CME Article

Biostatistics 303. Discriminant analysis

Y H Chan

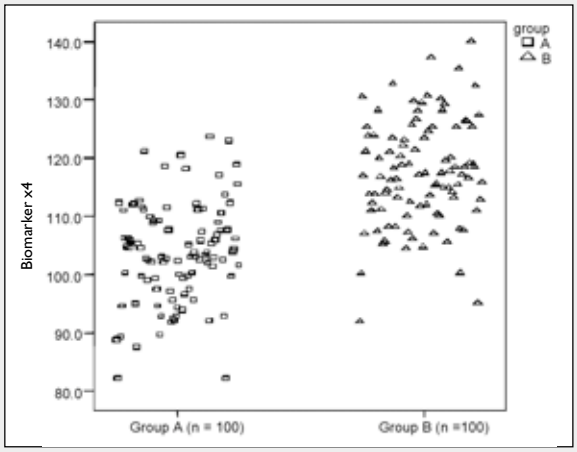
In this article, it was planned that we shall discuss Discriminant and Cluster analysis. While preparing the discussions for both topics, there was an overwhelming large amount of information and thus we shall concentrate on Discriminant analysis only and leave Cluster analysis to Biostatistics 304.

Discriminant analysis (DA) was the traditional statistical technique used for differentiating groups (categorical dependent variable) when the independent variables were quantitative. Consider the situation where a researcher hypothesised that four quantitative bio-markers, x1 to x4, could be used to differentiate two groups (A & B). Table I shows the differences between the two groups for each biomarker using 2-Sample t-test (after checking for normality and homogeneity of variance assumptions).

Table I. Mean differences (2 Sample t) between groups A and B.

Biomarker	Group	Mean (sd)	p-value	Total mean (sd)
x1	A	65.25 (3.79)	0.663	65.12 (3.67)
	B	65.00 (3.57)		
x2	A	44.59 (4.07)	0.660	44.46 (3.92)
	B	44.34 (3.79)		
x3	A	7.01 (3.09)	0.056	7.43 (3.11)
	B	7.85 (3.10)		
x4	A	103.72 (8.50)	<0.001	110.55 (11.09)
	B	117.37 (8.98)		

Fig. 1 Distribution of biomarker x4 for groups A and B.



Faculty of Medicine
National University of Singapore
Block MD11
Clinical Research Centre #02-02
10 Medical Drive
Singapore 117597

Y H Chan, PhD
Head
Biostatistics Unit

Correspondence to:
Dr Y H Chan
Tel: (65) 6874 3698
Fax: (65) 6778 5743
Email: medcyh@nus.edu.sg

Fig. 1 shows the distribution of x4 for both groups and although there is a significant difference ($p<0.001$), the demarcation is not obvious! What then is a good cut-off to differentiate the 2 groups? A recommendation is to use the total mean of x4 (=110.55); group A<110.55 and group B ≥ 110.55 giving a total accuracy of 78% with 77% and 79% accuracies for groups A and B, respectively (Table II). This may not be the optimal cut-off (giving the best accuracy) – an ROC analysis[1] should be performed.

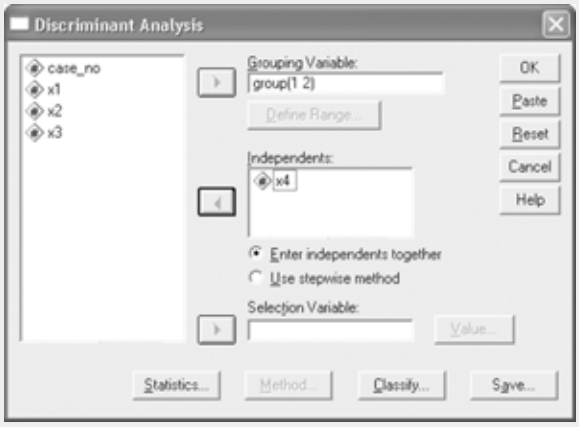
Table II. Accuracy with cutoff x4 = 110.55.

