

# Running a Discriminant Analysis with XLSTAT

[demoDA.xls](#)

## Dataset for running a Discriminant Analysis

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking [here](#).

The data are from [Fisher M. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179 -188] and correspond to 150 Iris flowers, described by four variables (sepal length, sepal width, petal length, petal width) and their species. Three different species have been included in this study: setosa, versicolor and virginica.

## Goal of this Discriminant Analysis

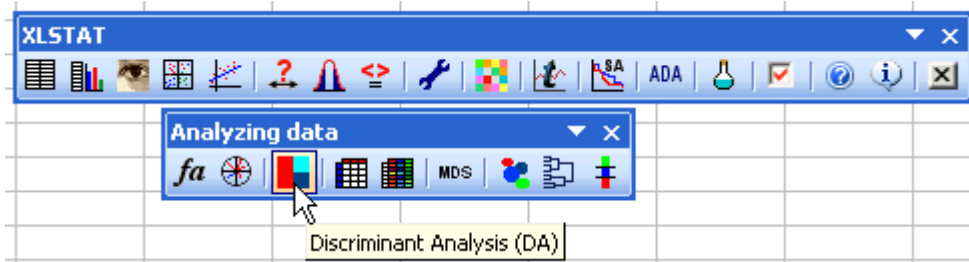
Our goal is to test if the four variables allow to discriminate the species, and to visualize the observations on a 2-dimensional map that shows as well as possible how separated the groups are.



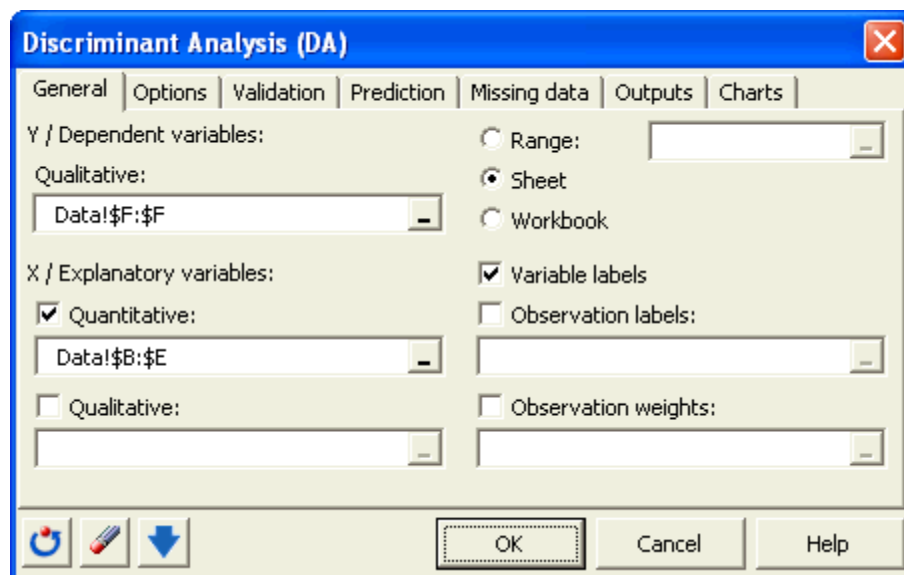
Iris setosa, versicolor and virginica.

## Setting up a Discriminant Analysis

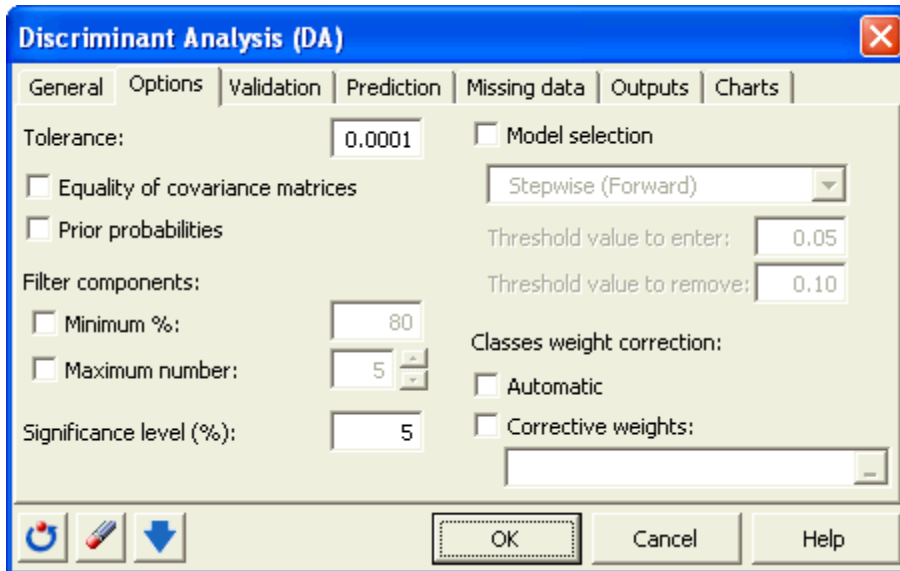
After opening XLSTAT, select the **XLSTAT / Analyzing data / Discriminant analysis** command, or click on the corresponding button of the **Analyzing data** toolbar (see below).



