

LI Analysis Training Series

Discriminant Analysis

(Last Revised: 4/17/2007)

Melissa Ives, Rodney Funk & Michael Dennis
 Chestnut Health Systems
 Bloomington IL 61701
 309-827-6026
www.chestnut.org

Acknowledgement: This document was developed under contract #270-2003-00006 from the Substance Abuse and Mental Health Services Administration (SAMHSA's) Center for Substance Abuse Treatment (CSAT). Any opinions about this data are those of the authors and do not represent official positions of the government or individual grantees.

Purpose: Discriminant function analysis is used to predict a known categorization variable (e.g., level of care, HIV risk groups, relapse, recidivism) The results from the classification function coefficients are then used to create a calculation rule to assign "predicted groups" for "new cases".

Data Requirements:

The variable to be predicted (the grouping variable) must be categorical. Predictor variables must be interval or dichotomous.

The DISCRIMINANT procedure requires one /GROUPS subcommand to identify the variable to be predicted and one /VARIABLES subcommand to identify the predictor variables. Other subcommands are optional and those we commonly use are discussed below. While not used in this example, additional subcommands are available for: selecting records, selecting a method for entering variables into the analysis, exporting model information, changing the tolerance level for including a variable (proportion of a variables within-group variance not accounted for by other variables in the analysis (default is .0001)), minimum and maximum probabilities and Rao's V for variable inclusion, maximum number of steps allowed in the analysis, rotation options, output history, plots and reading in or writing to a matrix.

Procedure:

```
DISCRIMINANT
  /GROUPS=<var to be predicted>(lowvalue,hivalue)
  /VARIABLES=<varlist of predictors>
  /ANALYSIS ALL
  /SAVE=CLASS=<varname predicted level>
  /PRIORS Size
  /STATISTICS=COEFF TABLE
  /CLASSIFY=NONMISSING POOLED.
```

The /ANALYSIS subcommand can be specified more than once and is used to produce multiple analyses for the same run. It can alter which predictor variables are included and/or the order in which the predictor variables are analyzed. Be sure that any variables listed in the

/ANALYSIS subcommand are included in the /VARIABLES subcommand. By default, all variables are entered simultaneously. The keyword ALL includes all variable listed on the /VARIABLES subcommand. If the /ANALYSIS subcommand is omitted, the following warning will be issued “Since ANALYSIS was omitted for the first analysis, all variables on the VARIABLES list will be entered at level 2 when METHOD = DIRECT¹ and at level 1 when a stepwise METHOD.”

By default, DISCRIMINANT assumes that the groups to be predicted have equal probability. The /PRIORS Size subcommand above tells SPSS that the probability should be based on the size of the group in the grouping variable. You can also assign specific values for each group provided the number of values equals the number of groups and the sum of all the values is 1.

The /SAVE subcommand is used to save the predicted group membership (CLASS), the actual discriminant scores and/or the probability of group membership. In the syntax above, we are saving only the predicted groups as a new variable with the name identified after ‘CLASS=’.

In order to create a new calculated rule to assign new cases to groups, it is necessary to include the /STATISTICS=COEFF subcommand. We also include the TABLE statistic to show a table of the actual vs. predicted classifications. A footnote to this table indicates the percent correctly classified.

The final subcommand /CLASSIFY is used to specify how records are handled. In this example, only records with no missing data are included. Since we wanted NONMISSING, we also added the POOLED subcommand, which is the default if /CLASSIFY is not included, but needed if CLASSIFY is specified.

In the example below, the following syntax was used to create groups of HIV risk by predicting HIVCluster4 based on the original scales and variables (items listed in the /VARIABLES subcommand. The calculated groups (dHIVpc) will then be compared to the original groups (HIVCluster4) that were identified using a Cluster analysis of the z-scored versions of these variables (or related proportional variables).

```
DISCRIMINANT
  /GROUPS=HIVCluster4 (1 4)
  /VARIABLES=NPS r1k r1m r1n
  SxRS sxprtrs spr VICTIM VICT_TGF VWORRY E9u
  /ANALYSIS ALL
  /SAVE=CLASS=dHIVpc
  /PRIORS Size
  /STATISTICS=COEFF TABLE
  /CLASSIFY=NONMISSING POOLED .
```

Output:

This is an example based on a simple 4-cluster solution for HIV Risk using the 2006 CSAT Adolescent dataset and the syntax example above. The following output uses SPSS 14.0.2. Our comments in the boxes below do not appear in actual output.