Group * Predicted group with cutoff = 110.55 Cross-tabulation

			Predicted Group with cutoff = 110.55		
			A	B	Total
group	A	Count	77	23	100
		% within group	77.0%	23.0%	100.0%
	B	Count	21	79	100
		% within group	21.0%	79.0%	100.0%
Total		Count	98	102	200
		% within group	49.0%	51.0%	100.0%

How does Discriminant analysis (DA) "discriminate" between the two groups? In SPSS, go to ***Analyze, Classify, Discriminant*** to get Template I.

Template I. Discriminant analysis definition.



Put the variable group (coded as 1=A, 2=B) into the Grouping Variable box; define range: minimum = 1 and maximum = 2 and put x4 into the Independents box. Click the Classify folder. In Template II, leave the Prior Probabilities to be "All groups equal" (when we are unsure that the sample is a representative of the population; otherwise use the "Compute from group sizes" option), use the Within-groups Covariance Matrix and tick the Summary table option which shows that the total accuracy of x4 to differentiate the 2 groups is 78% (Table IIa). For 1-variable only, DA uses the total mean (of x4 = 110.55) as the cutoff to discriminate between the two groups.

Template II. DA Classification options.

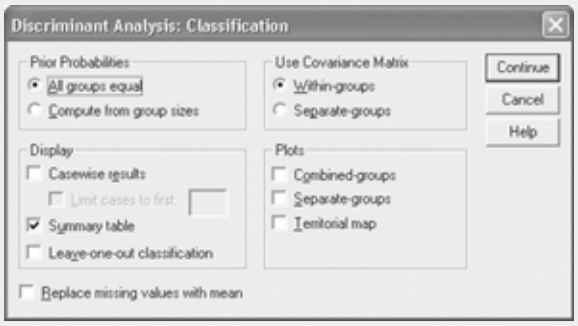


Table IIa. DA Accuracy of using biomarker x4.

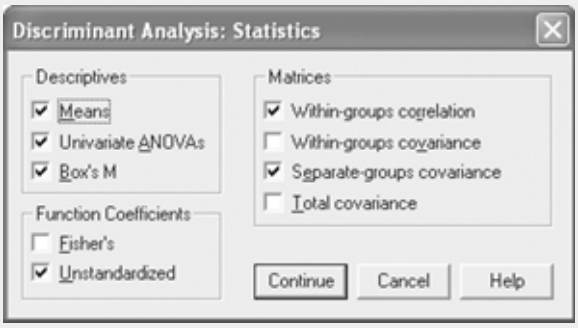
Classification Results[a]

			Predicted Group Membership		
		group	A	B	Total
Original	Count	A	77	23	100
		B	77.0%	23.0%	100.0%
	%	A	21	79	100
		B	21.0%	79.0%	100.0%

[a] 78.0% of original grouped cases correctly classified.

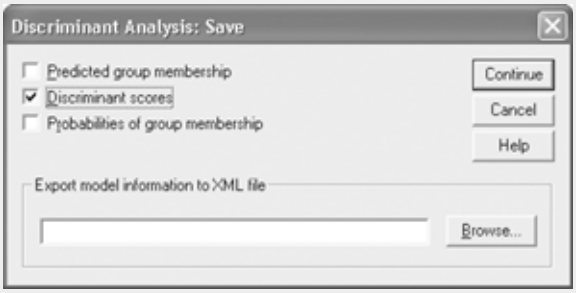
We can include the other biomarkers x1-x3 in DA to see whether the accuracy is enhanced. In Template I, now include x1-x3 to the Independents box. Click on the Statistics folder and check on the options shown in Template III.

Template III. DA Statistics options.



Click Continue. In Template I, click on the Save folder; check the Discriminant scores option (Template IV). Leave the Summary Table in Template II as checked.

Template IV. DA Save options.



