AFD - RESUMEN

Este enfoque conduce a s combinaciones lineales, donde s = Mín {p (número de variables), m-1 (número de grupos menos uno)}.
Encontrar estas combinaciones lineales, no correlacionadas dentro de los grupos, es un problema de valores propios.

- ✓ Se manejaron las pruebas de significación para determinar cuántas combinaciones lineales se necesitan para describir las diferencias de grupo. Algunas pruebas que se utilizan comúnmente pueden no proporcionar buenos resultados.
- ✓ Se analizaron los supuestos del análisis de la función discriminante estándar (normalidad e igualdad de matrices de covarianza intragrupo).

AFD - RESUMEN

- ✓ Se consideraron dos ejemplos.
- ✓ Se analizaron opciones para variar el AFD (probabilidades a-priori de pertenecer a un grupo, análisis paso a paso, y clasificación de las observaciones denominada "cuchillo de Jack").
- ✓ Extensiones a los métodos que cubre el curso (modelo cuadrático)

APLICACIONES CON XLSTAT

¿Cómo realizar un Análisis Discriminante con XLSTAT?

Datos: 150 flores de la familia Iris, definidas por 4 variables cuantitativas (Longitud-Sépalos, Anchura-Sépalos, Longitud-Pétalos, Anchura-Pétalos) y por su especie [Fisher M. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, pp 179 -188]. Tres diferentes especies forman parte de este estudio: setosa, versicolor y virginica.

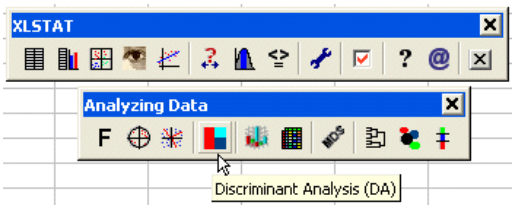
APLICACIONES CON XLSTAT

Iris setosa, versicolor y virginica

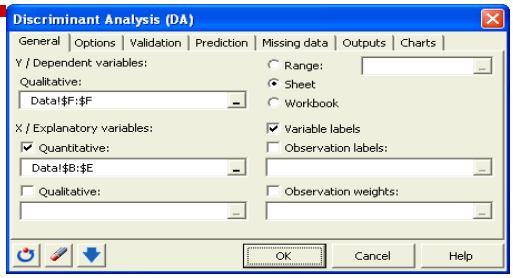
Objetivo: Probar si las cuatros variables descriptivas permiten identificar las especies, y visualizar los datos en un gráfico con el fin de comprobar que las tres especies pueden diferenciarse correctamente.

APLICACIONES CON XLSTAT

Inicie XLSTAT y seleccione el comando XLSTAT/Análisis de los datos/Análisis Factorial Discriminante o haga clic en el botón "Análisis Factorial Discriminante" de la barra de herramientas "Análisis de los datos".



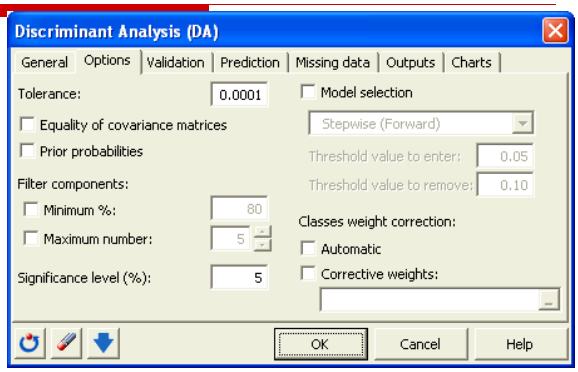
APLICACIONES CON XLSTAT



Cuadro de diálogo del AFD

Selección de datos en la hoja Excel: 1) La "Variable dependiente" representa la variable explicada (grupos), que en este caso es la especie de Iris. 2) Las "variables explicativas" son las cuatros variables medidas. La opción "Etiquetas de las columnas" se activa puesto que la primera fila de las columnas incluye el nombre de las variables.

APLICACIONES CON XLSTAT



Opciones: Hemos **deshabilitado** la opción "Igualdad de las matrices de covarianza intra-grupos", ya que como lo veremos más tarde (Prueba de Box), efectuar semejante hipótesis no es correcto.