¹ METHOD=DIRECT is the default.

Analysis 1

Summary of Canonical Discriminant Functions

This table is standard output for all discriminant analyses. It shows that there are 3 functions in this dataset, with the first function explaining nearly 60% of the variance. The second function adds another 22% and the third about 20%. Canonical correlation values close to 1 indicate a strong association between the groups and the discriminant scores.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	15.235(a)	58.8	58.8	.969
2	5.664(a)	21.9	80.6	.922
3	5.021(a)	19.4	100.0	.913

a First 3 canonical discriminant functions were used in the analysis.

This table is standard output for all discriminant analyses. The low values for Wilks'-Lambda in all 3 tests indicate that none of the function means are equal.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.002	68352.186	33	.000
2 through 3	.025	38948.943	20	.000
3	.166	18938.868	9	.000

This table is standard output for all discriminant analyses. The number of functions is equal to the 1 minus the number of groups or the number of variables; whichever is smaller. These values would be saved in the dataset if you specified /SAVE SCORES. The default name for the scores will be DSCn_1 where n is the function number.²

Standardized Canonical Discriminant Function Coefficients

	Function		
	1	2	3
nps Needle Problem Scale	.371	.173	.128
R1k P90 - days used needle inject anything	.194	1.088	.117
R1m P90 - How many people shared needles	-.050	.826	.074
R1n P90 - Days shared needles with others	-.119	-.097	-.014
sxrs Sex Risk Scale	.020	-.026	.028

² If you requested /SAVE PROBS, the default name for each probability is DSCn_2, where n=the number of the function.

sxprtrs Total number of sex partners-p90 days	-.113	.028	-.088
spr Sex Protection Ratio	-.095	-.067	.463
victim OMB Table: Ever victimized--New E9a-d, E9t	.267	.078	-1.054
vict_tgf Victimization: Traumagenic Factors	-.043	-.001	.096
vworry Current worries about being victimized	1.060	-.133	.299
E9u RERI P90 how many days attacked/abused	.420	-.049	.202

This table is standard output for all discriminant analyses. The Structure matrix organizes the variables by their correlations within function (beginning with the first function). In this case, vworry and e9u have the highest correlation with function 1.

Structure Matrix

	Function		
	1	2	3
vworry Current worries about being victimized	.814(*)	-.179	.190
E9u RERI P90 how many days attacked/abused	.098(*)	-.015	.009
R1k P90 - days used needle inject anything	.059	.649(*)	.065
R1m P90 - How many people shared needles	.024	.290(*)	.028
nps Needle Problem Scale	.054	.243(*)	.029
R1n P90 - Days shared needles with others	.015	.187(*)	.018
victim OMB Table: Ever victimized--New E9a-d, E9t	.264	.055	-.851(*)
vict_tgf Victimization: Traumagenic Factors	.157	.021	-.239(*)
sxrs Sex Risk Scale	.059	.032	-.120(*)
spr Sex Protection Ratio	-.035	-.030	.117(*)
sxprtrs Total number of sex partners-p90 days	.011	.004	-.045(*)

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

* Largest absolute correlation between each variable and any discriminant function

This table is standard output for all discriminant analyses. It shows the within-group (cluster) means for each function.

Functions at Group Centroids

HIVCluster4 HIV	Function
-----------------	----------

Risk clusters (z-scores of 10vars) - 4

cluster solution	1	2	3
1.00 A-Low Risk	-2.589	-.231	2.692
2.00 B-Mod Risk-low worry+traumatic factors,lower needle	-1.324	.153	-2.273
3.00 C-Mod Risk-high worry+trauma factors,higher needle	8.029	-.890	.613
4.00 D-High Risk	6.964	31.882	2.768

Unstandardized canonical discriminant functions evaluated at group means

The following output is in relation to the /PRIORS, /STATISTICS and /CLASSIFY subcommands.

Classification Statistics

Classification Processing Summary

Processed		29688
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	19130
Used in Output		10558

If the /PRIORS subcommand is used with the /STATISTICS subcommand, the following table will be reported, indicating the distribution of probabilities by group.

