KIWITEC.
HIGH QUALITY TECH COURSES

Curso:

SPSS STATISTICS

Módulo I:

ANÁLISIS DISCRIMINANTE

I. Análisis discriminante:

1. Introducción

Con independencia del área de conocimiento en la que se esté trabajando, es frecuente tener que enfrentarse con la necesidad de identificar las características que permiten diferenciar a dos o más grupos de sujetos. Y, casi siempre, para poder clasificar nuevos casos como pertenecientes a uno u otro grupo: ¿se beneficiará este paciente del tratamiento, o no?, ¿devolverá este cliente el crédito, o no?, ¿se adaptará este candidato al puesto de trabajo, o no?, etc.

A falta de otra información, cualquier profesional se limita a utilizar su propia experiencia o la de otros, o su intuición, para anticipar el comportamiento de un sujeto: el cliente devolverá el crédito o el candidato se adaptará al puesto de trabajo en la medida en se parezcan a los pacientes, clientes o candidatos que se benefician del tratamiento, que devuelven el crédito o que se adaptan a su puesto de trabajo. Pero a medida que los problemas se hacen más complejos y las consecuencias de una mala decisión más graves, las impresiones subjetivas basadas en la propia intuición o experiencia deben ser sustituidas por argumentos más consistentes. El análisis discriminante ayuda a identificar las características que diferencian (discriminan) a dos o más grupos y a crear una función capaza de distinguir con la mayor precisión posible a los miembros de uno u otro grupo.

Obviamente, para llegar a conocer en qué se diferencian los grupos necesitamos disponer de la información (cuantificada en una serie de variables) en la que suponemos que se diferencian. El análisis discriminante es una técnica estadística capaz de determinar qué variables permiten diferenciar a los grupos y cuántas de estas variables son necesarias para alcanzar la mejor clasificación posible. La pertenencia a los grupos, conocida de antemano, se utiliza como variable dependiente (una variable categórica con tantos valores discretos como grupos). Las variables en las que se supone que se diferencian los grupos se utilizan como variables independientes o variables de clasificación (también llamadas variables discriminantes), que deben ser variables cuantitativas continuas o, al menos, admitir un tratamiento numérico con significado.

El objetivo último del análisis discriminante es encontrar la combinación lineal de las variables independientes que mejor permite diferenciar (discriminar) a los grupos. Una vez encontrada esa combinación (la función discriminante) podrá ser utilizada para clasificar nuevos casos. Se trata de una técnica de análisis multivariante que es capaz de aprovechar las relaciones existentes entre una gran cantidad de variables independientes para maximizar la capacidad de discriminación.

