Discriminant analysis is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

For example, the dependent variable may be the choice of the make of a new car (A, B or C) and the independent variables may be ratings of attributes of PCs on a seven-point Likert scale.

The objectives of discriminant analysis are as follows:

1 Development of discriminant functions, or linear combinations of the predictor or independent variables, that best discriminate between the categories of the criterion or dependent variable (groups).

2 Examination of whether significant differences exist among the groups, in terms of the predictor variables.

3 Determination of which predictor variables contribute to most of the intergroup differences.

4 Classification of cases to one of the groups based on the values of the predictor variables.

5 Evaluation of the accuracy of classification.

Discriminant analysis techniques are described by the number of categories possessed by the criterion variable. When the criterion variable has two categories, the technique is known as **two-group discriminant analysis**. When three or more categories are involved, the technique is referred to as **multiple discriminant analysis**. The main distinction is that in the two-group case it is possible to derive only one **discriminant function**, but in multiple discriminant analysis more than one function may be computed.

Few Examples of discriminant analysis in marketing research.

● In terms of demographic characteristics, how do customers who exhibit loyalty to a particular car manufacturer differ from those who do not?

● Do heavy users, medium users and light users of soft drinks differ in terms of their consumption of frozen foods?

● what psychographic characteristics help differentiate between price-sensitive and non-price- sensitive buyers of groceries?

● Do market segments differ in their media consumption habits?

● what are the distinguishing characteristics of consumers who respond to direct mail offers?

Similarities and differences among ANOVA, regression and discriminant analysis

Table 21.1 Similarities and differences among ANOVA, regression and discriminant analysis

	ANOVA	*Regression*	*Discriminant analysis*
Similarities			
Number of dependent variables	One	One	One
Number of independent variables	Multiple	Multiple	Multiple
Differences			
Nature of the dependent variable	Metric	Metric	Categorical Binary
Nature of the independent variable	Categorical	Metric	Metric

The important statistics associated with discriminant analysis include the following:

Canonical correlation. Canonical correlation measures the extent of association between the discriminant scores and the groups. It is a measure of association between the single discriminant function and the set of dummy variables that define the group membership.

Centroid. The centroid is the mean values for the discriminant scores for a particular group. There are as many centroids as there are groups, as there is one for each group. The means for a group on all the functions are the *group centroids*.

Classification matrix. Sometimes also called confusion or prediction matrix, the classification matrix contains the number of correctly classified and misclassified cases. The correctly classified cases appear on the diagonal, because the predicted and actual groups are the same. The off-diagonal elements represent cases that have been incorrectly classified. The sum of the diagonal elements divided by the total number of cases represents the *hit ratio*.

Discriminant function coefficients. The discriminant function coefficients (unstandardized) are the multipliers of variables, when the variables are in the original units of measurement.

Discriminant scores. The unstandardized coefficients are multiplied by the values of the variables. These products are summed and added to the constant term to obtain the discriminant scores.

Eigenvalue. For each discriminant function, the eigenvalue is the ratio of between-group to within-group sums of squares. Large eigenvalues imply superior functions.

F values and their significance. *F* values are calculated from a one-way ANOVA, with the grouping variable serving as the categorical independent variable. Each predictor, in turn, serves as the metric-dependent variable in the ANOVA.

Group means and group standard deviations. Group means and group standard deviations are computed for each predictor for each group.

Pooled within-group correlation matrix. The pooled within-group correlation matrix is computed by averaging the separate covariance matrices for all the groups.

Standardised discriminant function coefficients. The standardised discriminant function coefficients are the discriminant function coefficients that are used as the multipliers when the variables have been standardised to a mean of 0 and a variance of 1.

Structure correlations. Also referred to as discriminant loadings, the structure correlations represent the simple correlations between the predictors and the discriminant function.

Total correlation matrix. If the cases are treated as if they were from a single sample and the correlations are computed, a total correlation matrix is obtained.

Wilks' λ . Sometimes also called the *U* statistic, Wilks' λ for each predictor is the ratio of the within-group sum of squares to the total sum of squares. Its value varies between 0 and 1. Large values of λ (near 1) indicate that group means do not seem to be different. Small values of λ (near 0) indicate that the group means seem to be different.

The assumptions in discriminant analysis are that each of the groups is a sample from a multivariate normal population and that all the populations have the same covariance matrix.

