

Discriminant analysis

An alternative medical example of a discriminant analysis problem

For people with primary traumatic brain injury (TBI), we can discriminate among those who make a good recovery, such that they are back to work at 6 months (GROUP 1), those who make a reasonable recovery, with many functions restored but not able to return to work at 6 months (GROUP 2) and those who are very dependent or dead at 6 months (GROUP 3). A study is carried out to see how well it is possible to predict, shortly after the injury has been sustained, who will be in which group at 6 months. The study is carried out on 54 people with primary TBI who have been scored (3-15) on the revised Glasgow coma scale. Data are obtained on the following variables: (1) an EEG-derived score (EEG), (2) the coma score (COMA) and (3) a pupil reactivity to light score (PUPIL). In addition, data derived from a scan is called LATER to remind us that it is not used in the initial analysis.

Some (fabricated) data

The data on the initial test results, the scan and their final classification appears in Table 9.3 (med.discriminant.sav on the book website). There are 26, 18 and 10 patients respectively in the three outcome groups. As a minimum, we need more cases in the smallest group than there are classification variables. Initially we have three classification variables and 10 cases in the smallest group, so we can proceed. Even when we use our extra variable LATER, we shall still have enough cases.

Table 9.3
Initial test results and outcome group of 54 brain injury patients
(med.discriminant.sav)

group	EEG	coma	pupil	later
1	8	11	7	6

1	4	8	7	5
1	4	9	6	4
1	7	10	9	5
1	6	10	7	7
1	7	9	8	6
1	6	10	6	6
1	5	9	6	3
1	8	8	6	6
1	8	10	7	7
1	7	11	8	6
1	6	8	4	6
1	4	10	5	6
1	5	8	7	4
1	5	10	7	7
1	7	9	6	6
1	8	10	7	6
1	7	11	7	9
1	8	8	7	7
1	6	10	7	5
1	7	10	8	6
1	5	11	7	6
1	8	10	8	7
1	5	7	8	2
1	7	8	9	7
1	6	9	7	7
2	7	8	4	6
2	8	8	6	5
2	6	6	4	2
2	8	6	4	5
2	7	9	5	4
2	8	7	7	6
2	9	9	6	6
2	9	7	5	5
2	10	8	6	5
2	7	7	4	4
2	7	7	7	4
2	9	8	6	6
2	7	7	5	5
2	8	7	6	4
2	9	8	6	4
2	8	8	7	6
2	8	8	6	7
2	6	6	5	6
3	7	6	5	4
3	4	5	6	5
3	3	5	5	3
3	4	5	7	4
3	5	6	5	4
3	6	6	6	3
3	4	5	5	4
3	5	4	6	2
3	5	6	8	4
3	7	6	7	3

Plots of pairs of the test results do show some separation of the three groups, though no pair of test scores gives a separation as good as in our initial example with weights and heights. Figure 9.3 shows the plot of EEG and COMA, with the three groups displayed as different symbols. Note that there are fewer symbols than there are cases because quite a lot of the cases (including cases in different groups) have identical values on the two tests. For example, there are two cases from GROUP 1 and 3 cases from GROUP 2 at the point where EEG = 8 and COMA = 8.