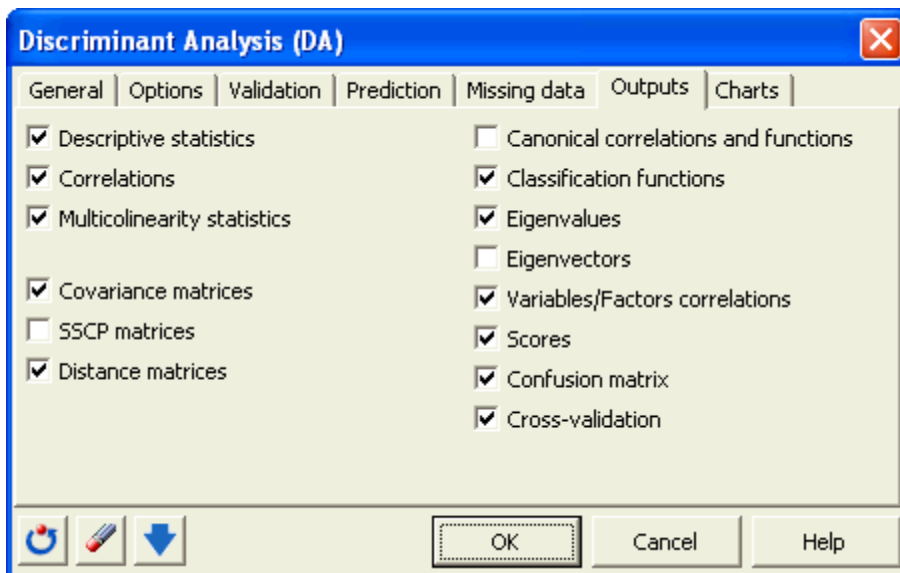
Once you've clicked on the button, the Discriminant analysis dialog box appears. The qualitative dependent variable corresponds here to the "Species" variable. The **quantitative Explanatory variables** are the four descriptive variables.



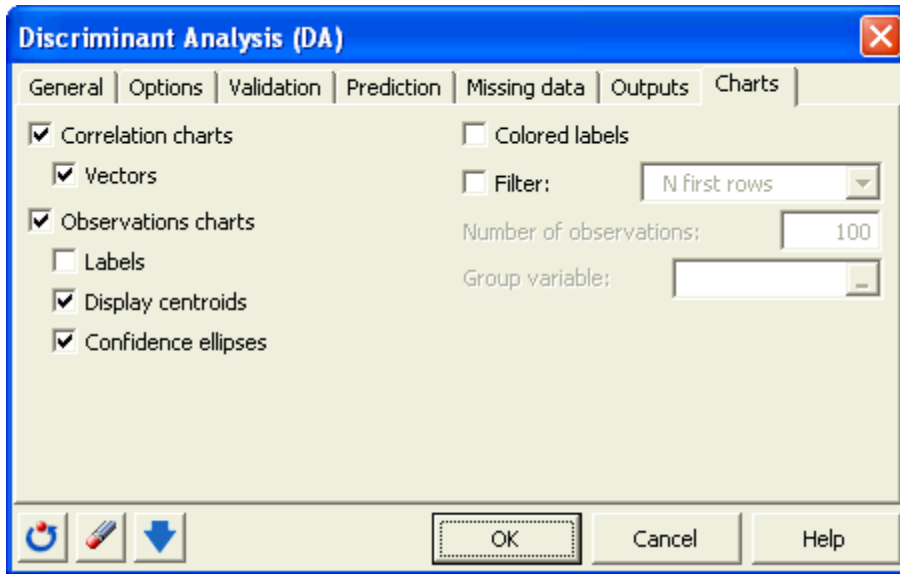
We uncheck the **Equality of covariance matrices** option because, as we will see with the Box's test, assuming that the covariance matrices of the three species are equal would be wrong.