Prior Probabilities for Groups

HIVCluster4 HIV Risk clusters (z- scores of 10vars) - 4 cluster solution	Cases Used in Analysis		
	Prior	Unweighted	Weighted
1.00 A-Low Risk	.346	3649	3649.000
2.00 B-Mod Risk-low worry+traumatic factors,lower needle	.465	4914	4914.000
3.00 C-Mod Risk-high worry+trauma factors,higher needle	.184	1938	1938.000
4.00 D-High Risk	.005	57	57.000
Total	1.000	10558	10558.00 0

This table results from the /STATISTICS=COEFF subcommand. This is the table of values that are used to calculate the discriminant function in the next part of this example.

Classification Function Coefficients

	HIVCluster4 HIV Risk clusters (z-scores of 10vars) - 4 cluster solution			
	1.00 A-Low Risk	2.00 B-Mod Risk- low worry+traumatic factors,lower needle	3.00 C-Mod Risk- high worry+trauma factors,higher needle	4.00 D-High Risk
nps Needle Problem Scale	-.229	-.432	6.872	17.950
R1k P90 - days used needle inject anythng	-.037	-.005	.379	13.803
R1m P90 - How many people shared needles	.177	-.706	-9.512	205.672
R1n P90 - Days shared needles with others	.061	-.016	-.701	-2.692
sxrs Sex Risk Scale	1.421	1.346	1.525	1.031
sxprtrs Total number of sex partners-p90 days	-.004	.090	-.324	-.062
spr Sex Protection Ratio	14.634	6.670	8.348	4.768
victim OMB Table: Ever victimized--New E9a-d, E9t	-7.097	20.322	17.241	17.167
vict_tgf Victimization: Traumagenic Factors	.067	-.393	-.501	-.299
vworry Current worries about being victimized	.608	-1.107	96.294	53.135
E9u RERI P90 how many days attacked/abused	.032	-.021	.468	.294
(Constant)	-9.051	-14.189	-62.194	-564.478

Fisher's linear discriminant functions

This table is the result of the /STATISTICS=TABLE subcommand, and shows the actual vs. predicted values for the groups (clusters). The footnote below the table shows that 96% of the cases were correctly classified.

Classification Results(a)

		Predicted Group Membership				Total	
		1.00 A-Low Risk	2.00 B-Mod Risk-low worry+traumatic factors,lower needle	3.00 C-Mod Risk-high worry+trauma factors,higher needle	4.00 D-High Risk		
Original	Count	1.00 A-Low Risk	3649	0	0	0	3649
		2.00 B-Mod Risk-low worry+traumatic factors,lower needle	320	4585	9	0	4914
		3.00 C-Mod Risk-high worry+trauma factors,higher needle	7	102	1826	3	1938
		4.00 D-High Risk	1	3	2	51	57
%		1.00 A-Low Risk	100.0	.0	.0	.0	100.0
		2.00 B-Mod Risk-low worry+traumatic factors,lower needle	6.5	93.3	.2	.0	100.0
		3.00 C-Mod Risk-high worry+trauma factors,higher needle	.4	5.3	94.2	.2	100.0
		4.00 D-High Risk	1.8	5.3	3.5	89.5	100.0

a 95.8% of original grouped cases correctly classified.

Comments:

Once the discriminant analysis has been run in SPSS, we will calculate a discriminant rule for classifying additional or future cases. To simplify the process, we have created an Excel file in the format below where the 'Classification Function Coefficients' table can be copied. Using a series of concatenate functions, the Excel file produces the syntax (example below) needed to create the discriminant rule. The Excel function will need to be edited to accommodate a larger number of groups or variables.