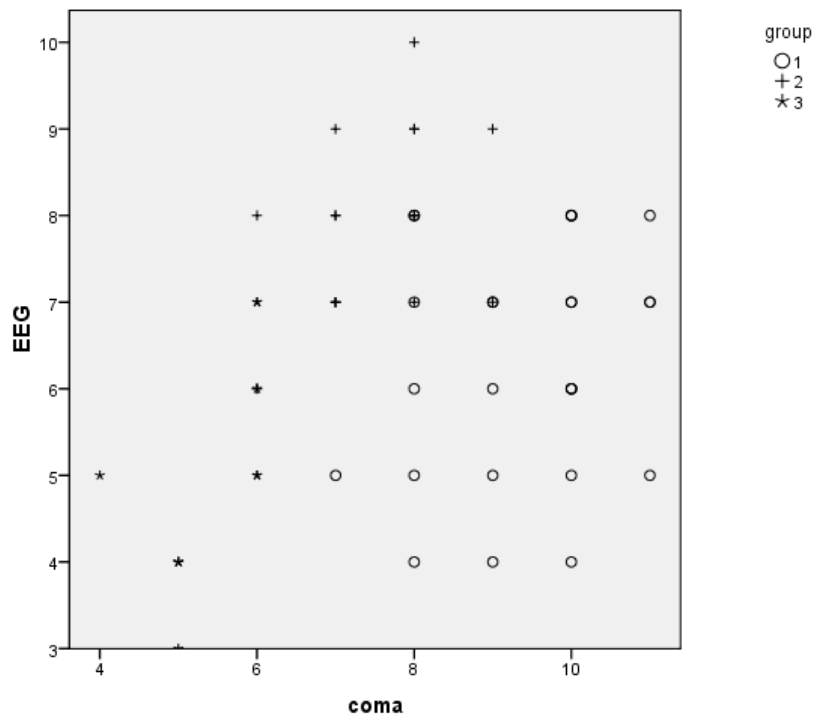
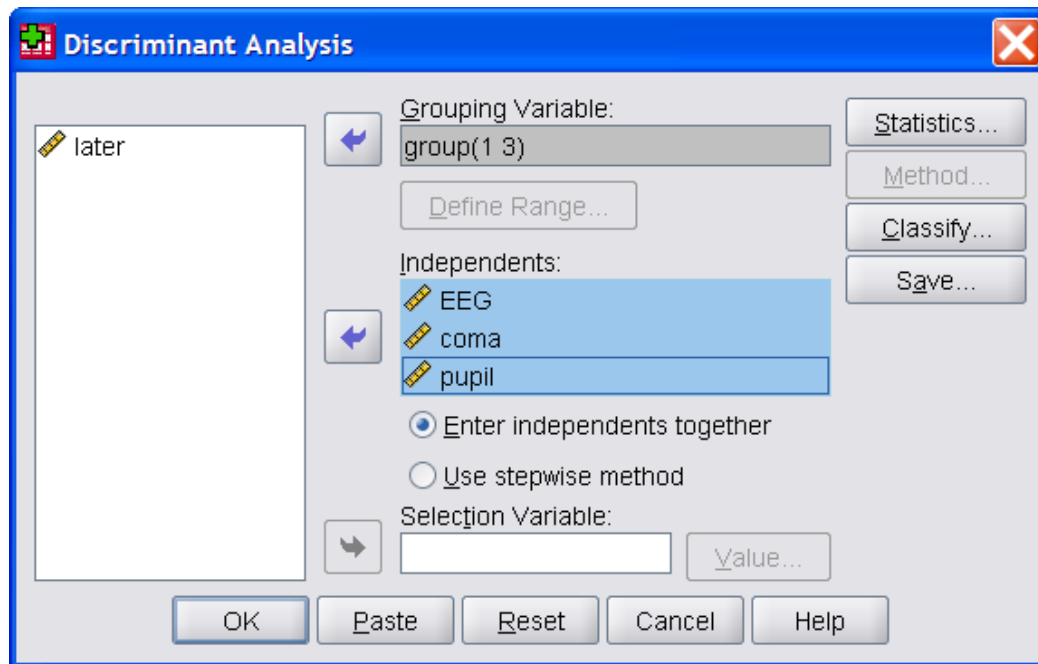


Figure 9.3. Results for EEG and COMA scores for the three outcome groups

Perhaps, using all three test results, the discriminant analysis can divide up the space of results in a way that assigns most cases correctly to their final group. We now tackle this in SPSS.

Requesting a discriminant analysis in SPSS

To do discriminant analysis in SPSS you need to enter the data for the training set just as in Table 9.1, with the correct group membership as one variable and the measurements to be used for classification as the other variables. From the menu bar we select **Analyze**, then **Classify**, then **Discriminant**. In SPSS Dialog Box 9.1, use the arrow to enter the **Grouping Variable** (GROUP in our case). Then click the **Define Range** button so that we can say how many groups there are. A small dialog box opens: enter 1 as **Minimum** and 3 as **Maximum** since we have three groups coded as 1, 2 and 3. Click **Continue** to return to the main dialog box. Then use the arrow to put the three tests into the **Independents** box and we have the Discriminant Analysis dialog box as shown.



