Chapter 25
Discriminant Analysis

Content list

When you have read this chapter you will understand:

1 The purposes of discriminant analysis.
2 How to use SPSS to perform discriminant analysis.
3 How to interpret the SPSS print out of discriminant analysis.

Introduction
This chapter introduces another extension of regression where the DV may have more than two conditions at a categorical level and IV's are scale data.

The purposes of discriminant analysis (DA)

Discriminant Function Analysis (DA) undertakes the same task as multiple linear regression by predicting an outcome. However, multiple linear regression is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values. But many interesting variables are categorical, such as political party voting intention, migrant/non-migrant status, making a profit or not, holding a particular credit card, owning, renting or paying a mortgage for a house, employed/unemployed, satisfied versus dissatisfied employees, which customers are likely to buy a product or not buy, what distinguishes Stellar Bean clients from Gloria Beans clients, whether a person is a credit risk or not, etc.

DA is used when:

- the dependent is categorical with the predictor IV's at interval level such as age, income, attitudes, perceptions, and years of education, although dummy variables can be used as predictors as in multiple regression. Logistic regression IV's can be of any level of measurement.
- there are more than two DV categories, unlike logistic regression, which is limited to a dichotomous dependent variable.

Discriminant analysis linear equation

DA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D = v_1X_1 + v_2X_2 + v_3X_3 = \ldots\ldots.. v_iX_i + a$$

Where D = discriminate function
v = the discriminant coefficient or weight for that variable
X = respondent's score for that variable
a = a constant
i = the number of predictor variables

This function is similar to a regression equation or function. The v's are unstandardized discriminant coefficients analogous to the b's in the regression equation. These v's maximize the distance between the means of the criterion (dependent) variable. Standardized discriminant coefficients can also be used like beta weight in regression. Good predictors tend to have large weights. What you want this function to do is maximize the distance between the categories, i.e. come up with an equation that has strong discriminatory power between groups. After using an existing set of data to calculate the discriminant function and classify cases, any new cases can then be classified. The number of discriminant functions is one less the number of groups. There is only one function for the basic two group discriminant analysis.

A discriminant score. *This is a weighted linear combination (sum) of the discriminating variables.*

Assumptions of discriminant analysis

The major underlying assumptions of DA are:

- the observations are a random sample;
- each predictor variable is normally distributed;

- each of the allocations for the dependent categories in the initial classification are correctly classified;
- there must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group);
- each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division;
- for instance, three groups taking three available levels of amounts of housing loan;
- the groups or categories should be defined before collecting the data;
- the attribute(s) used to separate the groups should discriminate quite clearly between the groups so that group or category overlap is clearly non-existent or minimal;
- group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

There are several purposes of DA:

- To investigate differences between groups on the basis of the attributes of the cases, indicating which attributes contribute most to group separation. The descriptive technique successively identifies the linear combination of attributes known as canonical discriminant functions (equations) which contribute maximally to group separation.
- Predictive DA addresses the question of how to assign new cases to groups. The DA function uses a person's scores on the predictor variables to predict the category to which the individual belongs.
- To determine the most parsimonious way to distinguish between groups.
- To classify cases into groups. Statistical significance tests using chi square enable you to see how well the function separates the groups.
- To test theory whether cases are classified as predicted.

***Discriminant analysis** – creates an equation which will minimize the possibility of misclassifying cases into their respective groups or categories.*

The aim of the statistical analysis in DA is to combine (weight) the variable scores in some way so that a single new composite variable, the discriminant score, is produced. One way of thinking about this is in terms of a food recipe, where changing the proportions (weights) of the ingredients will change the characteristics of the finished cakes. Hopefully the weighted combinations of ingredients will produce two different types of cake.

Similarly, at the end of the DA process, it is hoped that each group will have a normal distribution of discriminant scores. The degree of overlap between the discriminant score distributions can then be used as a measure of the success of the technique, so that, like the

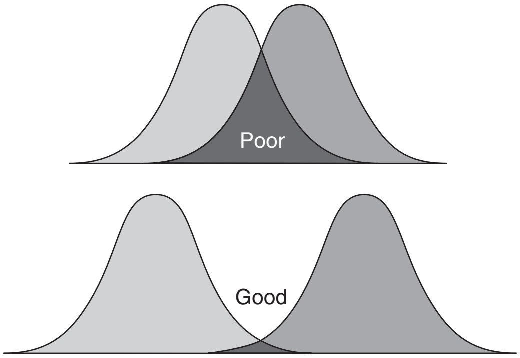


Figure 25.1 Discriminant distributions.

different types of cake mix, we have two different types of groups (Fig. 25.1). For example:

The top two distributions in Figure 25.1 overlap too much and do not discriminate too well compared to the bottom set. Misclassification will be minimal in the lower pair, whereas many will be misclassified in the top pair.

Standardizing the variables ensures that scale differences between the variables are eliminated. When all variables are standardized, absolute weights (i.e. ignore the sign) can be used to rank variables in terms of their discriminating power, the largest weight being associated with the most powerful discriminating variable. Variables with large weights are those which contribute mostly to differentiating the groups.

As with most other multivariate methods, it is possible to present a pictorial explanation of the technique. The following example uses a very simple data set, two groups and two variables. If scattergraphs are plotted for scores against the two variables, distributions like those in Figure 25.2 are obtained.