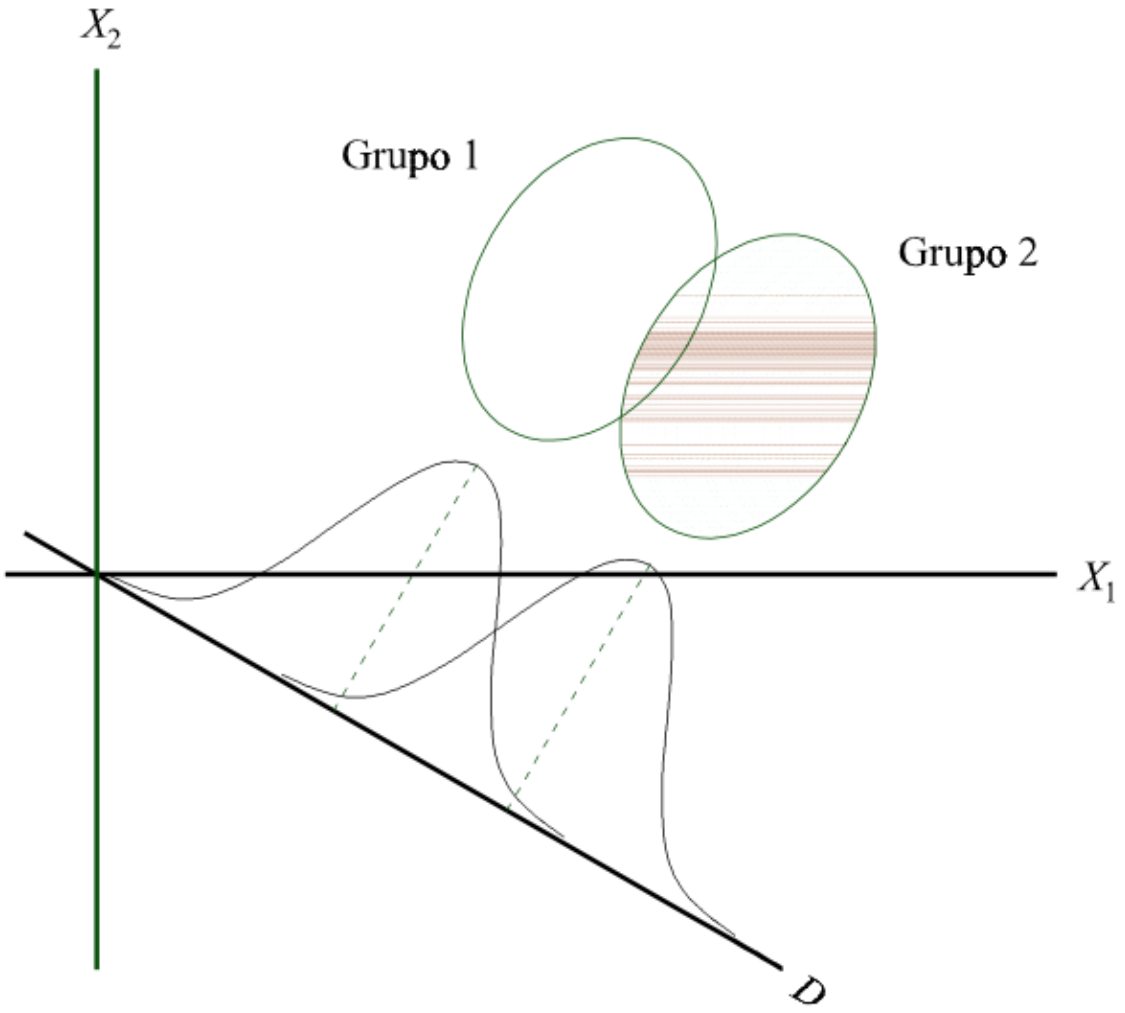
El análisis discriminante es conceptualmente muy similar al análisis de varianza multivariante de un factor. Su propósito es el mismo que el del análisis de regresión logística, pero a diferencia de él, sólo admite variables cuantitativas. Si alguna de las variables independientes es categórica, es preferible utilizar la regresión logística.

II. El caso de dos grupos

Según hemos señalado ya, el análisis discriminante permite diferenciar entre cualquier número de grupos. Sin embargo, por simplicidad, se trata primero el caso de dos grupos, para ampliar posteriormente el razonamiento al caso de k grupos.

En la figura abajo mostrada están representadas, en el espacio bivariante definido por las variables X_1 y X_2, las nubes de puntos correspondientes a dos grupos hipotéticos. Los dos grupos representados se diferencian entre sí en ambas variables, pero no por completo, pues de hecho, se solapan en una pequeña región situada entre ambos.

En la misma figura también está representada la función *D*, que es una combinación lineal de ambas variables. Sobre la función *D* se representa la proyección de las dos nubes de puntos de forma de histograma, como si la función *D* cortara a las dos nubes de puntos en la dirección de su eje. Las dos líneas punteadas de cada uno de los histogramas representan la ubicación proyectada de los puntos medios de cada grupo (los centroides).