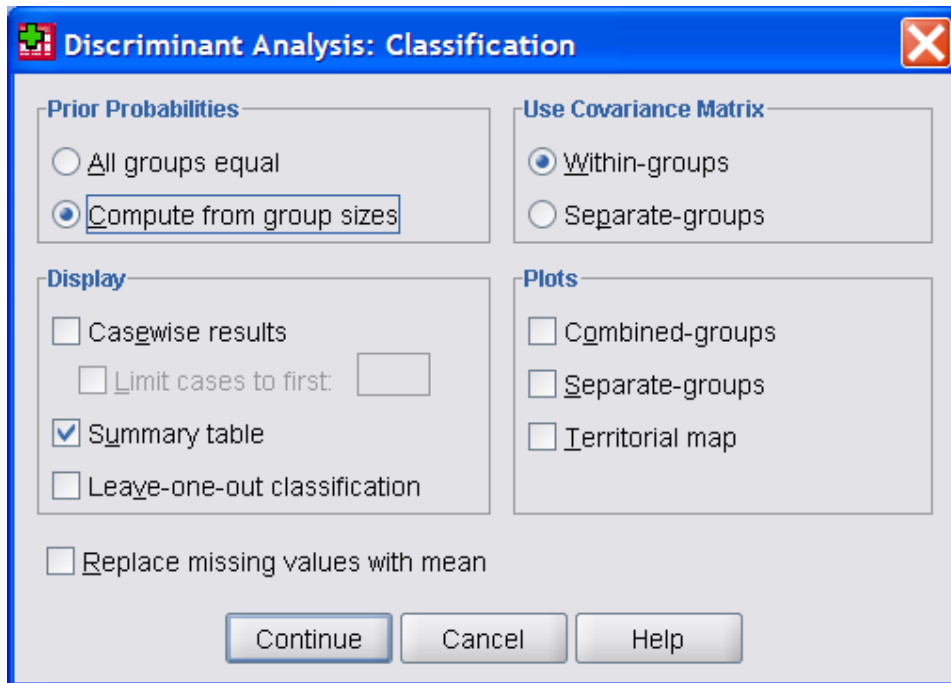
SPSS Dialog Box 9.1 Starting a discriminant analysis of brain injury data

The default options only provide the information needed to use discriminant analysis for separation, and to get what we need for classification we need to click the **Statistics** button. In the dialog box that opens up (not reproduced here), click **Fisher's**

in the **Function Coefficients** group, since this is what we need for discrimination. It may also be useful to see the **Means** of the test scores for each group, so click that box in the **Descriptives** group. This is all we need for now so click **Continue** to return to SPSS Dialog Box 9.1.

Near the bottom of SPSS Dialog Box 9.1 you see radio buttons for the options **Enter independents together** and **Use stepwise method**. The default, **Enter independents together**, will immediately try the classification using all the variables in the **Independents** list. An alternative approach, especially useful if you have a long list of variables, only some of which are likely to be useful for classification, is to use a *stepwise* method and we consider this option later. For now, we accept the default, **Enter independents together**.

Now, if we click the **Classify** button we get SPSS Dialog Box 9.2. The default choice for **Prior Probabilities** is **All groups equal**. This means that, if we had no classification variables, a randomly chosen case would be equally likely to belong to any of the groups. But if our groups were of unequal sizes, in the absence of any information on classification variables, the probability of a randomly chosen case belonging to a particular group would be proportional to the size of that group. Our groups are not all the same size so we choose **Compute from group sizes**. This will slightly improve our probability of successful classification. From the **Display** group we click **Summary table**. The summary table will give the results of classifying into groups 1 to 3 using the three tests. We accept the defaults for the other choices and click **Continue** to return to SPSS Dialog Box 9.1.