The new axis represents a new variable which is a linear combination of x and y, i.e. it is a discriminant function (Fig. 25.3). Obviously, with more than two groups or variables this graphical method becomes impossible.

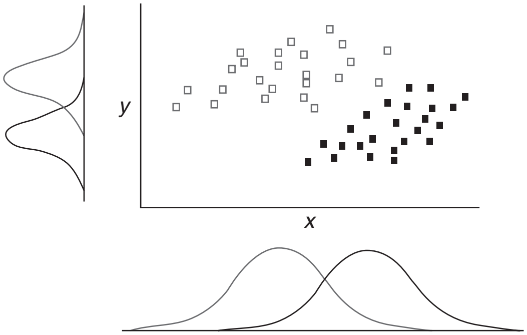


Figure 25.2 Scattergraph displaying distributions by axis.

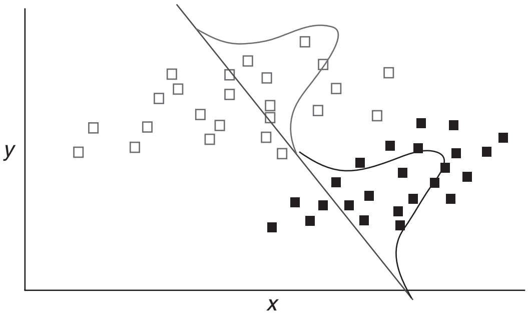


Figure 25.3 **New axis creating greater discrimination.**

Clearly, the two groups can be separated by these two variables, but there is a large amount of overlap on each single axis (although the y variable is the 'better' discriminator). It is possible to construct a new axis which passes through the two group centroids ('means'), such that the groups do not overlap on the new axis.

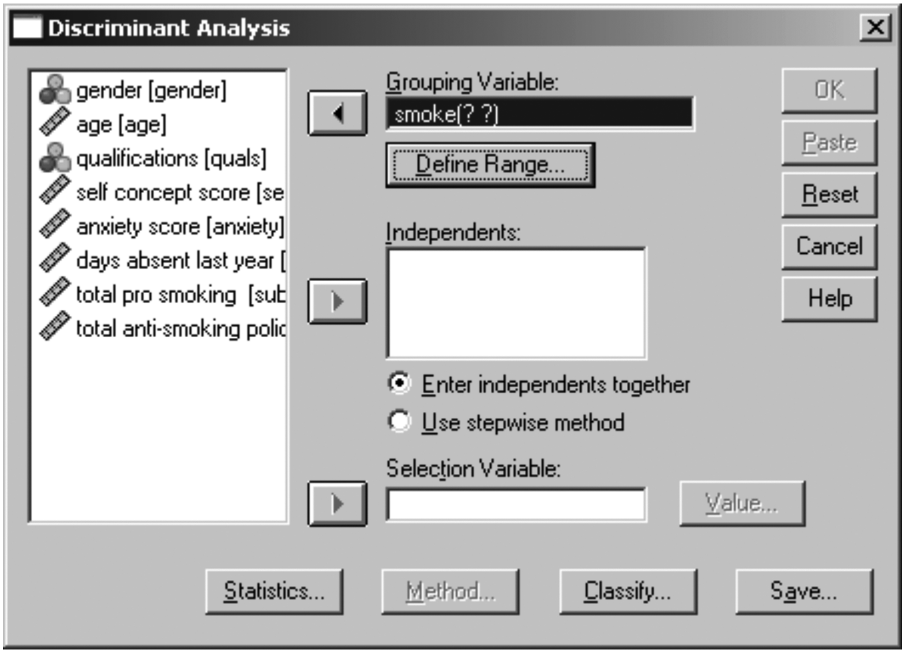
In a two-group situation predicted membership is calculated by first producing a score for D for each case using the discriminate function. Then cases with D values smaller than the cut-off value are classified as belonging to one group while those with values larger are classified into the other group. SPSS will save the predicted group membership and D scores as new variables.

The group centroid is the mean value of the discriminant score for a given category of the dependent variable. There are as many centroids as there are groups or categories. The cut-off is the mean of the two centroids. If the discriminant score of the function is less than or equal to the cut-off the case is classed as 0, whereas if it is above, it is classed as 1.

SPSS activity – discriminant analysis

Please access SPSS Chapter 25 Data File A on the web page. You will now be taken through a discriminant analysis using that data which includes demographic data and scores on various questionnaires. '*smoke*' is a nominal variable indicating whether the employee smoked or not. The other variables to be used are age, days absent sick from work last year, self-concept score, anxiety score and attitudes to anti-smoking at work score. The aim of the analysis is to determine whether these variables will discriminate between those who smoke and those who do not. This is a simple discriminant analysis with only two groups in the DV. With three or more DV groupings a multiple discriminant analysis is involved, but this follows the same process in SPSS as described below except there will be more than one set of eigenvalues, Wilks' Lambda's and beta coefficients. The number of sets is always one less than the number of DV groups.

1 ***Analyse*** >> ***Classify*** >> ***Discriminant***
2 Select '*smoke*' as your grouping variable and enter it into the ***Grouping Variable Box*** (Fig. 25.4).


Discriminant Analysis

Grouping Variable: smoke(? ?)