Discriminant classification Excel file³

	A	B	C	D	E	F
1	Classification Function Coefficients					
2	9 Variables-Add new as needed and update calculations in rows 19 and 20					
3			GROUPS Predicted			
4	Var Label	Var Name	1	2	3	4
5	nps Needle Problem Scale	NPS	-0.229217606	-0.431634272	6.871578867	17.95025326
6	R1k P90 - days used needle inject anyt	R1k	-0.036815671	-0.004987857	0.378842677	13.80323714
7	R1m P90 - How many people shared ne	R1m	0.177000561	-0.706448022	-9.511937517	205.6719309
8	R1n P90 - Days shared needles with ot	R1n	0.061016172	-0.015792706	-0.700587976	-2.692424459
9	sxrs Sex Risk Scale	sxrs	1.420562249	1.345572457	1.524718714	1.030573176
10	sxprtrs Total number of sex partners-p	sxprtrs	-0.004141473	0.090009855	-0.324462713	-0.061912141
11	spr Sex Protection Ratio	spr	14.63424353	6.670357122	8.348084067	4.768123543
12	victim OMB Table: Ever victimized--Ne	victim	-7.096759987	20.32210378	17.24069109	17.16726478
13	vict_tgf Victimization: Traumagenic Fa	vict_tgf	0.066582295	-0.392923421	-0.500604721	-0.299228385
14	vworry Current worries about being vi	vworry	0.607689792	-1.107096651	96.29374082	53.13513974
15	E9u RERI P90 how many days attackec	e9u	0.031513206	-0.021165619	0.467508013	0.294074231
16	(Constant)		-9.050783497	-14.18898528	-62.19350547	-564.4782987
17	Fisher's linear discriminant functions					
18	Calculations for determining new groups based on coefficients above					
19	Group 1	Group 2	Group 3	Group 4		
20	compute grp1=(NPS*-0.229217606466799+R1k*-0.0368156708662908+R1m*0.177000560896333+R1n*0.0610161717920791+sxrs*1.42056224877762					
21	+sxprtrs*-0.00414147250089271+spr*14.6342435319541+victim*-7.09675998654686+vict_tgf*0.0665822948358583+vworry*0.60768979173698+e9u*0.0					
22						
23	Copy this into SPSS					
24	compute grp1=(NPS*-0.229217606466799+R1k*-0.0368156708662908+R1m*0.177000560896333+R1n*0.0610161717920791+sxrs*1.42056224877762					
25	+sxprtrs*-0.00414147250089271+spr*14.6342435319541+victim*-7.09675998654686+vict_tgf*0.0665822948358583+vworry*0.60768979173698+e9u*0.0					

The formulas in the two cells in A20 and A21 - underneath

“Calculations for determining new groups based on coefficients above

Group 1 are as follows

In cell A20:

=CONCATENATE("compute grp",C4,"=", "(",\$B5,"*",C5,"+", \$B6,"*",C6,"+", \$B7,"*",C7,"+", \$B8,"*",C8,"+", \$B9,"*",C9)

In cell A21:

=CONCATENATE("+", \$B10,"*",C10,"+", \$B11,"*",C11,"+", \$B12,"*",C12,"+", \$B13,"*",C13,"+", \$B14,"*",C14,"+", \$B15,"*",C15,"+", \$B16,"*",C16,")")

To add more variables, first make space for the output by inserting the necessary number of additional rows before row 16”(Constant)”, THEN copy the ‘Classification Function Coefficients’ from SPSS into cell A1. After this, you will need to adjust the formula in what is now cell A21 to add the series: **"+",\$B16,"*",C16**, after C15,. This will need to be done for each additional variable (i.e. you may have variables in rows 17 and 18 as well, so you would add the series above three times using 16, 17 and 18 to include the new variables.) The commas and quotations are important! The last item in the formula adds the constant, so it will need to be changed from C16 to the new cell address for the constant (if you have added three new variables, this would change to C19). Finally, copy this formula over the existing formula in row 21 for each group you are calculating.

To add more groups, simply copy the calculations from A20 and A21 into the same cells underneath any additional columns (e.g. in column G to add a 5th group.).

³ Additional calculations under ‘Copy this into SPSS’ are not shown.

In cells A24-25 (and the cells below) are simply pointers to the longer calculations so they end up in column form so they can be easily copied into SPSS.

If you have added more groups and more variables, update the variable syntax first, then copy the formula to the cells for the new groups.