SPSS Dialog Box 9.2. Classification dialog box

Now, if we click the **Save** button, and then click the **Predicted group membership** box, SPSS will add a column to our data window, with the group assigned to each case by the discriminant functions. This makes it easy to identify the particular cases that are misclassified. Click **Continue** and **OK** to get the discriminant analysis.

Understanding the output and calculating the discriminant functions

There is a lot of information in the output window, much of it concerned with using discriminant analysis for separation, which does not interest us at the moment. The first table in the output just shows whether any cases were omitted because of missing data. We had no missing data, so all our cases were included in the analysis. Then there is a table showing the mean scores on the three tests for the members of each of our groups. This may be of interest if we are trying to understand the differences among the groups more clearly. However, these tables have not been reproduced here.

Now if we go right to the end of the output, the last two tables are shown in SPSS Output 9.1. Look at the last table first. This is the summary table we requested in SPSS Dialog Box 9.2. It shows the results of classifying the patients into our three groups using the data on the EEG, COMA and PUPIL scores. Of the actual GROUP 1 members, 22 were correctly classified into GROUP 1, three were put into GROUP 2 and one into GROUP 3. All of the GROUP 2 members were correctly classified. All but one of the GROUP 3 members were also correctly classified, with just one being put in GROUP 2. Overall, just over 90% were correctly classified using the data on the three tests. This could be part of the process of deciding which interventions may be appropriate. But how would you actually do the classification for a new patient for whom you have the test results? Look at the first table in SPSS Output 9.1. The columns give the coefficients for three discriminant functions, one for each group. For a new patient with test results of 6 on EEG, 10 on COMA and 8 on PUPIL, we calculate each discriminant function and assign the patient to the group with the highest value, as follows:

Function for group 1: $0.583*6+8.266*10+4.304*8-56.335=64.255$

Function for group 2: $2.615*6+5.805*10+2.701*8-40.374=54.974$

Function for group 3: $0.888*6+4.201*10+4.072*8-27.464=52.450$

So this patient is assigned to GROUP 1: most likely they will be able to return to work within six months. These calculations are quite laborious but SPSS will do them for us, and we consider this next.

Classification Function Coefficients

	group		
	1	2	3
EEG	.583	2.615	.888
coma	8.266	5.805	4.201
pupil	4.304	2.701	4.072
(Constant)	-56.335	-40.374	-27.464

Fisher's linear discriminant functions

Classification Results^a

	group	Predicted Group Membership			
		1	2	3	Total
Original	Count 1	22	3	1	26
	2	0	18	0	18
	3	0	1	9	10
% 1	1	84.6	11.5	3.8	100.0
	2	.0	100.0	.0	100.0
	3	.0	10.0	90.0	100.0

a. 90.7% of original grouped cases correctly classified.

SPSS Output 9.1. Part of the output from a discriminant analysis of reading data

Classifying new cases using SPSS

All you need to classify new cases is the first of the tables in SPSS Output 9.1, a calculator, the test results for the new cases, and a lot of patience. Alternatively, you can get SPSS to do the calculations using the **Selection Variable** box in SPSS Dialog Box 9.1 as follows. Add the test results for the 10 new cases shown in Table 9.4 to the bottom of the SPSS data window. The GROUP variable will be missing for these cases.

Table 9.4

Test results for new cases (med.discriminant.newcases.sav)

EEG	coma	pupil
7	10	8
4	8	9
7	7	5
7	8	7
6	9	5
5	6	8
6	10	6
5	5	7
4	6	5
6	8	7

We also need a new variable that can be used to select the 54 cases in the training set for the discriminant analysis, then SPSS will use the discriminant functions to assign

the new cases to their predicted groups. We could call the new variable CHOOSE and give it the value 1 for the cases in the training set and 2 for the new cases. In SPSS Dialog Box 9.1, select the CHOOSE variable and move it into the **Selection Variable** box at the bottom using the arrow. Then press the **Value** button and insert 1. The dialog box now has CHOOSE = 1 in the **Selection Variable** box at the bottom.