Independents:

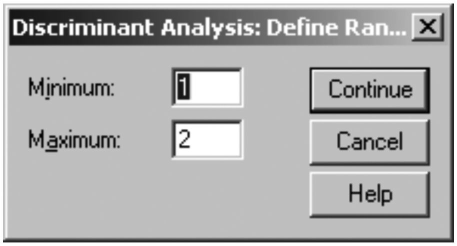
● Enter independents together

○ Use stepwise method

Selection Variable:

Figure 25.4 **Discriminant analysis dialogue box.**

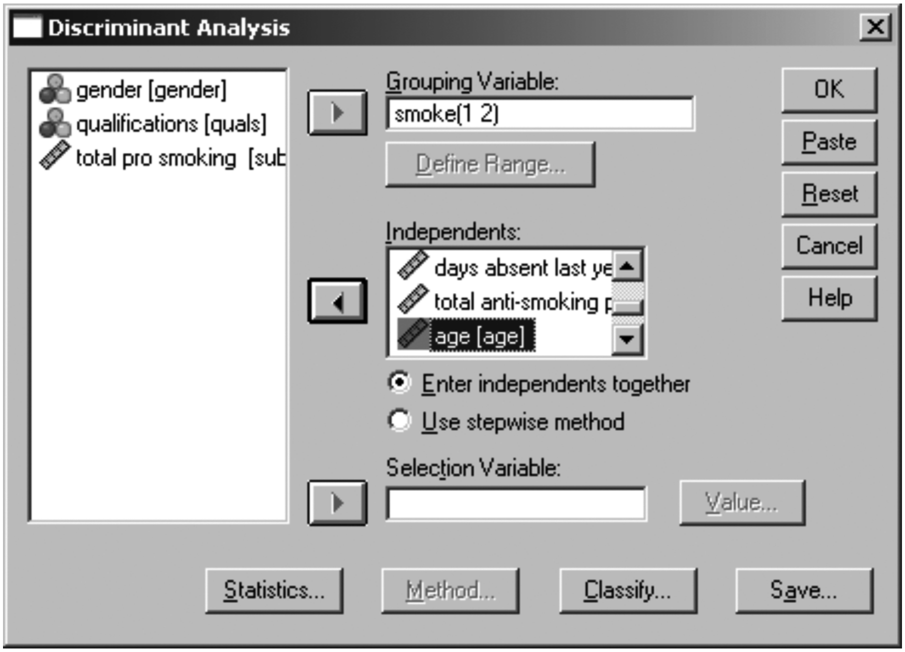
3 Click ***Define Range*** button and enter the lowest and highest code for your groups (here it is 1 and 2) (Fig. 25.5).
4 Click ***Continue***.
5 Select your predictors (IV's) and enter into ***Independents*** box (Fig. 25.6) and select ***Enter Independents Together***. If you planned a stepwise analysis you would at this point select ***Use Stepwise Method*** and not the previous instruction.
6 Click on ***Statistics*** button and select ***Means, Univariate Anovas, Box's M, Unstandardized*** and ***Within-Groups Correlation*** (Fig. 25.7).


Discriminant Analysis: Define Ran...

Minimum: 1

Maximum: 2

Figure 25.5 **Define range box.**



gender [gender]
qualifications [quals]
total pro smoking [sub

Grouping Variable: smoke(1 2)

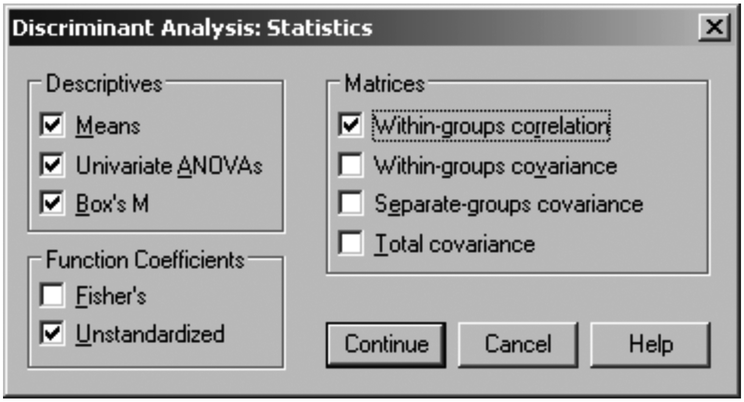
Independents:
days absent last ye
total anti-smoking p
age [age]

● Enter independents together
○ Use stepwise method

Selection Variable:

Figure 25.6 **Discriminant analysis dialogue box.**

7 ***Continue*** >> ***Classify***. Select ***Compute From Group Sizes, Summary Table, Leave One Out Classification, Within Groups***, and all ***Plots*** (Fig. 25.8).
8 ***Continue*** >> ***Save*** and select ***Predicted Group Membership*** and ***Discriminant Scores*** (Fig. 25.9).
9 ***OK***.