Steps for conducting Discriminant Analysis

1. Formulate the problem

The first step in discriminant analysis is to formulate the problem by identifying the objectives, the criterion variable and the independent variables. The criterion variable must consist of two or more mutually exclusive and collectively exhaustive categories. When the dependent variable is interval or ratio scaled, it must first be converted into categories. For example, attitude towards the brand, measured on a seven-point scale, could be categorized as unfavorable (1, 2, 3), neutral (4) or favorable (5, 6, 7).

Alternatively, one could plot the distribution of the dependent variable and form groups of equal size by determining the appropriate cut-off points for each category. The predictor variables should be selected based on a theoretical model or previous research, or in the case of exploratory research, the experience of the researcher should guide their selection.

The next step is to divide the sample into two parts. One part of the sample, called as **analysis sample**, is used for estimation of the discriminant function. The other part, called as **validation sample**, is reserved for validating the discriminant function. The roles of the halves are then interchanged and the analysis is repeated. This is called double cross-validation.

Often, the distribution of the number of cases in the analysis and validation samples follows the distribution in the total sample. For instance, if the total sample contained 50% loyal and 50% non-loyal consumers, then the analysis and validation samples would each contain 50% loyal and 50% non-

loyal consumers. On the other hand, if the sample contained 25% loyal and 75% non-loyal consumers, the analysis and validation samples would be selected to reflect the same distribution (25% vs. 75%).
The validation of the discriminant function should be conducted repeatedly. Each time, the sample should be split into different analysis and validation parts. The discriminant function should be estimated and the validation analysis carried out. Thus, the validation assessment is based on a number of trials.

2. Estimate the discriminant function coefficients

Once the analysis sample has been identified, we can estimate the discriminant function coefficients. Two broad approaches are available. The **direct method** involves estimating the discriminant function so that all the predictors are included simultaneously. In this case, each independent variable is included, regardless of its discriminating power. This method is appropriate when, based on previous research or a theoretical model, the researcher wants the discrimination to be based on all the predictors.
An alternative approach is the stepwise method. In **stepwise discriminant analysis**, the predictor variables are entered sequentially, based on their ability to discriminate among groups. This method is appropriate when the researcher wants to select a subset of the predictors for inclusion in the discriminant function.

3. Determine the significance of the discriminant function

It would not be meaningful to interpret the analysis if the discriminant functions estimated were not statistically significant. The null hypothesis that, in the population, the means of all discriminant functions in all groups are equal can be statistically tested. In SPSS, this test is based on Wilks' λ. If several functions are tested simultaneously (as in the case of multiple discriminant analysis), the Wilks' λ statistic is the product of the univariate λ for each function. The significance level is estimated based on a chi-square transformation of the statistic. If the null hypothesis is rejected, indicating significant discrimination, one can proceed to interpret the results.

4. Interpret the results

The interpretation of the discriminant weights, or coefficients, is similar to that in multiple regression analysis. The value of the coefficient for a particular predictor depends on the other predictors included in the discriminant function. The signs of the coefficients are arbitrary, but they indicate which variable values result in large and small function values and associate them with particular groups.

We can obtain some idea of the relative importance of the variables by examining the absolute magnitude of the standardized discriminant function coefficients. Generally, predictors with relatively large standardized coefficients contribute more to the discriminating power of the function, as compared with predictors with smaller coefficients, and are therefore more important. Some idea of the relative importance of the predictors can also be obtained by examining the structure correlations, also called *canonical loadings* or *discriminant loadings*. These simple correlations between each predictor and the discriminant function represent the variance that the predictor shares with the function.

5. Assess the validity of discriminant analysis

Many computer programs, such as SPSS, offer a leave-one-out cross-validation option. In this option, the discriminant model is re-estimated as many times as there are respondents in the sample. Each re-estimated model leaves out one respondent and the model is used to predict for that respondent. When a large holdout sample is not possible, this gives a sense of the robustness of the estimate using each respondent in turn, as a holdout (validation).

As explained earlier, the data are randomly divided into two subsamples. One, the analysis sample, is used for estimating the discriminant function, and the validation sample is used for developing the classification matrix. The discriminant weights, estimated by using the analysis sample, are multiplied by the values of the predictor variables in the holdout (validation) 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 appropriate decision rule. For example, in two-group discriminant analysis, a case will be assigned to the group whose centroid is the closest. The **hit ratio**, or the percentage of cases correctly classified, can then be determined by summing the diagonal elements and dividing by the total Number of cases.