

DISCRIMINANT ANALYSIS: ASSUMPTIONS*

John R. Slate
Ana Rojas-LeBouef

This work is produced by The Connexions Project and licensed under the
Creative Commons Attribution License [†]



NOTE: This chapter is published by NCPEA Press¹ and is presented as an NCPEA/Connexions publication "print on demand book." Each chapter has been peer-reviewed, accepted, and endorsed by the National Council of Professors of Educational Administration (NCPEA) as a significant contribution to the scholarship and practice of education administration.

About the Authors

John R. Slate is a Professor at Sam Houston State University where he teaches Basic and Advanced Statistics courses, as well as professional writing, to doctoral students in Educational Leadership and Counseling. His research interests lie in the use of educational databases, both state and national, to reform school practices. To date, he has chaired and/or served over 100 doctoral student dissertation committees. Recently, Dr. Slate created a website (Writing and Statistical Help²) to assist students and faculty with both statistical assistance and in editing/writing their dissertations/theses and manuscripts.

Ana Rojas-LeBouef is a Literacy Specialist at the Reading Center at Sam Houston State University where she teaches developmental reading courses. Dr. LeBoeuf recently completed her doctoral degree in Reading, where she conducted a 16-year analysis of Texas statewide data regarding the achievement gap. Her research interests lie in examining the inequities in achievement among ethnic groups. Dr. Rojas-LeBouef also assists students and faculty in their writing and statistical needs on the Writing and Statistical Help website.

In this set of steps, readers will learn how to conduct a canonical discriminant analysis procedure. For detailed information regarding the assumptions underlying use of a discriminant analysis, readers are referred to the Hyperstats Online Statistics Textbook at <http://davidmlane.com/hyperstat/>³; to the *Electronic Statistics Textbook* (2011) at <http://www.statsoft.com>