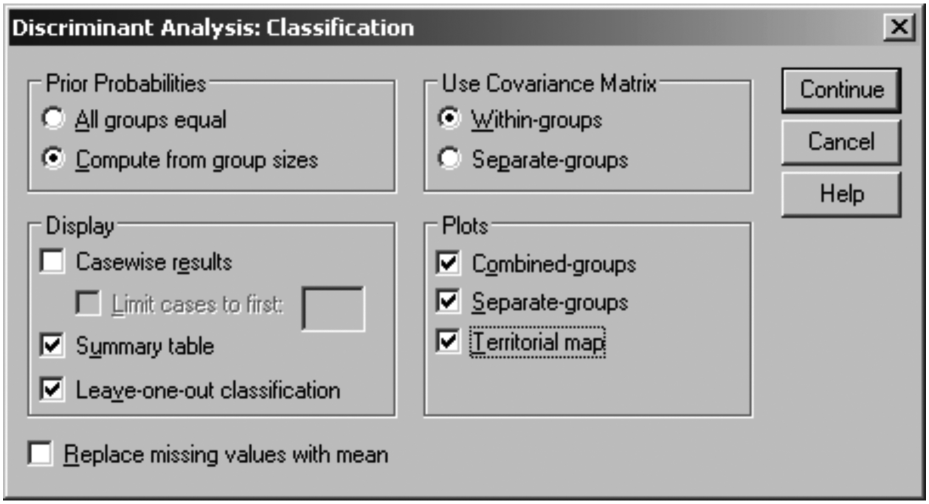


Descriptives
☑ Means
☑ Univariate ANOVAs
☑ Box's M

Function Coefficients
□ Fisher's
☑ Unstandardized

Matrices
☑ Within-groups correlation
□ Within-groups covariance
□ Separate-groups covariance
□ Total covariance

Figure 25.7 **Discriminant analysis statistics box.**



Prior Probabilities
○ All groups equal
● Compute from group sizes

Use Covariance Matrix
● Within-groups
○ Separate-groups

Display
□ Casewise results
□ Limit cases to first:
☑ Summary table
☑ Leave-one-out classification

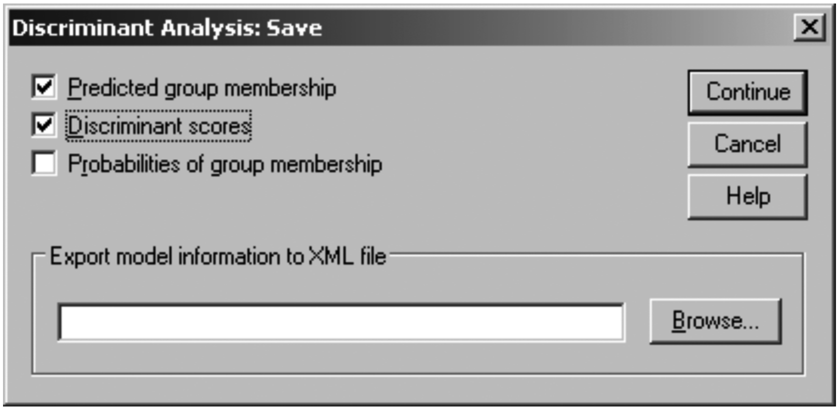
Plots
☑ Combined-groups
☑ Separate-groups
☑ Territorial map

□ Replace missing values with mean

Figure 25.8 **Discriminant analysis classification box.**

Interpreting the printout Tables 25.1 to 25.12
The initial case processing summary as usual indicates sample size and any missing data.

Group statistics tables
In discriminant analysis we are trying to predict a group membership, so firstly we examine whether there are any significant differences between groups on each of the independent variables using group means and ANOVA results data. The Group Statistics and Tests of Equality of Group Means tables provide this information. If there are no significant group differences it is not worthwhile proceeding any further with the analysis. A rough idea of variables that may be important can be obtained by inspecting the group means and



☑ Predicted group membership
☑ Discriminant scores
□ Probabilities of group membership

Export model information to XML file

Figure 25.9 **Discriminant analysis save box.**

Table 25.1 **Group statistics table**

Group Statistics					
				Valid N (listwise)	
smoke or not		Mean	Std. deviation	Unweighted	Weighted
non-smoker	age	38.7665	9.23647	257	257.000
	self concept score	46.6148	11.16826	257	257.000
	anxiety score	19.6848	5.23565	257	257.000
	days absent last year	4.8482	5.39643	257	257.000
	total anti-smoking policies subtest B	22.6770	2.56036	257	257.000
smoker	age	36.1934	8.52325	181	181.000
	self concept score	28.2818	6.54159	181	181.000
	anxiety score	28.5028	7.25153	181	181.000
	days absent last year	8.3481	7.53107	181	181.000
	total anti-smoking policies subtest B	20.6409	3.15670	181	181.000
Total	age	37.7032	9.02823	438	438.000
	self concept score	39.0388	13.12921	438	438.000
	anxiety score	23.3288	7.52428	438	438.000
	days absent last year	6.2945	6.58773	438	438.000
	total anti-smoking policies subtest B	21.8356	2.99204	438	438.000

standard deviations. For example, mean differences between self-concept scores and anxiety scores depicted in Table 25.1 suggest that these may be good discriminators as the separations are large. Table 25.2 provides strong statistical evidence of significant differences between means of smoke and no smoke groups for all IV's with self-concept and anxiety producing very high value F's. The Pooled Within-Group Matrices (Table 25.3) also supports use of these IV's as intercorrelations are low.

Table 25.2 **Tests of equality of group means table**

Tests of Equality of Group Means					
	Wilks' Lambda	F	df1	df2	Sig.
age	.980	8.781	1	436	.003
self concept score	.526	392.672	1	436	.000
anxiety score	.666	218.439	1	436	.000
days absent last year	.931	32.109		436	.000
total anti-smoking policies subtest B	.887	55.295	1	436	.000

Table 25.3 **Pooled within-groups matrices**

Pooled Within-Groups Matrices						
		age	Self concept score	Anxiety score	Days absent last year	Total anti-smoking policies subtest B
Correlation	age	1.000	–.118	.060	.042	.061
	self concept score	–.118	1.000	.042	–.143	–.044
	anxiety score	.060	.042	1.000	.118	.137
	days absent last year	.042	–.143	.118	1.000	.116
	total anti-smoking policies subtest B	.061	–.044	.137	.116	1.000

Log determinants and Box's M tables

In ANOVA, an assumption is that the variances were equivalent for each group but in DA the basic assumption is that the variance-co-variance matrices are equivalent. Box's M tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent. The researcher wants this test not to be significant so that the null hypothesis that the groups do not differ can be retained.