The relevant outputs are shown in Tables IIIa - IIIl.

Table IIIa (obtained by ticking the Means option in Template III) gives the descriptive statistics of x1 – x4 by group.

Table IIIa. Descriptive statistics.

Group statistics

				Valid N (listwise)	
Group		Mean	Std. deviation	Unweighted	Weighted
A	x1	65.249	3.7931	100	100.000
	x2	44.586	4.0669	100	100.000
	x3	7.005	3.0875	100	100.000
	x4	103.722	8.4994	100	100.000
B	x1	65.000	3.5692	100	100.000
	x2	44.341	3.7932	100	100.000
	x3	7.847	3.1025	100	100.000
	x4	117.373	8.9823	100	100.000
C	x1	65.125	3.6757	200	200.000
	x2	44.463	3.9245	200	200.000
	x3	7.426	3.1159	200	200.000
	x4	110.548	11.0859	200	200.000

Table IIIb (obtained by ticking the Univariate ANOVAs option in Template III) tests which biomarker is statistically different between the two groups (exactly the same as Table I). A key assumption of DA is that the independent variables should be from a multivariate normal distribution. Thus, it is necessary to check the normality of the variables (already checked for x1 – x4) before using DA.

Table IIIb. DA ANOVA tests.

Tests of equality of group means					
	Wilks' Lambda	F	df1	df2	Sig.
x1	.999	.229	1	198	.633
x2	.999	.194	1	198	.660
x3	.982	3.701	1	198	.056
x4	.619	121.860	1	198	.000

Another key assumption of DA is that the independent variables should not be highly correlated, see Table IIIc (Within-groups correlation, Template III).

Table IIIc. Correlation matrix.

Pooled within-group matrices					
		x1	x2	x3	x4
Correlation	x1	1.000	.293	-.010	-.272
	x2	.293	1.000	-.029	.192
	x3	-.010	-.29	1.000	.076
	x4	-.272	.192	.076	1.000

Table IIId. Covariance matrix.

Pooled within-group matrices					
Group		x1	x2	x3	x4
A	x1	14.388	4.572	.025	-6.169
	x2	4.572	16.540	-1.411	10.947
	x3	.025	-1.411	9.533	.295
	x4	-6.169	10.947	.295	72.240
B	x1	12.739	3.906	-.251	-11.372
	x2	3.906	14.388	.696	2.292
	x3	-.251	.696	9.625	3.815
	x4	-11.372	2.292	3.815	80.681

Table IIIe. Box's M test.

Test results		
Box's M		7.683
F	Approx.	.752
	df1	10
	ddf2	187429.482
	Sig.	.676

Table IIId (Separate-groups covariance, template III) shows the covariance matrix with Table IIIe testing the assumption of equal covariance (Box's M test, template III). We want the p-value (in this case Sig 0.676) not to be significant (>0.05). Unequal covariance causes observations to be "overclassified" to the groups with a larger covariance.

Tables IIIa – IIIe check the various assumptions of DA which if violated may affect the accuracy of the classification. Tables IIIf – IIIk show the "usefulness" of DA for this study.

In Template IV, we asked for the Discriminant scores to be saved. SPSS creates a new variable Dis1_1 which is a calculated score based on the Unstandardised canonical discriminant function coefficients (Table IIIf) where

$$\text{Discriminant score} = -16.164 + 0.097(x1) - 0.088(x2) + 0.023(x3) + 0.123(x4)$$

with Table IIIg showing the mean of the Discriminant score for each group. The assignment of the Predicted Group membership (see Template IV), a new variable Dis_1 will be created, will assign Discriminant scores ≥0 to group B and negative scores to group A.

Table IIIf. Canonical discriminant function coefficients.

Canonical discriminant function coefficients	
	Function
	1
x1	.097
x2	-.088
x3	.023
x4	.123
(Constant)	-16.164

Unstandardised coefficients

Table IIIg. Means of the discriminant scores.