*Version 1.2: Aug 18, 2011 6:54 am -0500

[†]<http://creativecommons.org/licenses/by/3.0/>

¹<http://www.ncpeapublications.org/books.html>

²<http://cnx.org/content/m40733/latest/www.writingandstatisticalhelp>

³<http://davidmlane.com/hyperstat/>

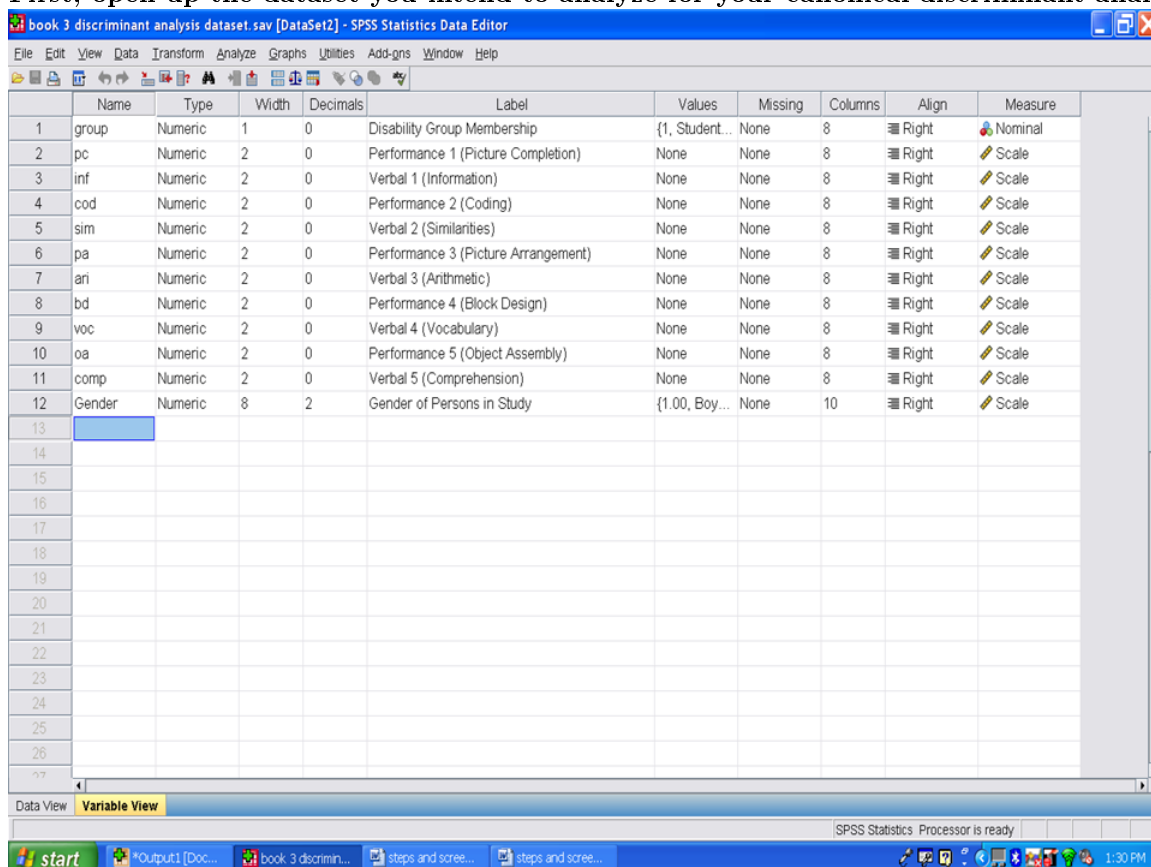
⁴<http://www.statsoft.com/textbook/>

; or to Andy Field's (2009) *Discovering Statistics Using SPSS* at http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967861⁵

Research questions for which a discriminant analysis procedure is appropriate involve determining variables that predict group membership. For example, if two groups of persons are present such as completers and non-completers and archival data are available, then a discriminant analysis procedure could be utilized. Such a procedure could identify specific variables that differentiate group membership. As such, interventions could be developed and targeted toward the variables that predicted group membership. Other sample research questions for which a discriminant analysis might be appropriate: (a) What factors differentiates successful from unsuccessful students?; (b) What factors differentiate delinquents from nondelinquents?; (c) What set of test scores best differentiates students with LD, students who are failing, and students with MR?; and (d) What set of factors differentiates drop-outs from persisters?

For purposes of this chapter, our research question is: “What scholastic variables differentiate boys from girls?”

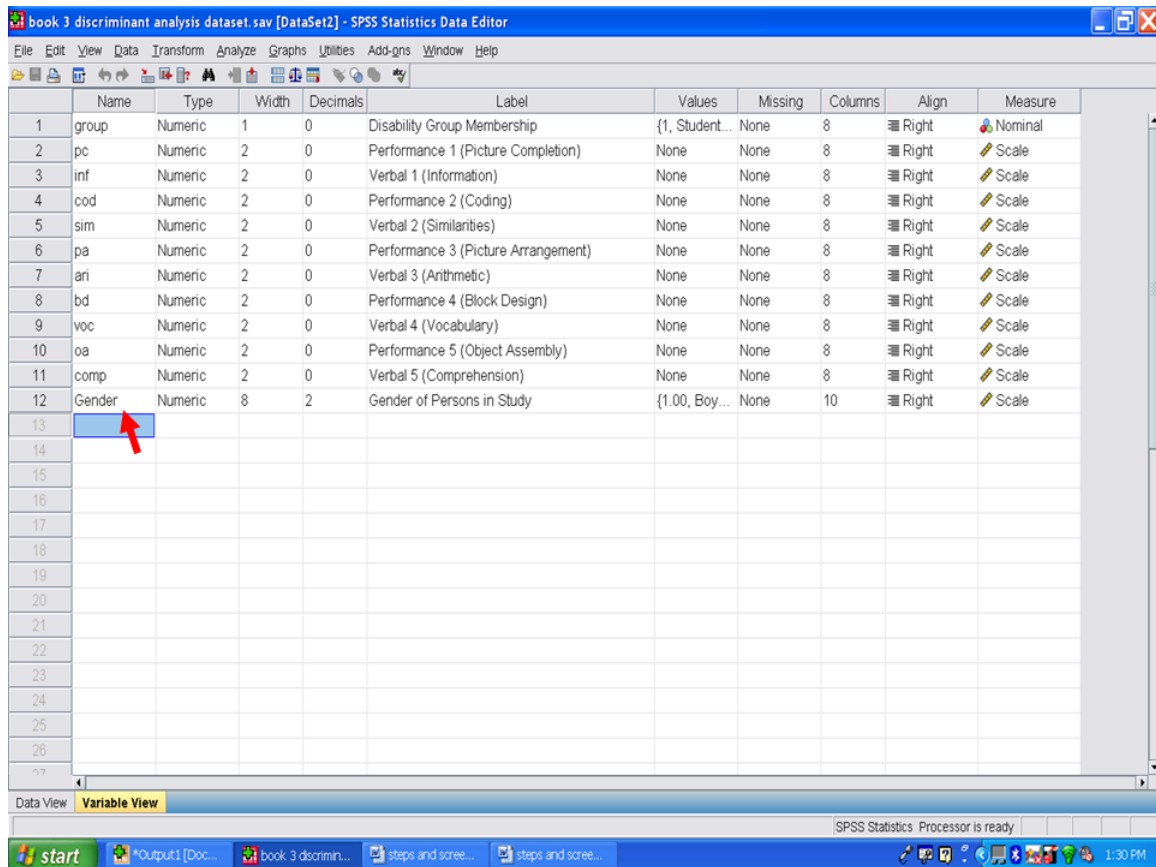
First, open up the dataset you intend to analyze for your canonical discriminant analysis.