For this assumption to hold, the log determinants should be equal. When tested by Box's M, we are looking for a non-significant M to show similarity and lack of significant differences. In this case the log determinants appear similar and Box's M is 176.474 with F = 11.615 which is significant at p < .000 (Tables 25.4 and 25.5). However, with large samples, a significant result is not regarded as too important. Where three or more groups exist, and M is significant, groups with very small log determinants should be deleted from the analysis.

Table of eigenvalues

This provides information on each of the discriminate functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus 1. We are only using two groups here, namely 'smoke' and 'no smoke', so only one function is displayed. The canonical correlation is the multiple correlation between the predictors and the discriminant function. With only one function it provides an index of overall model fit which is interpreted as being the proportion of variance explained (R^2). In our

Table 25.4 **Log determinants table**

Log Determinants		
Smoke or not	Rank	Log determinant
non-smoker	5	17.631
smoker	5	18.058
Pooled within-groups	5	18.212

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Table 25.5 **Box's M test results table**

Test Results		
Box's M		176.474
F	Approx.	11,615
	df1	15
	df2	600825.3
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

example (Table 25.6) a canonical correlation of .802 suggests the model explains 64.32% of the variation in the grouping variable, i.e. whether a respondent smokes or not.

Table 25.6 **Eigenvalues table**

Eigenvalues				
Function	**Eigenvalue**	**% of variance**	**Cumulative %**	**Canonical correlation**
1	1.806[a]	100.0	100.0	.802

[a] First 1 canonical discriminant functions were used in the analysis.

Wilks' lambda

Wilks' lambda indicates the significance of the discriminant function. This table (Table 25.7) indicates a highly significant function (p < .000) and provides the proportion of total variability not explained, i.e. it is the converse of the squared canonical correlation. So we have 35.6% unexplained.

Table 25.7 **Wilks' lambda table**

Wilks' Lambda				
Test of function(s)	**Wilks' Lambda**	**Chi-square**	**df**	**Sig.**
1	.356	447.227	5	.000

The standardized canonical discriminant function coefficients table

The interpretation of the discriminant coefficients (or weights) is like that in multiple regression. Table 25.8 provides an index of the importance of each predictor like the standardized regression coefficients (beta's) did in multiple regression. The sign indicates the direction of the relationship. Self-concept score was the strongest predictor while low anxiety (note –ve sign) was next in importance as a predictor. These two variables with large coefficients stand out as those that strongly predict allocation to the smoke or do not smoke group. Age, absence from work and anti-smoking attitude score were less successful as predictors.

Table 25.8 Standardized canonical discriminant function coefficients table

Standardized Canonical Discriminant Function Coefficients

	Function
	1
age	.212
self concept score	.763
anxiety score	–.614
days absent last year	–.073
total anti-smoking policies subtest B	.378

The structure matrix table

Table 25.9 provides another way of indicating the relative importance of the predictors and it can be seen below that the same pattern holds. Many researchers use the structure matrix correlations because they are considered more accurate than the Standardized Canonical Discriminant Function Coefficients. The structure matrix table (Table 25.9) shows the corelations of each variable with each discriminate function. These Pearson coefficients are structure coefficients or discriminant loadings. They serve like factor loadings in factor analysis. By identifying the largest loadings for each discriminate function the researcher gains insight into how to name each function. Here we have self-concept and anxiety (low scores) which suggest a label of personal confidence and effectiveness as the function that discriminates between non-smokers and smokers. Generally, just like factor loadings, 0.30 is seen as the cut-off between important and less important variables. Absence is clearly not loaded on the discriminant function, i.e. is the weakest predictor and suggests that work absence is not associated with smoking behaviour but a function of other unassessed factors.

Table 25.9 Structure matrix table

Structure Matrix

	Function
	1
self concept score	.706
anxiety score	–.527
total anti-smoking policies subtest B	.265
days absent last year	–.202
age	.106

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function.

Table 25.10 **Canonical Discriminant Function Coefficients table**

Canonical discriminant function coefficients	
	Function
	1
age	.024
self concept score	.080
anxiety score	–.100
days absent last year	–.012
total anti-smoking policies subtest B	.134
(Constant)	–4.543

Unstandardized coefficients.

The canonical discriminant function coefficient table

These unstandardized coefficients *(b)* are used to create the discriminant function (equation). It operates just like a regression equation. In this case we have (Table 25.10):

D = (.024 × *age*) + (.080 × *self-concept*) + (–.100 × *anxiety*) + (–.012 days *absent*) + (.134 *anti smoking score*) – 4.543.

The discriminant function coefficients *b* or standardized form *beta* both indicate the partial contribution of each variable to the discriminate function controlling for all other variables in the equation. They can be used to assess each IV's unique contribution to the discriminate function and therefore provide information on the relative importance of each variable. If there are any dummy variables, as in regression, individual beta weights cannot be used and dummy variables must be assessed as a group through hierarchical DA running the analysis, first without the dummy variables then with them. The difference in squared canonical correlation indicates the explanatory effect of the set of dummy variables.