Many results are optionally displayed by XLSTAT. We can see below which options have been activated for this particular case.



In order to avoid adding too much information on the plots, we have unchecked the **Labels** option in the **Charts** tab.



The computations begin once you have clicked on **OK**. The results will then be displayed.

## Interpreting the results of a Discriminant Analysis

The first results displayed are the various matrices used for the computations. The two Box's tests confirm that we need to reject the hypothesis that the covariance matrices are equal between the groups.

Box test (chi-square asymptotic approximation):	
-2Log(M)	146,663
Chi-square	140,943
Chi-square	31,410
DF	20
p-value	< 0,0001
alpha	0,05
Test interpretation:	
H0: The within-class covariance matrices are equal.	
Ha: The within-class covariance matrices are different.	
As the computed p-value is lower than the significance level alpha=0,05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.	
The risk to reject the null hypothesis H0 while it is true is lower than 0,01%.	
Box test (Fisher's F asymptotic approximation):	
-2Log(M)	146,663
F (Observed)	7,045
F (Critical value)	1,571
DF1	20
DF2	77567
p-value	< 0,0001
alpha	0,05
Test interpretation:	
H0: The within-class covariance matrices are equal.	
Ha: The within-class covariance matrices are different.	
As the computed p-value is lower than the significance level alpha=0,05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.	
The risk to reject the null hypothesis H0 while it is true is lower than 0,01%.	

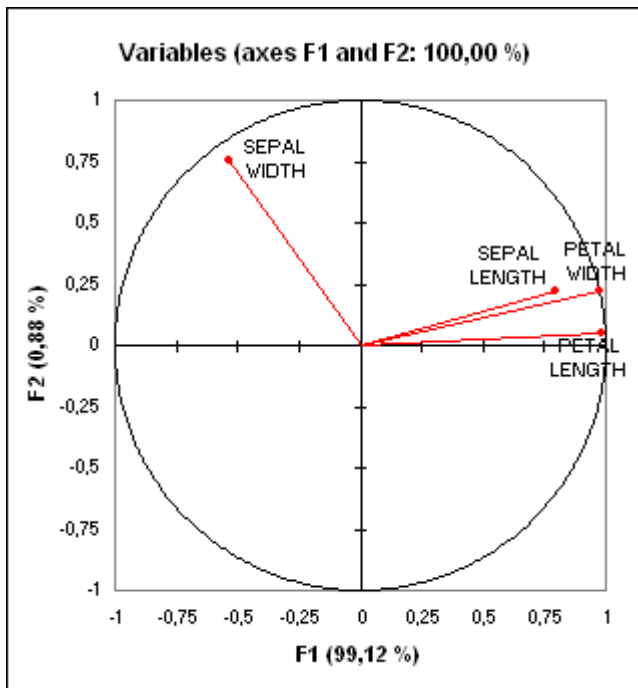
The Wilks' Lambda test allows to test if the vector of the means for the various groups are equal or not (you can understand it as a multidimensional version of the Fisher's LSD or the Tukey's HSD tests). We see that the difference between the means vectors of the groups is significant.

Wilks' Lambda test (Rao's approximation):	
Lambda	0,023
F (Observed)	199,145
F (Critical value)	1,971
DF1	8
DF2	288
p-value	< 0,0001
alpha	0,05
Test interpretation:	
H0: The means vectors of the 3 classes are equal.	
Ha: At least one of the means vector is different from another.	
As the computed p-value is lower than the significance level alpha=0,05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.	
The risk to reject the null hypothesis H0 while it is true is lower than 0,01%.	

The next table shows the eigenvalues and the corresponding % of variance. We can see that 99% of the variance is represented with the first factor. There are only two factors: the maximum number of factors is equal to k-1, when  $n > p > k$ , where n is the number of observations, p the number of explanatory variables, and k the number of groups.

Eigenvalues:		
	F1	F2
Eigenvalue	32,192	0,285
Discrimina	99,121	0,879
Cumulative	99,121	100,000

The following chart shows how the initial variables are correlated with the two factors (this chart corresponds to the factor loadings table). We can see that the factor F1 is correlated with Sepal length, Petal length, and Petal width and that F2 is correlated with Sepal width.