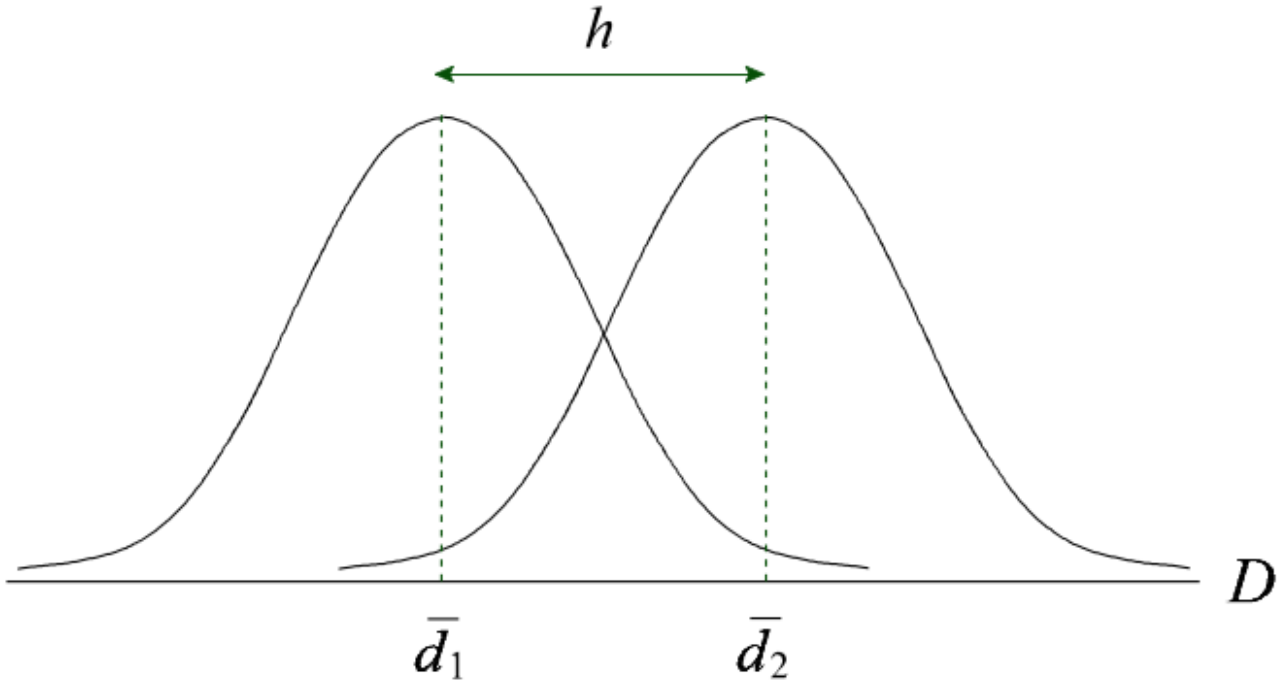
Diagramas de dispersión de dos grupos en dos variables de clasificación.

El propósito del análisis discriminante consiste en aprovechar la información contenida en las variables independientes para crear una función *D* combinación lineal de X_1 y X_2 capaz de diferenciar lo más posible a ambos grupos. La función discriminante es de la forma:

$$D = b_1X_1 + b_2X_2$$

Donde b_1 y b_2 son las ponderaciones de las variables independientes que consiguen hacer que los sujetos de uno de los grupos obtengan puntuaciones máximas en *D*, y los sujetos del otro grupo puntuaciones mínimas.

Una vez hallada la función discriminante D, carece de sentido intentar representar la situación de los grupos en el espacio definido por las variables X_1 y X_2. Conviene más bien centrar el interés en la representación de la función discriminante, que es unidimensional. La representación en *p* dimensiones resulta complicada cuando *p* es mayor de 2 y añade poco o nada a la interpretación de la función. En la siguiente figura se representa la función discriminante *D* extraída del espacio de las variables X_1 y X_2. Los grupos aparecen representados por sus histogramas y las proyecciones de los centroides aparecen marcadas por líneas de puntos.