Group centroids table

A further way of interpreting discriminant analysis results is to describe each group in terms of its profile, using the group means of the predictor variables. These group means are called centroids. These are displayed in the Group Centroids table (Table 25.11). In our example, non-smokers have a mean of 1.125 while smokers produce a mean of –1.598. Cases with scores near to a centroid are predicted as belonging to that group.

Table 25.11 **Functions at group centroids table**

Functions at Group Centroids	
	Function
smoke or not	**1**
non-smoker	1.125
smoker	–1.598

Unstandardized canonical discriminant functions evaluated at group means.

Classification table

Finally, there is the classification phase. The classification table, also called a confusion table, is simply a table in which the rows are the observed categories of the dependent and the columns are the predicted categories. When prediction is perfect all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. The cross validated set of data is a more honest presentation of the power of the discriminant function than that provided by the original classifications and often produces a poorer outcome. The cross validation is often termed a 'jack-knife' classification, in that it successively classifies all cases but one to develop a discriminant function and then categorizes the case that was left out. This process is repeated with each case left out in turn. This cross validation produces a more reliable function. The argument behind it is that one should not use the case you are trying to predict as part of the categorization process.

The classification results (Table 25.12) reveal that 91.8% of respondents were classified correctly into 'smoke' or 'do not smoke' groups. This overall predictive accuracy of the discriminant function is called the 'hit ratio'. Non-smokers were classified with slightly better accuracy (92.6%) than smokers (90.6%). What is an acceptable hit ratio? You must compare the calculated hit ratio with what you could achieve by chance. If two samples are equal in size then you have a 50/50 chance anyway. Most researchers would accept a hit ratio that is 25% larger than that due to chance.

Table 25.12 **Classification results table**

Classification Results[b,c]

			Predicted Group Membership		
		smoke or not	**non-smoker**	**smoker**	**Total**
Original	Count	non-smoker	238	19	257
		smoker	17	164	181
	%	non-smoker	92.6	7.4	100.0
		smoker	9.4	90.6	100.0
Cross-validated[a]	Count	non-smoker	238	19	257
		smoker	17	164	181
	%	non-smoker	92.6	7.4	100.0
		smoker	9.4	90.6	100.0

[a] Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

[b] 91.8% of original grouped cases correctly classified.

[c] 91.8% of cross-validated grouped cases correctly classified.

Saved variables

As a result of asking the analysis to save the new groupings, two new variables can now be found at the end of your data file. *dis_1* is the predicted grouping based on the discriminant analysis coded 1 and 2, while *dis1_1* are the D scores by which the cases were coded into their categories. The average D scores for each group are of course the group centroids reported earlier. While these scores and groups can be used for other analyses, they are useful as visual demonstrations of the effectiveness of the discriminant function. As an example, histograms (Fig. 25.10) and box plots (Fig. 25.11) are alternative ways of illustrating the distribution of the discriminant function scores for each group. By reading the range of scores on the axes, noting (group centroids table) the means of both as well as the very minimal overlap of the graphs and box plots, a substantial discrimination is revealed. This suggests that the function does discriminate well, as the previous tables indicated.

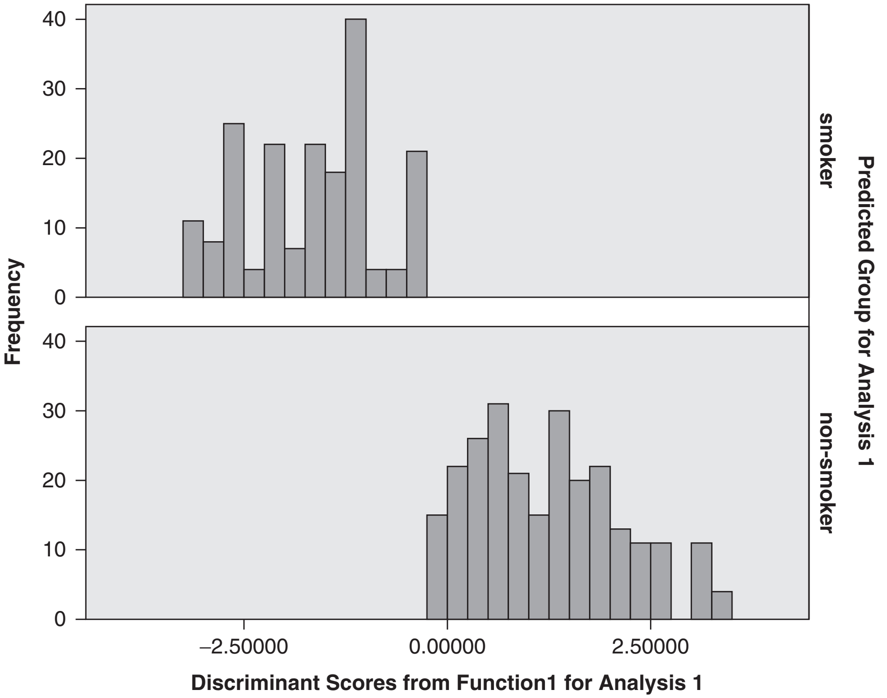


Figure 25.10 Histograms showing the distribution of discriminant scores for smokers and non-smokers.