The following is the syntax created from the Excel file⁴.

```
compute grp1=(NPS*-0.229217606466797 + R1k*-0.0368156708662909 +
  R1m*0.177000560896339 + R1n*0.0610161717920794+sxrs*1.42056224877761 +
  sxppd*-0.372732525078394+spr*14.6342435319541 +
  victim*-7.09675998654684 + vict_tgf*0.0665822948358586 +
  vworry*0.607689791736901 + e9u*0.0315132060756891 + -9.05078349715663).
compute grp2=(NPS*-0.431634271810806 + R1k*-0.00498785705748934 +
  R1m*-0.706448022326871 + R1n*-0.0157927061974964 + sxrs*1.34557245728823 +
  sxppd*8.10088698448553 + spr*6.67035712242361 + victim*20.322103778546 +
  vict_tgf*-0.392923420948189 + vworry*-1.10709665122328 +
  e9u*-0.0211656194155032 + -14.1889852776864).
compute grp3=(NPS*6.87157886739511 + R1k*0.378842676697873 + R1m*-9.51193751717916 +
  R1n*-0.700587975929445 + sxrs*1.52471871397815 + sxppd*-29.2016441660261 +
  spr*8.348084066733 + victim*17.2406910947572 + vict_tgf*-0.500604721194182 +
  vworry*96.2937408247949 + e9u*0.467508013010264 + -62.1935054719625).
compute grp4=(NPS*17.950253260416 + R1k*13.8032371378502 + R1m*205.671930875068 +
  R1n*-2.69242445899122 + sxrs*1.03057317616586 + sxppd*-5.57209268286663 +
  spr*4.76812354292021 + victim*17.1672647775792 + vict_tgf*-0.299228385439944 +
  vworry*53.1351397361728 + e9u*0.294074230620417 + -564.478298707234).
```

To create the group variable using the discriminant function, select the calculated group variables (e.g. grp1 to grp4) that has the highest value and set the discriminant cluster variable to that group number.

```
if (grp1 > max(grp2,grp3,grp4)) dHIVcf=1.
if (grp2 > max(grp1,grp3,grp4)) dHIVcf=2.
if (grp3 > max(grp1,grp2,grp4)) dHIVcf=3.
if (grp4 > max(grp1,grp2,grp3)) dHIVcf=4.
```

Using more clusters and/or more variables will result in a longer discriminant rule as in the example below from Dennis, Wechsberg, et al. (2001) where there were 8 groups and 21 variables.