Functions at group centroids	
	Function
Group	1
A	-.849
B	.849

For a 2-group analysis, only one function is needed to discriminate, thus 1 eigenvalue (which will explain 100% of the variance, Table IIIh) is given. The Canonical correlation measures the association between the Discriminant scores and the groups; a high value (near 1) shows that the function discriminates well.

Wilk's Lambda (Table IIIi) shows the proportion of the total variance (57.9%) in the Discriminant scores not explained by differences among groups. A small Lambda value (near 0) indicates that the group's mean Discriminant scores differ. The Sig ($p<0.001$) is for the Chi-square test which indicates that there is a highly significant difference between the groups' centroids. Tables IIIh & IIIi give an indication on how discriminating this DA model is but provides little information regarding the accuracy.

Table IIIh. Canonical correlation.

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical correlation
1	.729[a]	100.00	100.0	.649

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

Table IIIi. Wilk's Lambda.

Wilks' Lambda				
Test of function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.579	107.267	4	.000

Table IIIj shows the impact of each variable on the discriminant function after "standardising" – putting each variable on the same platform since each variable may have different units. Here x4 has the greatest impact which is also reflected in Table IIIk which shows the correlation (in order of importance) of each variable with the discriminant function.

Table IIIj. Impact of each variable.

Standardised Canonical discriminant function coefficients	
	Function
	1
x1	.356
x2	-.346
x3	.072
x4	1.077

Table IIIk. Correlation of each variable to the Discriminant function.

Structure matrix	
	Function
	1
x4	.919
x3	.160
x1	-.040
x2	-.037

Table IIIl shows that there is an improvement in the accuracy of the model with x1-x4 (81.5%) compared to x4 alone (78%) – note that it does not mean that as more variables are included in DA, the accuracy will improve!

Table IIIl. Classification table with biomarkers x1-x4.

Classification results[a]					
			Predicted group membership		
		group	A	B	Total
Original	count	A	83	17	100
		B	20	80	100
	%	A	83.0	17.0	100.0
		B	20.0	80.0	100.0

[a] 81.5% of original grouped cases correctly classified.

Question: is this discriminatory power of the classification statistically better than chance (50% assignment)? We can use Press's Q statistic to compare with the critical value (= 6.63) from the Chi-square distribution with 1 degree of freedom.

$$\text{Press's Q statistic} = \frac{[N-(nK)]^2}{N(K-1)}$$

where N = total sample size
n = number of observations correctly classified
K = number of groups

For the above example, N = 200, n = 163 and K = 2, giving Press's Q = 79.38>6.63; thus the results exceed the classification accuracy expected by chance at a statistically significant level. However, one must be careful as Press's Q is adversely affected by sample size.

Another technique is to use a Binomial test with p = 0.5 on the accuracy obtained. This is to compare the 81.5% success to a 50% chance assignment. Before we can perform the analysis, we have to create a new variable (let us call it "correct") to specify whether the classification is correct for that case. We can use the following syntax (group & Dis_1 are the actual and predicted classifications respectively; the symbol "~=" means "not-equal"):

```
IF (group = Dis_1) correct = 1.
EXECUTE.
IF (group ~= Dis_1) correct = 0.
EXECUTE.
```

In SPSS go to Analyze, Nonparametric Tests, Binomial to get Template V. Put the variable "correct" in the Test Variable list, leave the Test Proportion = 0.5. Table IV shows that the accuracy of 81.5% is statistically different from a 50-50% chance of classification.

Template V. Binomial test.

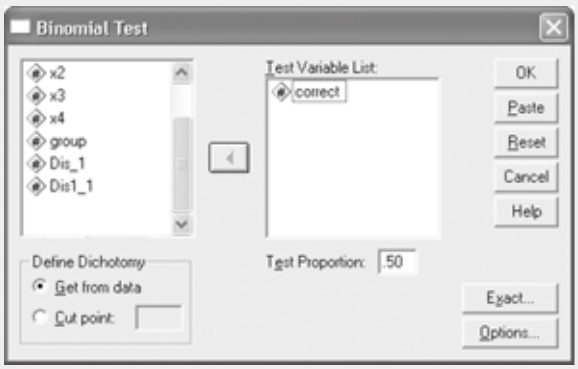


Table IV. Binomial test results.