New cases

Mahalanobis distances (obtained from the **Method Dialogue Box**) are used to analyse cases as it is the distance between a case and the centroid for each group of the dependent. So a new case or cases can be compared with an existing set of cases. A new case will have one distance for each group and therefore can be classified as belonging to the group for which its distance is smallest. Mahalanobis distance is measured in terms of SD from the centroid, therefore a case that is more than 1.96 Mahalanobis distance units from the centroid has a less than 5% chance of belonging to that group.

How to write up the results

'A discriminant analysis was conducted to predict whether an employee was a smoker or not. Predictor variables were age, number of days from work in previous year, self-concept score,

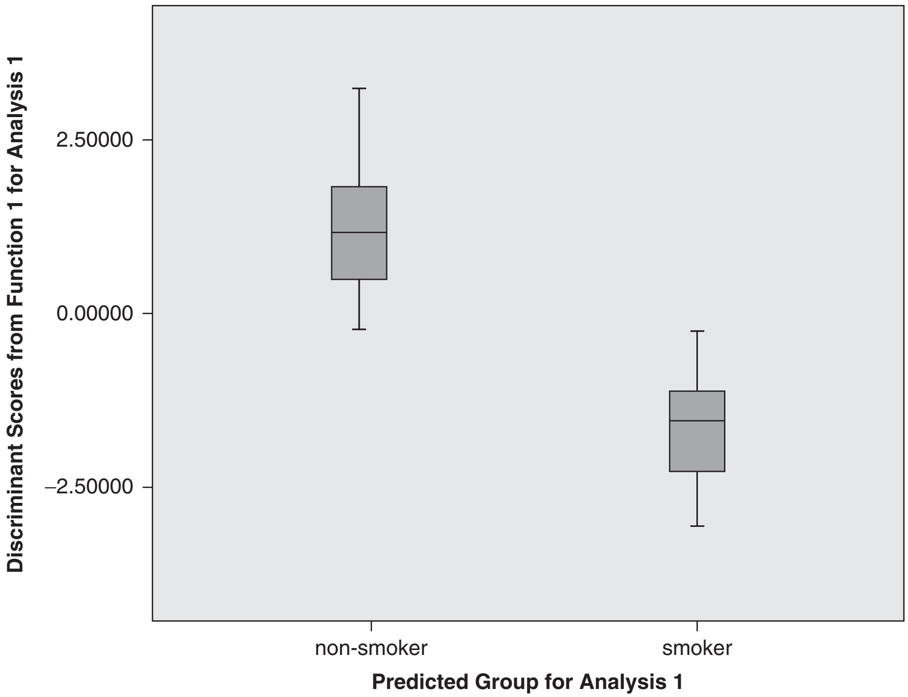


Figure 25.11 **Box plots illustrating the distribution of discriminant scores for the two groups.**

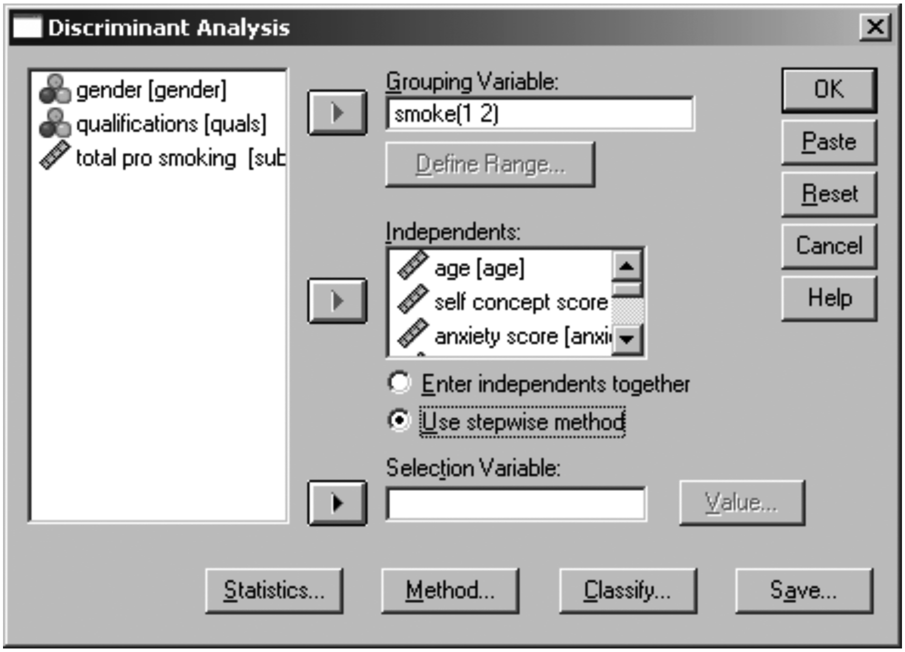
anxiety score, and attitude to anti-smoking workplace policy. Significant mean differences were observed for all the predictors on the DV. While the log determinants were quite similar, Box's M indicated that the assumption of equality of covariance matrices was violated. However, given the large sample, this problem is not regarded as serious. The discriminate function revealed a significant association between groups and all predictors, accounting for 64.32% of between group variability, although closer analysis of the structure matrix revealed only two significant predictors, namely self-concept score (.706) and anxiety score (–.527) with age and absence poor predictors. The cross validated classification showed that overall 91.8% were correctly classified'.