⁴ For ease of printing, a space has been added before and after each plus sign (+) that is not included in the concatenations or in the SPSS syntax.

```

*** Create DERG8c based on the discriminant runs from the old cohort data.
*** Discriminant ERG 8 Groups computed.

Compute grp1=(-0.0167313725147621*abdyijd
-0.0010522780603889*actijj30
0.1454098297067800*abdylstc
0.5657241126251330*ahriskf
0.0987900844841663*vag_a
-0.0163367351919234*afsexp
0.1016080882785490*vag_r
-9.5256810926435500).

Compute grp2=(-0.0508542964388557*abdyijd
0.0011298117228633*actijj30
0.1541701700541000*abdylstc
0.3557715646318480*ahriskf
0.2346209471853690*vag_a
-0.0211577283658239*afsexp
0.08814814211559048*vag_r
-14.2354952434888000).

Compute grp3=(-0.0611581921695071*abdyijd
-0.0021239696488629*actijj30
0.1065052212329810*abdylstc
0.7587280725630050*ahriskf
1.1739707156165200*vag_a
-0.0906294113222242*afsexp
0.0240878980711963*vag_r
-107.6834212944770000).

Compute grp4=(0.0155421656847390*abdyijd
-0.0022537121518590*actijj30
0.0802778796510271*abdylstc
1.3505696646390800*ahriskf
0.2263290343100620*vag_a
-0.0175449574039482*afsexp
0.1254821924688380*vag_r
-23.9992212317254000).

Compute grp5=(0.0373514619650277*abdyijd
-0.0040801920380040*actijj30
0.0501067791942429*abdylstc
1.2305000658055000*ahriskf
0.0711336984395082*vag_a
-0.0194833529216982*afsexp
0.1277310533729960*vag_r
+0.0116773448755138*abdyije
+0.0126153988934649*actused
+1.2426918487676200*dcnsw30
+0.0752216619986581*abdylista
+7.0019700898686100*vag_apr
+0.0272316044201766*aftddsc
+9.3239296893343400*vag_rpr
+0.0135438359503893*abdyije
+0.0130732214009489*actused
+1.2065854988355500*dcnsw30
+0.0992712400126697*abdylista
+5.1781147152779400*vag_apr
+0.0098714291374300*aftddsc
+10.5729857270835000*vag_rpr
+0.0261488392029916*abdyije
+0.0092146235038456*actused
+6.2934651909979500*dcnsw30
+0.0086143295001728*abdylista
+12.3856639820369000*vag_apr
+0.3234192800361470*aftddsc
+8.8112934580360000*vag_rpr
+0.0244880060514282*abdyije
+0.0305794243535812*actused
+14.921845138631800*dcnsw30
+0.1132084226418170*abdylista
+5.2405423484630700*vag_apr
+0.0042885377003840*aftddsc
+10.9127428976376000*vag_rpr
+0.0693352919502520*abdyije
+0.0321654908381404*actused
+14.796554342301700*dcnsw30
+0.0836162514345284*abdylista
+7.1327368300854400*vag_apr
+0.0042721749399283*aftddsc
+10.5104163729199000*vag_rpr
+0.0126572549784067*abdyijf+
+1.0604183409638300*dcrow30+
+0.2706532054610210*ahriskc+
+0.9537802047218760*dgegdst+
+0.0239294586762797*ana_a+
+1.2280305540547900*dgegsgdt+
+0.2670791161460980*ana_r+
+0.0011324895499435*abdyijf+
+0.6499748323659190*dcrow30+
+0.6026369473591360*ahriskc+
+6.7853522436461100*dgegdst+
+0.4187241675971390*ana_a+
+0.0977336448005431*dgegsgdt+
+0.1950067982683480*ana_r+
+0.1044831327350390*abdyijf+
+1.5912349468012100*dcrow30+
+0.0085517319260578*ahriskc+
+12.3069080439360000*dgegdst+
+6.4717658103344700*ana_a+
+3.5023939169816300*dgegsgdt+
+0.1570424956440350*ana_r
+0.0167040269553382*abdyijf+
+9.6545386673885900*dcrow30+
+1.6470680984466400*ahriskc+
+4.8898304089124800*dgegdst+
+0.2289125875659420*ana_a+
+0.4094149495296880*dgegsgdt+
+0.1000703330279440*ana_r+
+0.0247315026815315*abdyijf+
+10.6243216706962000*dcrow30+
+2.7048648809615900*ahriskc+
+0.8241449693821640*dgegdst+
+0.2470253200760360*ana_a+
+0.0743620831361340*dgegsgdt+
+0.05746616074666005*ana_r+

```

```

-22.15233118815876000).
Compute grp6=(0.1517761287833010*abdyi.jd
0.0159642281492469*acti.j30
0.1087165836672910*abdy.lstc
2.1226580749092700*ahriskf
0.0457593042087807*vag_a
0.6941933867528930*afsexp
0.5666543047009710*vag_r
-92.0019890426765000).
Compute grp7=(0.1971306777977430*abdyi.jd
0.0173840377901986*acti.j30
0.0564465049150939*abdy.lstc
1.1160562611362700*ahriskf
0.0866663113307844*vag_a
-0.0386775917816750*afsexp
0.1442764727301740*vag_r
-28.7661983863303000).
Compute grp8=(0.4986980879939240*abdyi.jd
0.0726617136887980*acti.j30
0.0917287351283340*abdy.lstc
1.6708755397645300*ahriskf
0.0605611905429530*vag_a
-0.0685958720495480*afsexp
0.1898006555050370*vag_r
-58.2098691188530000).
execute.

if (grp1 > max(grp2, grp3, grp4, grp5, grp6, grp7, grp8)) derg8cf=1.
if (grp2 > max(grp1, grp3, grp4, grp5, grp6, grp7, grp8)) derg8cf=2.
if (grp3 > max(grp1, grp2, grp4, grp5, grp6, grp7, grp8)) derg8cf=3.
if (grp4 > max(grp1, grp2, grp3, grp5, grp6, grp7, grp8)) derg8cf=4.
if (grp5 > max(grp1, grp2, grp3, grp4, grp6, grp7, grp8)) derg8cf=5.
if (grp6 > max(grp1, grp2, grp3, grp4, grp5, grp7, grp8)) derg8cf=6.
if (grp7 > max(grp1, grp2, grp3, grp4, grp5, grp6, grp8)) derg8cf=7.
if (grp8 > max(grp1, grp2, grp3, grp4, grp5, grp6, grp7)) derg8cf=8.

variable labels derg8cf 'Discriminant ERG 8 Groups computed w/follow-up data'.
value labels derg8cf 1 'PCU' 2 'CRS' 3 'HPRT2' 4 'PDSR' 5 'PNU' 6 'HPRT1' 7 'HFNU' 8 'HRNU'.
freq derg8cf.