Binomial test

		Category	N	Observed prop.	Test prop.	Asymp. sig. (2-tailed)
Correct	Group 1	1.00	163	.82	.50	.000[a]
	Group 2	.00	37	.19		
	Total		200	1.00		

[a] Based on Z Approximation.

VALIDATION OF THE RESULTS

The above example shows a "balanced" accuracy for both groups (total = 81.5%, A = 83%, B = 80%). There are situations where the total accuracy is 70% with A = 90% but B = 50% only. One has to assess the models "clinically" to determine its usefulness.

The results obtained from DA may only be applicable to the sample used. We want a discriminant model which has both external and internal validity. DA provides a leave-one-out classification (see Template II) as a cross-validation check on the propensity to inflate the accuracy if only 1 sample is being used. Table V shows the leave-one-out cross-validation which still gives a 81.5% accuracy - which may still be overly optimistic!

Table V. Leave-one-out cross-validation.

Classification results[b,c]

			Predicted group membership		
		Group	A	B	Total
Original	Count	A	83	17	100
		B	20	80	100
	%	A	83.0	17.0	100
		B	20.0	80.0	100.0%
Cross-validated[a]	Count	A	83	17	100
		B	20	80	100
	%	A	83.0	17.0	100
		B	20.0	80.0	100.0%

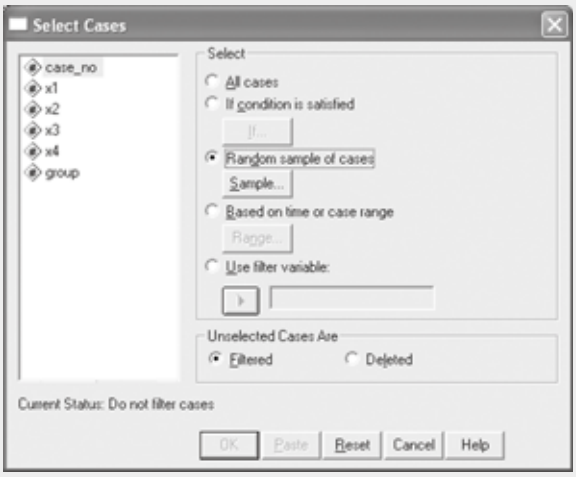
[a] Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

[b] 81.5% of original grouped cases correctly classified.

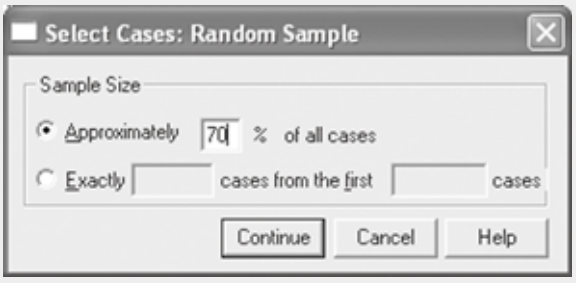
[c] 81.5% of cross-validated grouped cases correctly classified.

Another cross-validation procedure it to divide the dataset into two samples (a test sample and a retest/hold sample) which means that one needs a sizeable number of cases. To perform this procedure, in SPSS, go to Data, Select Cases – in Template VI, tick the Random sample of cases option, click on Sample to get Template VII. Let us say we take approximately 70% of the cases as the test sample – a new variable filter_$ (having 1 or 0) will be created.

Template VI. Choosing a Random sample.



Template VII. Specifying the percentage of cases to be randomly chosen.



Before performing DA, go back to ***Data, Select Cases*** – click on All cases (template VI). Then do the usual steps for DA but now put the variable filter_$ in the Selection variable, click on Value and enter 1 (see Template VIII).