Histogramas de cada grupo y centroides representados sobre la función discriminante.

Sustituyendo en la función discriminante el valor de las medias del grupo I en las variables X_1 y X_2, obtenemos el centroide del grupo 1:

$$\bar{d}_1 = b_1\bar{x}_1^{(1)} + b_2\bar{x}_2^{(1)}$$

De igual modo, sustituyendo las medias del grupo 2, obtenemos el centroide del grupo 2:

$$\bar{d}_2 = b_1\bar{x}_1^{(2)} + b_2\bar{x}_2^{(2)}$$

La función *D* debe ser tal que la distancia d entre los dos centroides sea máxima, consiguiendo de esta forma que los grupos estén lo más distantes posible. Podemos expresar esta distancia de la siguiente manera:

$$h = \bar{d}_1 - \bar{d}_2$$

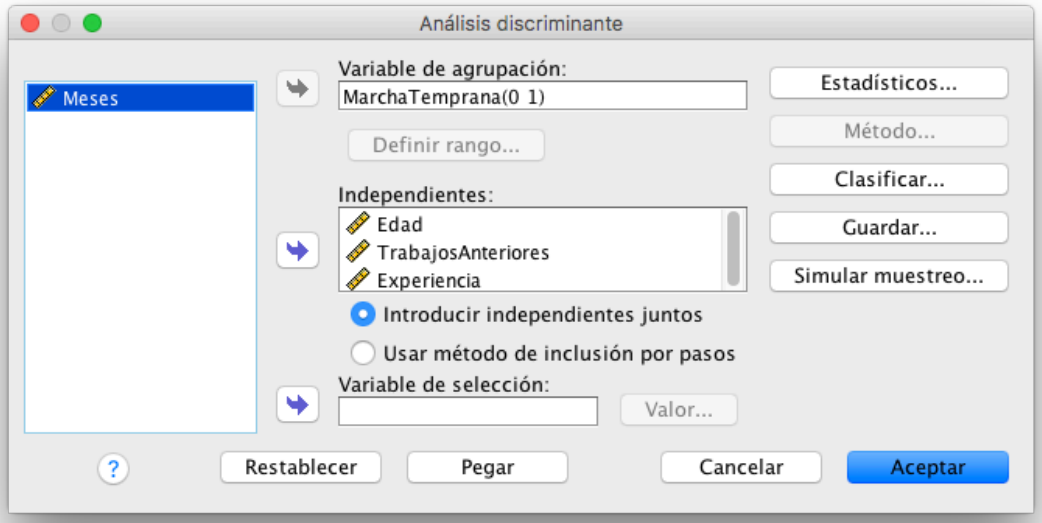
donde $\bar{d}_1$ e $\bar{d}_1$ son las medias del grupo 1 y del grupo 2 en la función D.

El problema radica en reducir la dimensionalidad de las *p* variables independientes a una sola dimensión (la de la combinación lineal *D*) en la que los grupos se diferencien lo más posible. Las puntuaciones de los sujetos en esa nueva dimensión (denominadas puntuaciones discriminantes) serán las que permitan llevar a cabo la clasificación de los sujetos.

Es importante señalar que los grupos deben diferenciarse de antemano en las variables independientes. El análisis busca diferenciar los dos grupos al máximo combinando las variables independientes pero si los grupos no difieren en las variables independientes, el análisis será infructuoso: no podrá encontrar una dimensión en la que los grupos difieran. Dicho de otro modo, si el solapamiento entre los casos de ambos grupos es excesivo, los centroides se encontrarán en la misma o parecida ubicación en el espacio *p*-dimensional y, en esas condiciones, no será posible encontrar una función discriminante útil para la clasificación. Es decir, si los centroides están muy próximos, las medias de los grupos en la función discriminante serán tan parecidas (osea, el valor de *d* será tan pequeño) que no será posible distinguir a los sujetos de uno y otro grupo.