Stepwise discriminant analysis using Chapter 25 SPSS data file D

Stepwise discriminate analysis, like its parallel in multiple regression, is an attempt to find the best set of predictors. It is often used in an exploratory situation to identify those variables from among a larger number that might be used later in a more rigorous theoretically driven study. In stepwise DA, the most correlated independent is entered first by the stepwise programme, then the second until an additional dependent adds no significant amount to the canonical R squared. The criteria for adding or removing is typically the setting of a critical significance level for '*F to remove*'.

To undertake this example, please access SPSS Ch 25 Data File A. It is the same file we used above. On this occasion we will enter the same predictor variables one step at a time to see which combinations are the best set of predictors, or whether all of them are retained. Only one of the SPSS screen shots will be displayed, as the others are the same as those used above.

1 Click ***Analyze*** >> ***Classify*** >> ***Discriminant***.
2 Select grouping variable and transfer to ***Grouping Variable*** box. Then click ***Define Range*** button and enter the lowest and highest codes for your grouping variable define range.
3 Click ***Continue*** then select predictors and enter into ***Independents box***. Then click on ***Use Stepwise Methods***. This is the important difference from the previous example (Fig. 25.12).
4 ***Statistics*** >> ***Means, Univariate Anovas, Box's M, Unstandardized and Within Groups Correlation.***
5 Click ***Classify.*** Select ***Compute From Group Sizes, Summary Table, Leave One Out Classification, Within Groups,*** and all ***Plots.***
6 ***Continue*** >> ***Save*** and select ***Predicted Group Membership*** and ***Discriminant Scores.***
7 ***OK***.



gender [gender]
qualifications [quals]
total pro smoking [sub

Grouping Variable: smoke(1 2)

Independents:
age [age]
self concept score
anxiety score [anxi

○ Enter independents together
● Use stepwise method

Selection Variable:

Figure 25.12 **Discriminant analysis dialogue box selected for stepwise method.**

Interpretation of printout Tables 25.13 and 25.14

Many of the tables in this stepwise discriminant analysis are the same as those for the basic analysis, and we will therefore only comment on the extra stepwise statistics tables.

Stepwise statistics tables

The Stepwise Statistics Table (25.13) shows that four steps were taken, with each one including another variable and therefore these four were included in the Variables in the Analysis and Wilks Lambda tables because each was adding some predictive power to the function. In some stepwise analyses only the first one or two steps might be taken, even though there are more variables, because succeeding additional variables are not adding to the predictive power of the discriminant function.

Table 25.13 Variables in the analysis table

Variables in the Analysis				
Step		**Tolerance**	**F to Remove**	**Rao's V**
1	self concept score	1.000	392.672	
2	self concept score	.998	277.966	218.439
	anxiety score	.998	128.061	392.672
3	self concept score	.996	255.631	309.665
	anxiety score	.979	138.725	461.872
	total anti-smoking policies subtest B	.979	45.415	636.626
4	self concept score	.982	264.525	320.877
	anxiety score	.976	139.844	485.614
	total anti-smoking policies subtest B	.977	41.295	677.108
	age	.980	12.569	748.870

Wilks' lambda table

This Table (25.14) reveals that all the predictors add some predictive power to the discriminant function as all are significant with p < .000.

The remaining tables providing the discriminant function coefficients, structure matrix, group centroids and the classification are the same as above.

Table 25.14 Wilks' lambda table

Wilks' Lambda									
						Exact F			
Step	**Number of Variables**	**Lambda**	**df1**	**df2**	**df3**	**Statistic**	**df1**	**df2**	**Sig.**
1	1	.526	1	1	436	392.672	1	436.000	.000
2	2	.406	2	1	436	317.583	2	435.000	.000
3	3	.368	3	1	436	248.478	3	434.000	.000
4	4	.358	4	1	436	194.468	4	433.000	.000

***SPSS Activity**. Please access SPSS Chapter 25 Data File B on the Web page and conduct both a normal DA and a stepwise DA using all the variables in both analyses. Discuss your results in class. The dependent or grouping variable is whether the workplace is seen as a beneficial or unpleasant environment. The predictors are mean opinion scale scores on dimensions of workplace perceptions.*

What you have learned from this chapter

Discriminant analysis uses a collection of interval variables to predict a categorical variable that may be a dichotomy or have more than two values. The technique involves finding a linear combination of independent variables (predictors) – the discriminant function – that creates the maximum difference between group membership in the categorical dependent variable. Stepwise DA is also available to determine the best combinations of predictor variables. Thus discriminant analysis is a tool for predicting group membership from a linear combination of variables.

Review questions

Qu. 25.1
The technique used to develop an equation for predicting the value of a qualitative DV based on a set of IV's that are interval and categorical is:

(a) cluster analysis
(b) discriminant regression
(c) logistic regression
(d) multivariate analysis
(e) factor analysis

Qu. 25.2
The number of correctly classified cases in discriminant analysis is given by:

(a) the cut-off score
(b) the hit rate
(c) the discriminant score
(d) the F statistic
(e) none of these

Qu. 25.3
If there are more than 2 DV categories:

(a) you can use either discriminant analysis or logistic regression
(b) you cannot use logistic regression
(c) you cannot use discriminate analysis
(d) you should use logistic regression
(e) you should use discriminant analysis

Qu. 25.4
Why would you use discriminant analysis rather than regression analysis?

Check your answers in the information above.

Now access the Web page for Chapter 25 and check your answers to the above questions. You should also attempt the SPSS activity located there.

Further reading

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons.