \overline{X_1})(X_2 - \overline{X_2})$	8	8	16

The normal equations are,

$$K_1 \sum x_1^2 + K_2 \sum x_1 x_2 = \overline{X_1}(\text{likers}) - \overline{X_1}(\text{dislikers})$$

$$K_1 \sum x_1 x_2 + K_2 \sum x_2^2 = \overline{X_2}(\text{likers}) - \overline{X_2}(\text{dislikers})$$

Solving them we get,

K_1 =0.368 ; K_2 =0.147

Discriminant function

$$Z = 0.368X_1 - 0.147X_2$$

We also find the discriminant scores for the means of the two groups and the grand mean.

$\overline{Z}$(Dislikers) =0.368*4 -0.147*4.4=0.824

$\overline{Z}$(likers) =0.368*9 -0.147*6.4=2.368

$\overline{Z}$(grand mean) =0.368*6.5-0.147*5.4=1.596

We note that the discriminant function "favours" X_1 by giving about 2.5 times the (absolute value) weight (K_1 =0.368 versus K_2 =-0.147) to X_1 as is given to X_2

Person	**Evaluation**	**Discriminant Score**
1	Dislikers	0.148
2	"	0.809
3	"	0.735
4	"	1.250
5	"	1.176
Mean		**0.824**
6	Likers	1.691
7	"	2.353
8	"	2.279
9	"	2.721
10	"	2.721
Mean		**2.368**
Grand Mean		**1.596**

Between group variability: $5(0.824-1.596)^2 + 5(2.368-1.596)^2$ =5.96

Within group variability:

Dislikers: $(0.148-0.824)^2 + (0.809-0.824)^2 + ... + (1.176-0.824)^2 = 0.772$

Likers: $(1.169-2.368)^2 + ... + (2.721-2.368)^2 = 0.772$

Total=1.544

Discriminant score $C = \frac{5.96}{1.544} = 3.68$

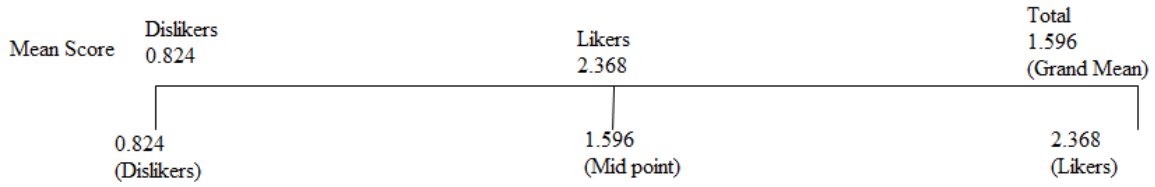
Since the normal equations for solving K_1 and K_2 obtained by maximizing the ratio between group and within group variance the discriminant criterion as calculated above 3.86 will be the maximum possible ratio.

If we suppress X_2 in the discriminant function and calculate another C it will be less than 3.86.

In discriminant function X_2 receives a negative weight bringing thereby the importance X_1 to the highest order.

This means protein is much more important than vitamin D.

Classifying the persons:-



It is all well and good to find the discriminant function, but the question is how to assigns the persons to the relevant groups.

- Assigns all cases with discriminant scores that are on the left of the midpoint (1.596) to the dislikers groups.
- Assigns all cases with discriminant scores that are on right of the midpoint (1.596) to the likers groups.

That is all true dislikers will be correctly classified as such and all true likers will be correctly classified. This can be shown by a 2x2 table: (assigned by the rule)

True state	Dislikers	Likers	Total
Dislikers	5	0	5
Likers	0	5	5
Total	5	5	10

Testing statistical significance:-

We test whether the group means differ significantly or not.

$$F = \frac{n_1 n_2 (n_1 + n_2 - m - 1)}{m(n_1 + n_2)(n_1 + n_2 - 2)} D^2$$

~F distribution with $n_1 + n_2 - m - 1$ degree of freedom.

Where

n_1 - number of observation in group 1

n_2 =no. Of observations in group 2.

m= no. Of independent variables

D^2 =mahalanobis square distance

In our problem

n_1=5, n_2=5, m=2(X_1 and X_2)

Simply way of calculating D^2 would be use the discriminant function.

D^2 = ($n_1 + n_2$ -2)*(0.368(5.0)-0.147(2)) =8(0.368*5-0.14*2) =12.353

In discriminant function

Z=0.368 X_1 -0.147 X_2

Where X_1 and X_2 are substituted by the respective group means differences:

$X_1(\text{likers}) - X_1(\text{dislikers})$ and $X_2(\text{likers}) - X_2(\text{dislikers})$

$$F = \frac{5*5(5+5-2-1)*12.353}{2*(5+5)(5+5-2)}$$

$$= \frac{25*7*12.353}{2*10*8} = 13.511$$

Table F(2,7)=4.74 at 5% level.

Since the calculated F exceeds table F at 5% level, reject H_0 and accept H_1 i.e, the group means are not equal in importance with a probability of 95%

This clearly validates the relatives importance of X_1 for higher than X_2.