Los supuestos del análisis son los mismos que los del análisis de regresión múltiple. En especial, debe cumplirse que la distribución de las variables independientes sea normal.

Para llevar a cabo un Análisis discriminante se utiliza la opción de menú **Clasificar >Discriminante…** del menú Analizar para acceder al cuadro de diálogo Análisis discriminante.

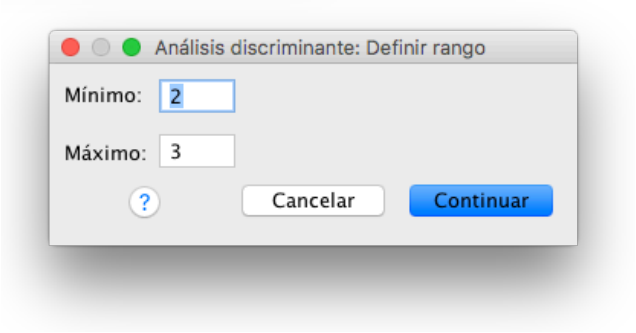


La lista de variables del archivo de datos contiene un listado con todas las variables del archivo excepto las que tienen formato de cadena.

Debe seleccionarse una variable categórica (nominal u ordinal) y trasladarla al cuadro Variable de agrupación. La variable de agrupación es aquella que define los grupos que se desea comparar.

Posteriormente debe seleccionarse al menos una variable cuantitativa (escala) y trasladarla a la lista Independientes. Las variables independientes son aquellas en las que se desea comparar los grupos.

Por último es necesario pulsar sobre el botón **Definir rango…** para acceder al subcuadro de diálogo Definir rango:



Tras seleccionar la variable de agrupación es necesario introducir los códigos que identifican a los grupos que se desea comparar. El análisis incluirá tantos grupos como números enteros consecutivos contenga la variable de agrupación entre los límites del rango definido (ambos límites incluidos).

El análisis discriminante no sólo permite averiguar en qué variables se diferencian los grupos sino, además, construir una función para clasificarlos.

SPSS muestra ofrece un resumen con el total de casos procesados, el número de casos válidos para el análisis y el número de casos excluidos. Dentro de los casos excluidos se distingue entre los que son excluidos porque su código en la variable de agrupación no está dentro del rango seleccionado, los que son excluidos porque tienen un valor perdido en al menos una variable discriminante, y los que cumplen las dos condiciones anteriores.

Resumen de procesamiento de casos de análisis

Casos sin ponderar		N	Porcentaje
Válido		65	100,0
Excluido	Códigos de grupo perdidos o fuera de rango	0	,0
	Como mínimo, falta una variable discriminatoria	0	,0
	Faltan ambos códigos de grupo, los perdidos o los que están fuera de rango y, como mínimo, una variable discriminatoria	0	,0
	Total	0	,0
Total		65	100,0

Posteriormente se muestra otra tabla con un resumen del número de casos válidos en cada variable discriminante. La información de esta tabla posee un interés especial, pues un número desigual de casos en cada uno de los grupos puede afectar a la clasificación.

Estadísticas de grupo

MarchaTemprana		Media	Desviación estándar	N válido (por lista) No ponderados	N válido (por lista) Ponderados
,00	Edad	31,8235	2,91777	34	34,000
	TrabajosAnteriores	1,1471	,78363	34	34,000
	Experiencia	2,6471	2,14451	34	34,000
1,00	Edad	30,5161	3,02072	31	31,000
	TrabajosAnteriores	,8710	1,14723	31	31,000
	Experiencia	1,2581	1,86132	31	31,000
Total	Edad	31,2000	3,01662	65	65,000
	TrabajosAnteriores	1,0154	,97616	65	65,000
	Experiencia	1,9846	2,11758	65	65,000

También se muestra una tabla con los *autovalores* y algunos estadísticos descriptivos multivariantes. Cuando se trabaja con más de dos grupos, se obtiene más de una función discriminante. En estas tablas es posible comparar de manera global la capacidad discriminativa de cada función. En la tabla aparece una fila numerada por cada función discriminante.