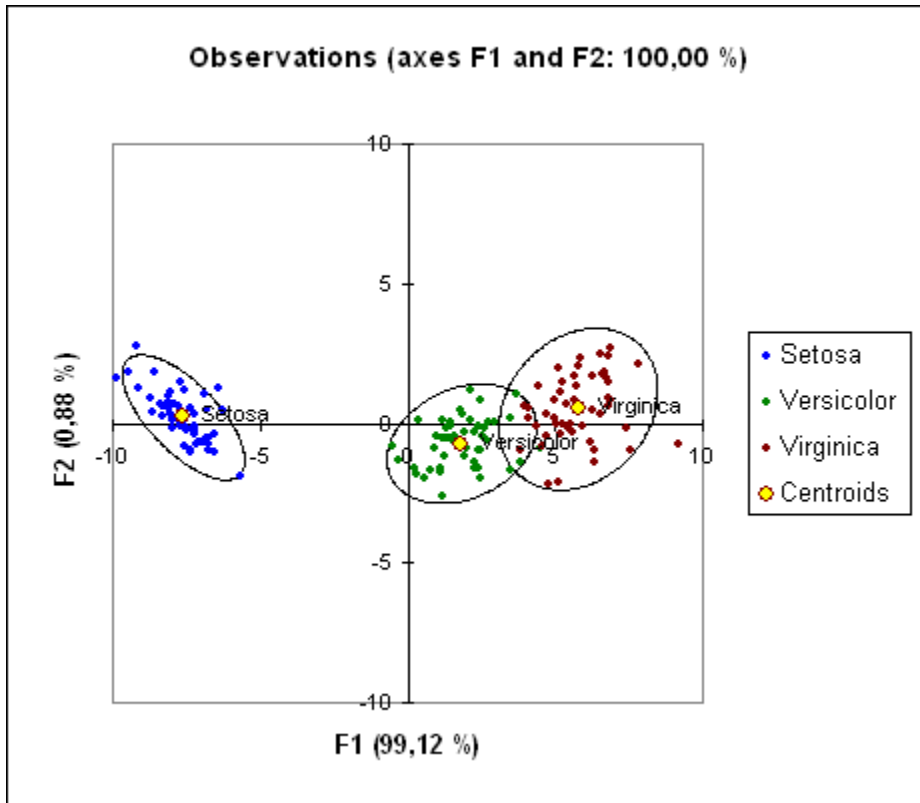
The following table displays the discriminant functions. When we assume the equality of the covariance matrices, the discriminant functions are linear. When the equality is not assumed, which is the case in this tutorial, the discriminant functions are quadratic. The rule based on these functions is that we allocate an observation to the group corresponding to the function that gives the greatest value. These functions can be used in predictive mode on new observations to allocate them to a group.

Classification functions:			
	Setosa	Versicolor	Virginica
Intercept	-121,826	-76,549	-75,821
SEPAL LENGTH	4,455	1,801	0,737
SEPAL WIDTH	-0,762	1,596	1,325
PETAL LENGTH	3,356	0,327	0,623
PETAL WIDTH	-3,126	-1,471	0,966
SEPAL LENGTH*SEPAL LENGTH	-0,095	-0,048	-0,053
SEPAL LENGTH*SEPAL WIDTH	0,124	0,037	0,035
SEPAL LENGTH*PETAL LENGTH	0,045	0,086	0,100
SEPAL LENGTH*PETAL WIDTH	0,048	-0,065	-0,018
SEPAL WIDTH*SEPAL WIDTH	-0,078	-0,099	-0,079
SEPAL WIDTH*PETAL LENGTH	-0,011	-0,021	-0,011
SEPAL WIDTH*PETAL WIDTH	0,021	0,195	0,085
PETAL LENGTH*PETAL LENGTH	-0,194	-0,099	-0,067
PETAL LENGTH*PETAL WIDTH	0,179	0,269	0,029
PETAL WIDTH*PETAL WIDTH	-0,530	-0,436	-0,097

The next table lists for each observation the factor scores (the coordinates of the observations in the new space), the probability to belong to each group, and the squared Mahalanobis distances to the centroid of the group. Each observation is classified into the group for which the probability of belonging is the greatest. The probabilities are posterior probabilities that take into account the prior probabilities through the Bayes formula. We notice that three observations (5,9,12) have been reclassified. There are several ways in which these results can be interpreted: either the person who made the measures made an error when recording the values, or the corresponding iris flowers have had a very unusual growth or the criteria used by the specialist to determine the species was not precise enough, or some information necessary to discriminate the flowers is not available here.