Our independent variable is gender. Boys are labeled as group 1 and girls are labeled as group 2.

⁵http://www.amazon.com/Discovering-Statistics-Introducing-Statistical-Method/dp/1847879071/ref=sr_1_1?s=books&ie=UTF8&qid=1304967861

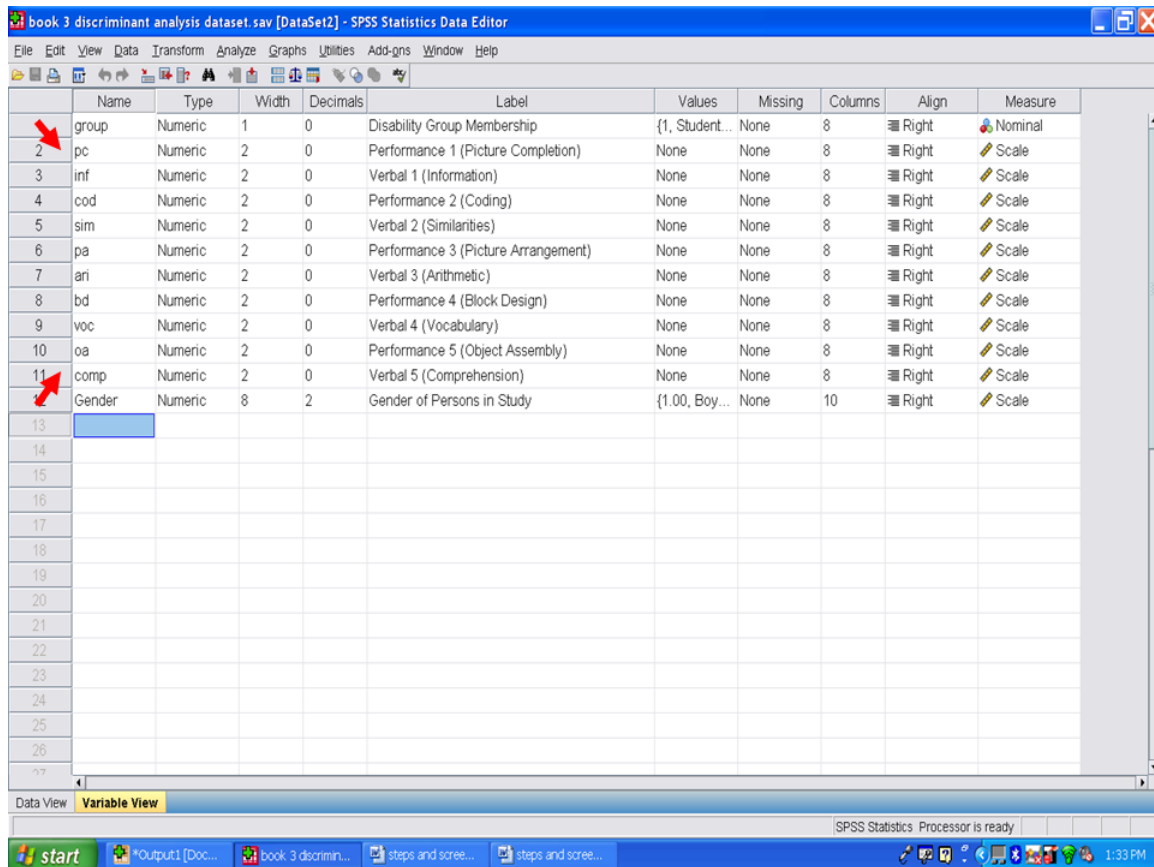
⁶<http://cnx.org/content/m40733/latest/13.1.png/image>