Template VIII. DA on test sample.

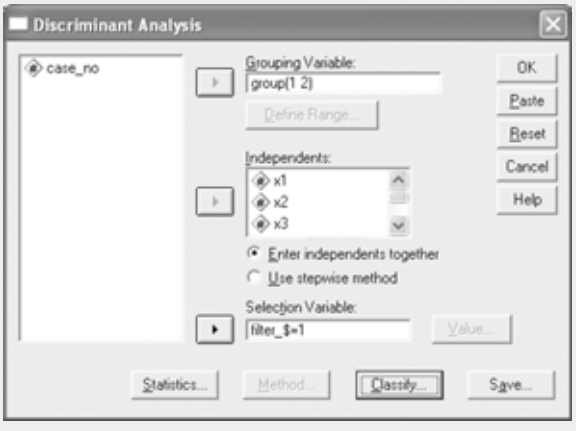


Table VI shows the test-retest results with the leave-one-out classification option invoked (this will not be performed for the retest sample). The three results are consistent with that when the whole sample was used. Thus our discriminating equation from the whole sample could be used to "discriminate" new cases. This test-retest could be performed several times!

Table VI. Test-retest results.

Classification results[b,c,d]

				Predicted group membership		
			Group	A	B	Total
Cases selected	Original	Count	A	62	12	74
			B	15	59	74
		%	A	83.8	16.2	100.0
			B	20.3	79.7	100.0
	Cross-validated[a]	Count	A	62	12	74
			B	16	58	74
		%	A	83.8	16.2	100.0
			B	21.6	78.4	100.0
Cases not selected	Original	Count	A	21	5	26
			B	4	22	26
		%	A	80.8	19.2	100.0
			B	15.4	84.6	100.0
	Cross-validated[a]	Count	A			
			B			
		%	A			
			B			

[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] 81.8% of selected original grouped cases correctly classified.

[c] 82.7% of unselected original grouped cases correctly classified.

[d] 81.1% of selected cross-validated grouped cases correctly classified.

For completeness, we can ask for the Fisher's function coefficients (Template III) – usually not necessary – which gives the weights of each biomarker for the individual group (see Table VII). We can calculate the Fisher's score for each group (manually) and assign the classification of a new case to the group with the higher value.

Table VII. Fisher's discriminating functions.

Classification function coefficients

	Group	
	A	B
x1	5.982	6.145
x2	.397	.248
x3	.388	.427
x4	1.998	2.207
(Constant)	-309.662	-337.117

Fisher's linear discriminant functions.

MULTIPLE GROUPS CLASSIFICATION

For a n-group (n>2) discrimination, DA provides n -1 discriminating functions. We shall discuss for n = 3 using four biomarkers, x1-x4. Since there are three groups, two discriminating functions will be given. We shall only highlight the tables which are "different" from the 2-group analysis.

Table VIIIa shows that 1st function has a high canonical correlation (0.919) and explains 99.5% of the variance. Is it worth keeping the 2nd function? Table VIIIb shows that using both functions (1 through 2), the hypothesis that the means of both functions are equal in the 3 groups could be rejected. Similarly, after removing function 1, function 2 (p = 0.036) was still significant - thus it is worthwhile to keep both functions.

Table VIIIa. DA 3-group canonical correlation.

Eigenvalues				
Function	Eigenvalue	% of variance	Cumulative %	Canonical Correlation
1	5.461[a]	99.5	99.5	.919
2	.028[a]	.5	100.0	.166

[a] First 2 canonical discriminant functions were used in the analysis.

Table VIIIb. DA 3-group Wilk's Lambda.

Wilks' Lambda				
Test of function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.150	576.672	8	.000
2	.972	8.528	3	.036

Table VIIIc. DA 3-group impact of each variable.