Remember to use the Save button to get the predicted group membership for all cases, including the new ones. The predicted group membership is in a new column in the data sheet, and, if you look at the output (not shown here) you will see that for the new cases the discriminant functions assign them to groups 1, 1, 2, 1, 1, 3, 1, 3, 3, and 1 respectively.

Estimating the probability of misclassifying new cases

When we did our discriminant analysis on the 54 cases from the original study, the classification using the three test results was correct in 90.7% of the cases. When we apply the results to new cases we are unlikely to achieve quite such a high proportion of correct assignments because the original data, and any future data, will contain some random variation as well as the relationships among the test scores and the group membership. When we perform the discriminant analysis on any set of data, the results will be best for that dataset, including its random components. The next set of data will have different random elements and classification will have a slightly increased probability of being wrong. There are several ways to deal with this, all of which are used in practice.

First, you can just go ahead and apply the discriminant functions to new data, then, when actual group membership is eventually ascertained, you can find what

proportion of classifications your discriminant functions got right. This is a great method if it is available to you. However, it can be a problem if ascertaining actual group membership for your new cases is difficult or expensive, and not just a matter of waiting until later results are available. This could apply in our example, because the predicted outcome may have some influence on which interventions are selected.

Another approach is available if you have a large body of data to start with. In this case you could randomly assign each case to one of two datasets. You would then use one of them as the training set to derive the discriminant functions, and the other one as the *cross-validation set* to calculate misclassification probabilities. The snag about this method is that you get better estimates of your discriminant functions by using a bigger dataset to calculate them, so it seems wasteful to use only half of the available data for this task just so that you can get a better estimate of the proportion of misclassifications. However, if you want to do this, SPSS makes it easy. All you need to do is make a new variable (you could call it XVAL for cross-validation) and give this new variable the value 1 for the cases you want in the training set, and 2 for all the others. You would make the assignment to training set and validation set at random. Then you would use the **Selection Variable** box just as we did for classifying new cases, only this time you would be classifying the members of the cross-validation set (coded 2). You would end by clicking **Save** and then **Predicted group membership** in order easily to see how many were misclassified both in the training set and in the validation set.

The last approach is called *leave-one-out* classification in SPSS, though some packages use the term *cross-validation*. Here the full dataset is used to calculate the

discriminant functions, then each case is classified using the discriminant functions calculated from all cases other than the one being classified. The probability of misclassification is then calculated from these results. To get this done, we tick the **Leave-one-out** classification box in SPSS Dialog Box 9.2. SPSS Output 9.2 shows the results for our example. From this we would expect to get 87% (see below table) of new cases correctly classified using the three tests. The snag about this approach is that there is some evidence that it still gives somewhat optimistic estimates of misclassification rates.

Classification Results^{b,c}

		Predicted Group Membership				
		1	2	3	Total	
Original	Count	1	22	3	1	26
		2	0	18	0	18
		3	0	1	9	10
	%	1	84.6	11.5	3.8	100.0
		2	.0	100.0	.0	100.0
		3	.0	10.0	90.0	100.0
Cross-validated ^a	Count	1	22	3	1	26
		2	1	16	1	18
		3	0	1	9	10
	%	1	84.6	11.5	3.8	100.0
		2	5.6	88.9	5.6	100.0
		3	.0	10.0	90.0	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. 90.7% of original grouped cases correctly classified.

c. 87.0% of cross-validated grouped cases correctly classified.

SPSS Output 9.2. Leave-one-out classification for the brain injury data

Assumptions used in classification by discriminant analysis

The classification method we have been using in this chapter has the optimal property that it minimises the total probability of misclassification, but only if one assumption is satisfied. That assumption is that the *covariance matrices* of the independent (classification) variables are the same for all groups. This requires the IVs to be measured on an interval scale so dichotomous variables are not possible. However, even if that assumption is not satisfied, you may still achieve a useful classification, and you can estimate the misclassification probabilities for new cases using one of the methods described in the previous section.