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	group	Numeric	1	0	Disability Group Membership	{1, Student...	None	8	Right	Nominal
2	pc	Numeric	2	0	Performance 1 (Picture Completion)	None	None	8	Right	Scale
3	inf	Numeric	2	0	Verbal 1 (Information)	None	None	8	Right	Scale
4	cod	Numeric	2	0	Performance 2 (Coding)	None	None	8	Right	Scale
5	sim	Numeric	2	0	Verbal 2 (Similarities)	None	None	8	Right	Scale
6	pa	Numeric	2	0	Performance 3 (Picture Arrangement)	None	None	8	Right	Scale
7	ari	Numeric	2	0	Verbal 3 (Arithmetic)	None	None	8	Right	Scale
8	bd	Numeric	2	0	Performance 4 (Block Design)	None	None	8	Right	Scale
9	voc	Numeric	2	0	Verbal 4 (Vocabulary)	None	None	8	Right	Scale
10	oa	Numeric	2	0	Performance 5 (Object Assembly)	None	None	8	Right	Scale
11	comp	Numeric	2	0	Verbal 5 (Comprehension)	None	None	8	Right	Scale
12	Gender	Numeric	8	2	Gender of Persons in Study	{1.00, Boy...	None	10	Right	Scale
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										

Our dependent variables, the ones we will use to differentiate boys from girls are 10 subscales from the Wechsler Intelligence Scale for Children-Third Edition: Picture Completion (pc), Information (inf), Coding (cod), Similarities (sim), Picture Arrangement (pa), Arithmetic (ari), Block Design (bd), Vocabulary (voc), Object Assembly (oa), and Comprehension (comp).

⁷<http://cnx.org/content/m40733/latest/13.2.png/image>



In the previous screenshots, we were in the variable view screen. Click on data view, shown below, so that your screen looks like the one below.

⁸<http://cnx.org/content/m40733/latest/13.3.png/image>

book 3 discriminant analysis dataset.sav [DataSet2] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1: pc 10.0 Visible: 12 of 12 Variables

	group	pc	inf	cod	sim	pa	ari	bd	voc	oa	comp	Gender	var	v
1	1	10	7	8	7	10	5	7	7	6	7	2.00		
2	1	8	8	7	6	7	9	9	4	4	7	2.00		
3	1	10	8	2	9	10	11	9	6	11	8	2.00		
4	1	3	4	11	4	7	6	3	4	10	6	1.00		
5	1	5	7	17	6	8	10	7	5	9	1	1.00		
6	1	2	5	8	8	9	7	10	3	5	5	1.00		
7	1	7	4	5	8	7	11	11	2	12	5	2.00		
8	1	7	7	7	7	14	8	10	8	11	3	2.00		
9	1	9	6	10	6	9	8	4	5	8	5	1.00		
10	1	5	9	7	10	9	12	11	7	8	11	1.00		
11	1	6	7	5	8	7	7	7	4	7	5	1.00		
12	1	13	8	5	13	10	5	6	9	9	7	2.00		
13	1	7	8	11	6	8	4	3	3	4	4	1.00		
14	1	6	6	8	12	9	9	9	10	6	8	1.00		
15	1	7	8	3	1	9	8	5	10	3	9	1.00		
16	1	10	7	5	5	9	5	5	8	3	6	2.00		
17	1	12	8	7	8	13	10	11	7	6	5	1.00		
18	1	15	7	5	10	15	8	7	8	7	10	2.00		
19	1	9	6	4	8	3	5	8	6	10	3	2.00		
20	1	6	7	10	6	9	6	8	6	5	2	2.00		
21	1	13	7	8	7	9	6	3	6	6	4	2.00		
22	1	10	7	12	8	8	8	12	5	11	4	2.00		
23	1	10	9	10	10	11	11	14	11	11	10	1.00		
24	1	6	6	6	2	5	7	6	6	11	6	2.00		
25	1	14	7	6	9	10	6	7	6	8	5	2.00		

Data View Variable View

SPSS Statistics Processor is ready

start *Output1 [Doc... book 3 discrimin... steps and scree... steps and scree... 1:34 PM

Prior to conducting a canonical discriminant function, we need to check the assumptions that underlie its use.