Standardised canonical discriminant function coefficients		
	Function	
	1	2
x1	-1.675	.833
x2	1.885	.180
x3	..049	-.027
x4	-.098	.048

Table VIIId. DA 3-group canonical discriminant function coefficients.

Canonical discriminant function coefficients		
	Function	
	1	2
x1	-.484	.241
x2	.511	.049
x3	.009	-.005
x4	-.029	.014
(Constant)	-.395	-23.919

Unstandardised coefficients.

Table VIIIc shows the impact of each variable on the two functions. Tables VIIId and VIIIe give the two Discriminating functions and the mean discriminant score of each function, with the model accuracy given in Table VIIIf. Figure II is obtained by ticking the Combine-groups under the Plots option in Template II. Fig. 3 is the territorial map (edited-reduced version presented – SPSS provides a text version of this map which is not graphical-transferable) of Fig. 2 which shows the "border lines" of the three groups.

Table VIIIe. DA 3-group means of discriminant scores.

Functions at group centroids		
	Function	
group	1	2
1	-.490	-.240
2	-2.523	.144
3	3.072	.085

Table VIIIf. DA 3-group classification table.

Classification results[a]						
			Predicted group membership			
		group	1	2	3	Total
Original	Count	1	90	10	0	100
		2	0	106	0	106
		3	7	0	96	103
	%	1	90.0	10.0	.0	100.0
		2	.0	100.0	.0	100.0
		3	6.8	.0	93.2	100.0

Fig. 2 3-group Discriminating plot.

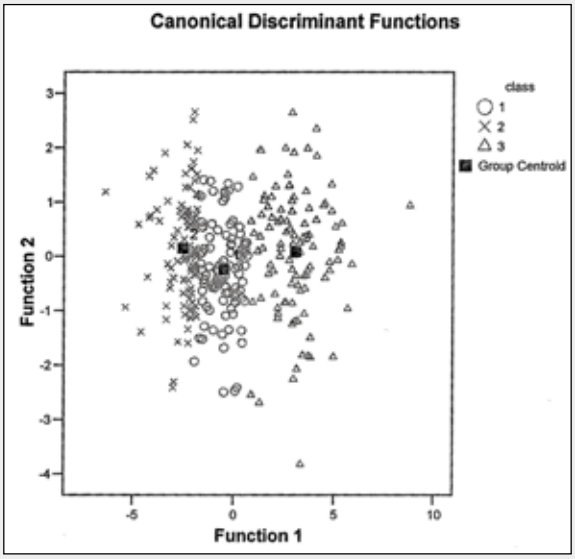
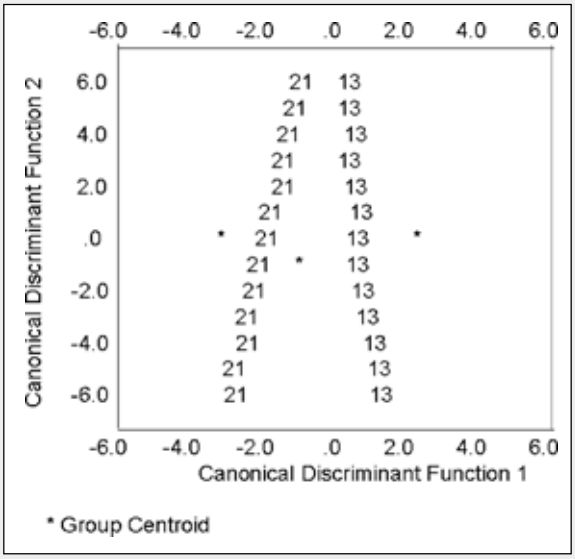


Fig. 3 3-group Territorial map.



DA also provides the option of a Stepwise analysis (see Template I). Performing a Stepwise analysis on the above 3-group analysis shows that only x1 and x2 (see Table IX) were used in the discriminating model with a total accuracy of 93.9%.

Table IX. Discriminant function – stepwise.