You can check the assumption of equal covariance matrices using Box's M test for equality of covariance matrices. Click the Statistics button and then click **Box's M test** in the **Descriptives** group. The result appears at the start of the discriminant analysis, just after the table of group means. For our example the test statistic is shown in SPSS Output 9.3. We see that the result is not significant ($p > 0.05$) so we do not reject the hypothesis that the three covariance matrices are equal.

Box's M		8.918
F	Approx.	.664
	df1	12.000
	df2	3936.777
	Sig.	.787

Tests null hypothesis of equal population covariance matrices.

SPSS Output 9.3. Box's test of equality of covariance matrices

Other parts of the output

After the first few tables showing whether cases have been omitted and displaying group means and the Box test for equality of covariance matrices, if requested, the discriminant analysis begins with a summary of the *canonical discriminant functions*. This information is not of direct importance for our purpose of classification, but the following paragraph gives a brief account of the tables, though we have not reproduced them here.

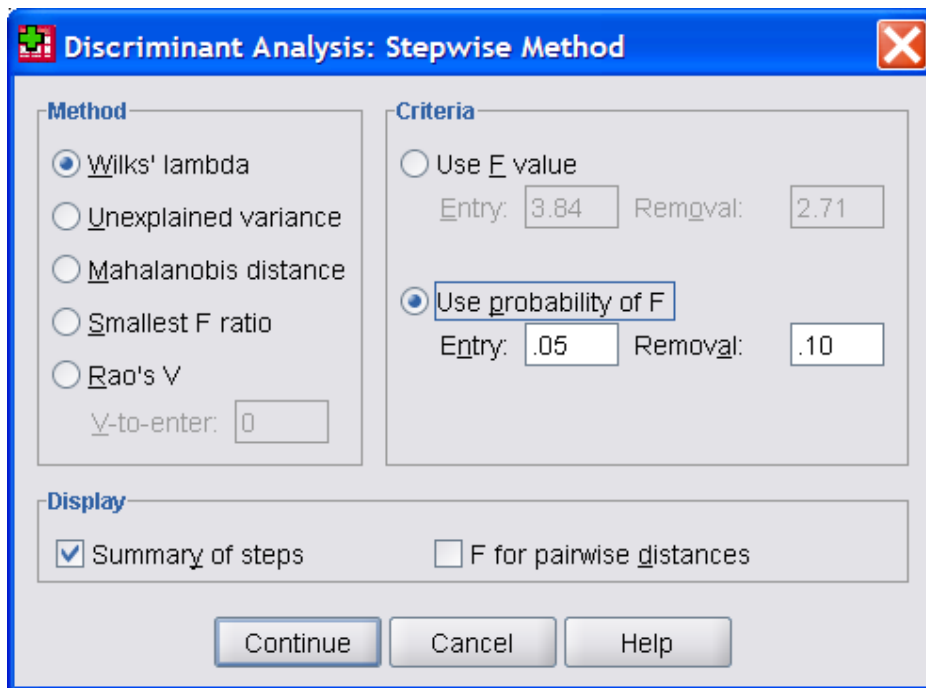
The idea behind the canonical discriminant functions is to find a set of linear combinations of the IVs that will separate the groups as well as possible, while being fewer in number than the IVs. In other words, is it possible to achieve good separation of the groups using fewer dimensions than the number of IVs? The first canonical discriminant function is the linear combination of IVs that separates the groups as much as possible, the second is the linear combination that achieves the best separation in a direction orthogonal to the first, and so on. Often, you can find a set of

linear combinations of independents that is fewer than the total number of independents, which will achieve good separation of the groups. The Wilks' lambda table shows whether the canonical discriminant functions achieve significant separation (look at the 'Sig' column), but the only relevance of this for our purpose of classification is that you are unlikely to achieve a useful classification unless there is significant separation. But we discussed above more direct ways to find out whether your classification is likely to be useful for new cases. The *eigenvalue* table shows the proportion of the within groups variance that is accounted for by each canonical discriminant function (each eigenvalue divided by the total of eigenvalues gives the proportion of variance accounted for by that canonical discriminant function). The coefficients for the canonical discriminant functions are given in another table, and also the correlation of each IV with each canonical discriminant function.