1 Normal Distribution

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution. You can examine whether or not variables are normally distributed with histograms of frequency distributions. However, note that violations of the normality assumption are usually not "fatal," meaning, that the resultant significance tests etc. are still "trustworthy." You may use specific tests for normality in addition to graphs. <http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>¹⁰

We recommend that you calculate the standardized skewness coefficients and the standardized kurtosis coefficients, as discussed in other chapters.

* Skewness [Note. Skewness refers to the extent to which the data are normally distributed around the mean. Skewed data involve having either mostly high scores with a few low ones or having mostly low scores with a few high ones.] Readers are referred to the following sources for a more detailed definition of skewness: http://www.statistics.com/index.php?page=glossary&term_id=356¹¹ and <http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb>¹²

To standardize the skewness value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the skewness value from the

⁹<http://cnx.org/content/m40733/latest/13.4.png/image>

¹⁰<http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>

¹¹http://www.statistics.com/index.php?page=glossary&term_id=356

¹²<http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb>

SPSS output and divide it by the Std. error of skewness. If the resulting calculation is within -3 to +3, then the skewness of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this ± 3 range, the dataset is not normally distributed.

* Kurtosis [Note. Kurtosis also refers to the extent to which the data are normally distributed around the mean. This time, the data are piled up higher than normal around the mean or piled up higher than normal at the ends of the distribution.] Readers are referred to the following sources for a more detailed definition of kurtosis: http://www.statistics.com/index.php?page=glossary&term_id=326 and <http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb>¹⁴

To standardize the kurtosis value so that its value can be constant across datasets and across studies, the following calculation must be made: Take the kurtosis value from the SPSS output and divide it by the Std. error of kurtosis. If the resulting calculation is within -3 to +3, then the kurtosis of the dataset is within the range of normality (Onwuegbuzie & Daniel, 2002). If the resulting calculation is outside of this ± 3 range, the dataset is not normally distributed.

2 Homogeneity of Variances/Covariances

It is assumed that the variance/covariance matrices of variables are homogeneous across groups. Again, minor deviations are not that important. <http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>¹⁵

3 Correlations between Means and Variances

The major "real" threat to the validity of significance tests occurs when the means for variables across groups are correlated with the variances (or standard deviations). Intuitively, if there is large variability in a group with particularly high means on some variables, then those high means are not reliable. However, the overall significance tests are based on pooled variances, that is, the average variance across all groups. Thus, the significance tests of the relatively larger means (with the large variances) would be based on the relatively smaller pooled variances, resulting erroneously in statistical significance. In practice, this pattern may occur if one group in the study contains a few extreme outliers, who have a large impact on the means, and also increase the variability. To guard against this problem, inspect the descriptive statistics, that is, the means and standard deviations or variances for such a correlation. <http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>¹⁶

After calculating the means and standard deviations for your variables for each of your groups, check them to determine if large variability is present in the means for one of your groups compared to the means for the other group.

4 The Matrix Ill-Conditioning Problem

Another assumption of discriminant function analysis is that the variables that are used to discriminate between groups are not completely redundant. As part of the computations involved in discriminant analysis, you will invert the variance/covariance matrix of the variables in the model. If any one of the variables is completely redundant with the other variables then the matrix is said to be *ill-conditioned*, and it cannot be inverted. For example, if a variable is

¹³http://www.statistics.com/index.php?page=glossary&term_id=326

¹⁴<http://www.statsoft.com/textbook/basic-statistics/#Descriptive%20statisticsb>

¹⁵<http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>

¹⁶<http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>

the sum of three other variables that are also in the model, then the matrix is ill-conditioned. <http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>¹⁷

What this assumption means is that each variable should be unique from any other variable in the analysis. Having one variable that includes another variable would be a violation of this assumption. An example of this would be using a total score that contains several subscale scores, all of which are used in the discriminant analysis.

5 Tolerance Values.

In order to guard against matrix ill-conditioning, constantly check the so-called tolerance value for each variable. This tolerance value is computed as *1 minus R-square* of the respective variable with all other variables included in the current model. Thus, it is the proportion of variance that is unique to the respective variable. In general, when a variable is almost completely redundant (and, therefore, the matrix ill-conditioning problem is likely to occur), the tolerance value for that variable will approach 0. <http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>¹⁸

We will check this assumption, the tolerance values, when we examine the SPSS output.

¹⁷<http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>

¹⁸<http://www.statsoft.com/textbook/discriminant-function-analysis/#assumptions>