APLICACIONES CON XLSTAT

Resultados

XLSTAT visualiza las matrices utilizadas en los cálculos. Las dos pruebas de Box permiten confirmar que no se puede suponer que las matrices de covarianza sean idénticas para las 3 especies.

Box's test (Chi-square asymptotic approximation):

Chi-square	142.373
Chi-square	31.410
DF	20
One-tailed	< 0.0001
Alpha	0.05

Decision:

At the level of significance Alpha= 0.050 the decision is to reject the null hypothesis of equality of the within-groups covariance matrices.
In other words, the difference between the within-groups covariance matrices is significant.

Box's test (Fisher's F asymptotic approximation):

F (observe	7.113
F (critical v	1.573
DF 1	20
DF 2	27150
One-tailed	< 0.0001
Alpha	0.05

Decision:

At the level of significance Alpha= 0.050 the decision is to reject the null hypothesis of equality of the within-groups covariance matrices.
In other words, the difference between the within-groups covariance matrices is significant.

APLICACIONES CON XLSTAT

La prueba "Lambda de Wilk" permite probar si los vectores de las medias para los diferentes grupos son iguales o no [esta prueba se puede percibir como un equivalente multidimensional de la prueba LSD ("menor diferencia significativa") de Fisher o de la prueba HSD ("diferencia honestamente significativa") de Tukey]. Aquí observamos que la diferencia entre los vectores es significativa para un nivel de significancia de 0.05.

Wilks' Lambda test:	
Lambda	0.023
F (observe	199.145
F (critical v	1.980
DF 1	8
DF 2	288
One-tailed	< 0.0001
Alpha	0.05
The F-value is computing according to the Rao's approximation	
Decision:	
At the level of significance Alpha=0.050 the decision is to reject the null hypothesis of equality of mean vectors of the 3 groups.	
In other words, the difference between the groups centroids is significant.	

APLICACIONES CON XLSTAT

En la siguiente tabla se visualizan las **funciones discriminantes**. Cuando se supone que las matrices de covarianza son iguales, estas funciones son lineales. En el caso contrario, son cuadráticas, como es el caso aquí.

Regla básica: se le atribuye una observación al grupo cuya F.D. proporciona el valor más elevado.

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

APLICACIONES CON XLSTAT

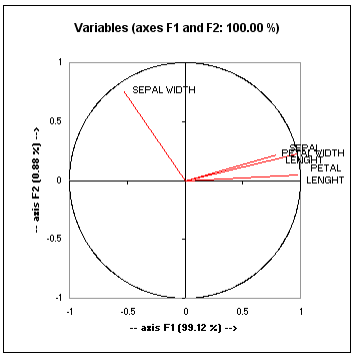
La siguiente tabla proporciona los **valores propios** y el **% de varianza** correspondiente. En este caso, el 99% de la varianza está explicada por el primer factor. Aparecen solamente dos factores: En efecto, el número máximo de factores no nulos es k-1, cuando n>p>k, donde n = nro. de observaciones, p = nro. de variables explicativas y k = nro. de grupos.

Eigenvalues and percentage of variance:

	F1	F2
Eigenvalue	32.192	0.285
% variance	99.121	0.879
% cumulat	99.121	100.000

APLICACIONES CON XLSTAT

El siguiente gráfico muestra que las cuatro primeras variables están correlacionadas con los dos factores obtenidos (este gráfico está construido a partir de la tabla de las coordenadas de las variables). Se puede observar que el factor F1 está correlacionado con Long. Sép., Long. Pét. y Anch. Pét. y que F2 está correlacionado con Anch. Pét.



APLICACIONES CON XLSTAT

La siguiente tabla enumera para cada flor, **sus coordenadas factoriales**, **la probabilidad de asignación a cada grupo, y el cuadrado de las distancias de Mahalanobis respecto al centroide de cada grupo.**

Cada observación es reclasificada en el grupo para el cual la probabilidad es máxima. Las probabilidades son probabilidades a posteriori, que toman en cuenta las probabilidades a priori a través de la fórmula de Bayes.