Using a stepwise method

Since there are correlations among our three tests (look ahead to Table 9.5) we may wonder whether they all do contribute significantly to the classification into our three groups. We can check this by using the stepwise method instead of entering all our independent (classification) variables together. Look again at SPSS Dialog Box 9.1. This time we will click the radio button **Use stepwise method**. This is rather like doing a stepwise regression, and is especially useful in similar circumstances, namely when we have rather a long list of possible classification variables and it is unlikely that all will make a useful contribution to a set of discriminant functions. We would like to find the best subset, or else something close to that. By taking a stepwise approach we ask SPSS to choose first the single variable that gives the best classification into our groups. Then it checks to see that it does achieve a significant classification (i.e., it does better than just assigning the cases to the groups in the

correct proportions but otherwise at random). Then it looks at the remaining variables and adds the one that gives the biggest improvement. It checks the two variables now in the discriminant functions and makes sure each makes a significant contribution in the presence of the other. At each step we see whether another variable can be added that will make a significant improvement, and whether any previous ones can be removed. The process stops when no more variables can be added or removed at the level of significance we are using. To choose the method by which variables will be added and removed, we click the **Method** button in SPSS Dialog Box 9.1 to obtain SPSS Dialog Box 9.3.



SPSS Dialog Box 9.3. Using the stepwise method

The default method uses **Wilks' lambda**, a *likelihood ratio* method. It has the advantage that it does not depend on the scale parameters of the variables and we will use it. The F values offered as defaults for entering and removing variables are the 5% and 10% levels for F with 1 and infinite degrees of freedom. If we have a fairly large

data set this will approximate using the F probability values 5% and 10% for entry and removal. As probabilities of 5% and 10% would be typical, we may as well use the default probabilities and we click the radio button, **Use probability of F**.

If we apply this process to our example data, all three test scores are entered into the discriminant functions and they remain there, so EEG, COMA and PUPIL do all contribute significantly to successful classification. The table of stepwise results appears at the beginning of the analysis and is shown in SPSS Output 9.4.

Variables Entered/Removed^{a,b,c,d}

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	coma	.293	1	2	51.000	61.503	2	51.000	.000
2	EEG	.168	2	2	51.000	36.075	4	100.000	.000
3	pupil	.129	3	2	51.000	29.182	6	98.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

- a. Maximum number of steps is 6.
- b. Maximum significance of F to enter is .05.
- c. Minimum significance of F to remove is .10.
- d. F level, tolerance, or VLN insufficient for further computation.

SPSS Output 9.4. Stepwise discrimination using EEG, COMA and PUPIL: variables Entered/Removed

Now at last we use the final column of data from Table 9.3, the one called LATER, to remind us to ignore it until now. This column actually shows a score derived from a scan, but you can see from the table of correlations in Table 9.5 that it is correlated with all of EEG, COMA and PUPIL.

Table 9.5
Correlations among EEG, COMA and PUPIL and LATER

	EEG	coma	pupil
coma	0.259		
pupil	0.045	0.415	
later	0.356	0.647	0.278

Perhaps including it as an extra IV will not improve the classification. In fact if we add it to the list of independents and use the **Stepwise** method, the results are exactly as before, the extra variable **LATER** never gets added to the discriminant functions. If we include it and use the **Enter independents together** method, we get the summary table shown in SPSS Output 9.5.

Classification Results^a

		group	Predicted Group Membership			
			1	2	3	Total
Original	Count	1	22	3	1	26
		2	0	17	1	18
		3	0	1	9	10
	%	1	84.6	11.5	3.8	100.0
		2	.0	94.4	5.6	100.0
		3	.0	10.0	90.0	100.0

a. 88.9% of original grouped cases correctly classified.

SPSS Output 9.5. Classification including an extra (redundant) variable

We can see that we actually achieve *less* successful classification than we did with just the three tests, all of which contributed useful information. The inclusion of redundant variables can reduce the proportion of correctly classified cases. It is therefore worth using the stepwise method, especially if you suspect that some of your variables may be redundant.