Prior and posterior classification, membership probabilities, scores and squared distances:										
Observatio	Prior	Posterior	Pr(Setosa)	(Versicolor)	(Virginica)	F1	F2	D <sup>2</sup> (Setosa)	(Versicolc <sup>2</sup>	(Virginica)
Obs1	Setosa	Setosa	1,000	0,000	0,000	-7,672	-0,135	5,848	108,391	179,145
Obs2	Virginica	Virginica	0,000	0,000	1,000	6,800	0,581	770,307	42,266	11,244
Obs3	Versicolor	Versicolor	0,000	0,997	0,003	2,549	-0,472	423,336	10,002	21,859
Obs4	Virginica	Virginica	0,000	0,000	1,000	6,653	1,805	813,456	53,028	11,666
Obs5	Virginica	<b>Versicol</b>	0,000	<b>0,605</b>	0,395	3,815	-0,943	520,064	12,926	13,778
Obs6	Setosa	Setosa	1,000	0,000	0,000	-7,213	0,356	8,778	103,550	163,512
Obs7	Virginica	Virginica	0,000	0,000	1,000	5,106	1,992	683,968	55,639	17,807
Obs8	Versicolor	Versicolor	0,000	0,813	0,187	3,498	-1,685	431,446	20,036	22,976
Obs9	Versicolor	<b>Virginica</b>	0,000	0,336	<b>0,664</b>	3,716	1,045	488,109	16,061	14,698
Obs10	Setosa	Setosa	1,000	0,000	0,000	-8,681	0,878	16,398	138,753	205,927
Obs11	Versicolor	Versicolor	0,000	0,997	0,003	2,292	-0,333	391,790	8,841	20,433
Obs12	Versicolor	<b>Virginica</b>	0,000	0,154	<b>0,846</b>	4,498	-0,883	534,065	15,635	12,233
Obs13	Virginica	Virginica	0,000	0,001	0,999	4,968	0,821	623,539	24,287	10,605

The following chart represents the observations on the factor axes. It allows you to confirm that the species are very well discriminated on the factor axes extracted from the original explanatory variables.



The confusion matrix summarizes the reclassification of the observations, and allows to quickly see the % of well classified observations, which is the ratio of the number of observations that have been well classified over the total number of observations. It is here equal to 98%.

Confusion matrix for the estimation sample:

from \ to	Setosa	Versicolor	Virginica	Total	% correct
Setosa	50	0	0	50	100,00%
Versicolor	0	48	2	50	96,00%
Virginica	0	1	49	50	98,00%
Total	50	49	51	150	98,00%

As the corresponding option has been activated in the "Outputs" tab of the dialog box, the predictions for the cross-validation are computed. Cross-validation allows to see what would be the prediction for a given observation if it is left out of the estimation sample. We can see here that only one more observation (Obs8) is miss-classified.

Cross-validation: Prior and posterior classification, membership probabilities, scores and squared distances:					
Observation	Prior	Posterior	Setosa	Versicolor	Virginica
Obs1	Setosa	Setosa	1,000	0,000	0,000
Obs2	Virginica	Virginica	0,000	0,000	1,000
Obs3	Versicolor	Versicolor	0,000	0,997	0,003
Obs4	Virginica	Virginica	0,000	0,000	1,000
Obs5	Virginica	<b>Versicolor</b>	0,000	0,665	0,335
Obs6	Setosa	Setosa	1,000	0,000	0,000
Obs7	Virginica	Virginica	0,000	0,000	1,000
Obs8	Versicolor	<b>Virginica</b>	0,000	0,290	0,710
Obs9	Versicolor	<b>Virginica</b>	0,000	0,154	0,846
Obs10	Setosa	Setosa	1,000	0,000	0,000
Obs11	Versicolor	Versicolor	0,000	0,997	0,003
Obs12	Versicolor	<b>Virginica</b>	0,000	0,068	0,932

The confusion matrix of the cross-validation is displayed below.

Confusion matrix for the cross-validation results:					
from \ to	Setosa	Versicolor	Virginica	Total	% correct
Setosa	50	0	0	50	100,00%
Versicolor	0	47	3	50	94,00%
Virginica	0	1	49	50	98,00%
Total	50	48	52	150	97,33%