APLICACIONES CON XLSTAT

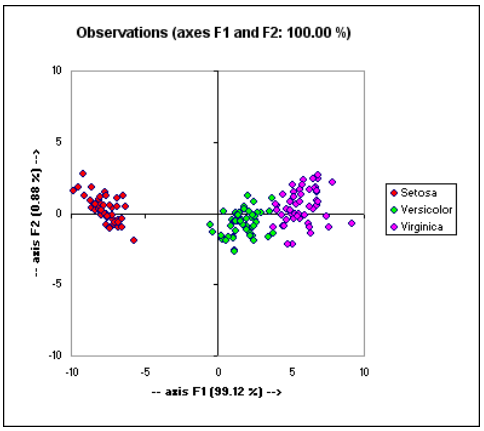
Prior-classification, post-classification, membership probability, observation scores and squared distances to the groups' centroids:

Observatio	Prior	Post	?rob. Setos	ob. Versic	rob. Virgini	F1	F2	entroid(Sentroid(Versentroid(Virginica))		
Obs1	Setosa	Setosa	1.000	0.000	0.000	-1.528	6.528	5.848	108.391	179.145
Obs2	Virginica	Virginica	0.000	0.000	1.000	2.444	7.244	770.307	42.266	11.244
Obs3	Versicolor	Versicolor	0.000	0.997	0.003	1.277	6.191	423.336	10.002	21.859
Obs4	Virginica	Virginica	0.000	0.000	1.000	2.404	8.469	813.456	53.028	11.666
Obs5	Virginica	Versicolor	0.000	0.605	0.395	1.625	5.720	520.064	12.926	13.778
Obs6	Setosa	Setosa	1.000	0.000	0.000	-1.402	7.019	8.778	103.550	163.512
Obs7	Virginica	Virginica	0.000	0.000	1.000	1.979	8.656	683.968	55.639	17.807
Obs8	Versicolor	Versicolor	0.000	0.813	0.187	1.538	4.978	431.446	20.036	22.976
Obs9	Versicolor	Virginica	0.000	0.336	0.664	1.598	7.708	488.109	16.061	14.698
Obs10	Setosa	Setosa	1.000	0.000	0.000	-1.805	7.541	16.398	138.753	205.927
Obs11	Versicolor	Versicolor	0.000	0.997	0.003	1.207	6.330	391.790	8.841	20.433
Obs12	Versicolor	Virginica	0.000	0.154	0.846	1.812	5.780	534.065	15.635	12.233
Obs13	Virginica	Virginica	0.000	0.001	0.999	1.941	7.485	623.539	24.287	10.605
Obs14	Versicolor	Versicolor	0.000	1.000	0.000	0.877	5.036	252.687	8.955	29.434
Obs15	Virginica	Virginica	0.000	0.033	0.967	1.968	6.637	652.105	17.382	10.633

Las observaciones (5,9,12) fueron reclasificadas. Puede haber varias razones: la persona que efectuó las mediciones ha cometido un error cuando medía, o los iris que corresponden a estos datos han tenido un crecimiento anormal por razones desconocidas, o el criterio de clasificación utilizado por el especialista no es correcto, o falta información para diferenciar perfectamente las especies entre sí.

APLICACIONES CON XLSTAT

El siguiente gráfico representa las observaciones sobre los ejes factoriales. Este gráfico permite confirmar que las observaciones están correctamente discriminadas sobre los ejes factoriales obtenidos a partir de las variables explicativas iniciales.



APLICACIONES CON XLSTAT

Por último, la **matriz de confusión** resume la información que concierne las reclasificaciones de observaciones, y se puede deducir el **índice de error aparente**, que corresponde a la razón del número de observaciones reclasificadas, sobre el número total de observaciones.

Confusion matrix (learning-sample):

	to Setosa	o Versicoloto Virginica		Sum
from Setos	50	0	0	50
	33.33%	0.00%	0.00%	33.33%
from Versi	0	48	2	50
	0.00%	32.00%	1.33%	33.33%
from Virgir	0	1	49	50
	0.00%	0.67%	32.67%	33.33%
Sum	50	49	51	150
	33.33%	32.67%	34.00%	100.00%

Apparent error rate (resubstitution error rate calculated with the learning-sample): 2.00 %