```

To identify the success of the prediction including kappa, use:

CROSSTAB TABLES=<actual> BY <Predicted>/CELLS=COUNT ROW COLUMN/STATS=KAPPA.

Kappa, or Cohen's kappa, is a measure of the percentage-of-agreement correcting for chance, usually between two raters, for categorical data. In this case, the measure of agreement is between two ways of arriving at discriminant groups (i.e. groups created using Cluster analysis vs. the predicted groups using discriminant analysis).

Describing These Procedures: These procedures can be described in a variety of ways. Below are three examples from published articles.

To evaluate the reliability and replicability of the solution, we used a discriminant analysis based on the variables used in the cluster solution to predict cluster membership; the Kappa between the original clusters and the predicted cluster was 0.86. (Godley, Dennis, Godley, Funk. 2004).

Finally, we did a discriminant analysis of these 21 items predicting the HIV risk groups used in this paper and were able to correctly classify 79.2% (Kappa . 0.734). We looked at the ability of the discriminant rule based on only 21 items to predict the group. As recommended by contemporary methodologists (Kraemer, 1992), we used a Kappa statistic to compare the cluster analysis based on all the data and just the 21 variables since neither is considered perfect or error free. A table of the variables and Fischer linear discriminant functions is available from <http://www.chestnut.org/LI/Posters/erg8disc.pdf> or from the first author. (Dennis, Wechsberg, et al. , 2001).

The third step was to conduct a discriminant analysis to predict cluster membership based on the intake ASI composite scores, time in treatment, and prior treatment episodes. Time in treatment was measured with dummy variables that were created for those clients staying in treatment more than one month and another for those staying more than 3 months. Using these 10 variables, we were able to correctly predict 67% outcomes in terms of patterns of change in the original 100 cases. With a kappa of .55, this is about as good as most psychiatric diagnostic tests (Kraemer, 1992). The resulting discriminant function equations were used to classify people in both the original sample of 100, as well as a subsequent of 39.

Fourth, we evaluated the validity of the cluster solution through (a) reverse validation (i.e., predicting the source variables from the cluster solution) and (b) replication of this solution by using the discriminant function to calculate the group membership on a subsequent sample of 39 clients. The discriminant function based cluster solution had similar predictive validity (percent of variance explained) for both the original and subsequent sample on both the Life Distress factor (61%, 49%) and Substance Use Severity factor (50%, 56%). (Godley, Funk, Dennis, et al. 2004).

Bibliography

- Dennis, M. L., Wechsberg, W. M., McDermeit (now Ives), M., Campbell, R. S., & Rasch, R. R. (2001). The correlates and predictive validity of HIV risk groups among drug users in a community-based sample: Methodological findings from a multi-site cluster analysis. *Evaluation and Program Planning*, 24, 187-206.
- Godley, S.H., Dennis, M.L., Godley, M.D., Funk, R.R. (2004). Thirty-month relapse trajectory cluster groups among adolescents discharged from out-patient treatment. *Addiction*. 99 (Suppl. 2), 129–139.
- Godley, S.H., Funk, R.R., Dennis, M.L., Oberg, D., Passetti, L. (2004). Predicting response to substance abuse treatment among pregnant and post partum women. *Evaluation and Program Planning*. 27:223-231.
- Statistical Program for the Social Sciences (SPSS 1999). SPSS Base 10.0 User's Guide. Chicago, IL: Author (www.spss.com). (NOTE: This is the most recent non-electronic SPSS manual describing the procedures for running discriminant analysis. See Chapter 27, pages 315-321. For the most part, the syntax has not changed from earlier versions.)
- Statistical Program for the Social Sciences version 14.0.2 (SPSS 2006). SPSS Tutorial "Discriminant Analysis".