Canonical discriminant function coefficients		
	Function	
	1	2
x1	-.498	.249
x2	.521	.043
(Constant)	-.750	-23.783

It has been shown that DA also works well with qualitative independent variables like gender (1 = M, 2 = F), race, etc. So what is the difference between DA and binary logistic regression[1]? It has been recommended that when DA's assumptions failed, logistic regression is to be used. Both techniques give us the saved predicted probabilities for group membership which allows a further ROC analysis for model probability cut-off. DA has the Discriminant score which could be useful if one wants to derive a scoring system – like a fitness score, for example. Perhaps the obvious advantage of DA over binary logistic regression is the ability to discriminate more than two groups (which have to be analysed by a multinomial logistic regression – Biostatistics 305). In summary, if our aim is to develop a model to "discriminate", as the saying goes, "don't care whether it's a black cat or white cat, as long as it can catch a mouse, it's a good cat!".

REFERENCE

1. Chan YH. Biostatistics 202. Logistic regression analysis. Singapore Med J 2004; 45:149-53.

SINGAPORE MEDICAL COUNCIL CATEGORY 3B CME PROGRAMME

Multiple Choice Questions (Code SMJ 200502A)

	True	False
Question 1. The assumptions for a Discriminant analysis are:		
(a) Independent quantitative variables must be of normal distribution.	❑	❑
(b) The covariance of the variables should be unequal.	❑	❑
(c) Variables should have high correlations.	❑	❑
(d) Only quantitative variables could be used in the analysis.	❑	❑
Question 2. Which of the following is used to calculate the Discriminant scores?		
(a) The standardized canonical discriminant function coefficients.	❑	❑
(b) The structure matrix.	❑	❑
(c) The unstandardised canonical discriminant function coefficients.	❑	❑
(d) The Fisher's linear discriminant functions.	❑	❑
Question 3. The following statements are true:		
(a) A high Wilk's Lambda (near 1) shows good model discrimination.	❑	❑
(b) A high canonical correlation (near 1) shows that a function will discriminate well.	❑	❑
(c) Including more variables in a model will improve the accuracy.	❑	❑
(d) The impact of a variable on a discriminant function is given by the unstandardised canonical discriminant function coefficients.	❑	❑
Question 4. Discriminant analysis is better than logistic regression because:		
(a) Higher accuracies could be obtained.	❑	❑
(b) The probabilities for discrimination are available.	❑	❑
(c) Can be used to "differentiate" more than 2 groups.	❑	❑
(d) Can use Press's Q statistic to check on the discriminatory power of the model.	❑	❑
Question 5. The following techniques could be used to cross-validate a model:		
(a) The Binomial test.	❑	❑
(b) The leave-one-out classification.	❑	❑
(c) The test-retest samples.	❑	❑
(d) Performing a stepwise analysis.	❑	❑

Doctor's particulars:

Name in full: _____

MCR number: _____ Specialty: _____

Email address: _____

Submission instructions:

A. Using this answer form

1. Photocopy this answer form.
2. Indicate your responses by marking the "True" or "False" box ☑
3. Fill in your professional particulars.
4. Either post the answer form to the SMJ at 2 College Road, Singapore 169850 OR fax to SMJ at (65) 6224 7827.

B. Electronic submission

1. Log on at the SMJ website: URL http://www.sma.org.sg/cme/smj
2. Either download the answer form and submit to smj.cme@sma.org.sg OR download and print out the answer form for this article and follow steps A. 2-4 (above) OR complete and submit the answer form online.

Deadline for submission: (February 2005 SMJ 3B CME programme): 12 noon, 25 March 2005

Results:

1. Answers will be published in the SMJ April 2005 issue.
2. The MCR numbers of successful candidates will be posted online at http://www.sma.org.sg/cme/smj by 20 April 2005.
3. Passing mark is 60%. No mark will be deducted for incorrect answers.
4. The SMJ editorial office will submit the list of successful candidates to the Singapore Medical Council.