El autovalor es el cociente entre la variación debida a las diferencias entre los grupos (medida mediante la *suma de cuadrados inter-grupos*) y la variación que se da dentro de cada grupo combinada en una única cantidad (medida mediante la *suma de cuadrados intra-grupos*). Este estadístico se diferencia de la *F* del análisis de varianza multivariante en que no intervienen los grados de libertad. Su interés principal radica en que permite comparar cómo se distribuye la dispersión *inter-grupos* cuando existe más de una función. Aunque un autovalor tiene un mínimo de cero, no tiene un máximo, lo cual lo hace difícilmente interpretable por sí solo. Por esta razón se acostumbra a utilizar el estadístico *lambda de Wilks*, que se encuentra estrechamente relacionado con los autovalores.

La *correlación canónica* es la correlación entre la combinación lineal de las variables independientes (la función discriminante) y una combinación lineal de variables *indicador* (unos y ceros) que recogen la pertenencia de los sujetos a los grupos. En el caso de dos grupos, la correlación canónica es la correlación simple entre las puntuaciones discriminantes y una variable con códigos 1 y 0 según cada caso pertenezca a un grupo o a otro. Una correlación canónica alta indica que las variables discriminantes permiten diferenciar entre los grupos. Con más de dos grupos, la correlación canónica es equivalente al estadístico eta utilizado en el análisis de varianza de un factor (eta=raíz cuadrada del cociente entre la suma de cuadrados inter-grupos y la suma de cuadrados total).

El estadístico *lambda de Wilks* expresa la proporción de variabilidad total no debida a las diferencias entre los grupos; permite contrastar la hipótesis nula de que las medias multivariantes de los grupos (los centroides) son iguales.

$$\Lambda = \frac{Suma\ de\ cuadrados\ intragrupos}{Suma\ de\ cuadrados\ total}$$

Por tanto, valores próximos a 1 indicarán un gran parecido entre los grupos, mientras que valores próximos a 0 indicarán una gran diferencia entre ellos. Nótese que lambda + eta^2=1.

Es más frecuente utilizar una transformación de Λ que posee distribución aproximada conocida. Bartlett (1947) ha demostrado que el estadístico:

$$V = \left| N - 1 - \frac{(p+g)}{2} \right| 1n\Lambda$$

se aproxima a la distribución *chi-cuadrado* con (*p-k*) (*g-k-1*) grados de libertad: *p* es el número de variables independientes o discriminantes, *g* es el número de grupos, y *k* es el número de funciones discriminantes obtenidas con anterioridad al contraste (cuando sólo existe una función, porque sólo hay dos grupos-, k=0).

Resumen de funciones discriminantes canónicas

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	,149[a]	100,0	100,0	,361

a. Se utilizaron las primeras 1 funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,870	8,562	3	,036

También se muestra una tabla con los *coeficientes estandarizados*, que contiene una versión estandarizada de los coeficientes de la función canónica discriminante. Estos coeficientes estandarizados son independientes de la métrica original de las variables discriminantes y, por tanto, son preferibles a los coeficientes *brutos* cuando las variables poseen una métrica distinta. Son los coeficientes que el programa ofrece por defecto, mientras que los coeficientes *brutos* deben solicitarse de manera explícita.

Coeficientes de función discriminante canónica estandarizados

	Función 1
Edad	,135
TrabajosAnteriores	-,654
Experiencia	1,287

Para interpretar los signos de las ponderaciones resulta útil inspeccionar primero la ubicación de los centroides de cada grupo.

La matriz de estructura contiene las correlaciones entre las variables discriminantes y la función discriminante estandarizada. Mientras que los coeficientes estandarizados muestran la contribución neta de cada variable independiente a la función discriminante (de manera similar a como lo hacen los coeficientes beta de un análisis de regresión múltiple), las correlaciones muestran la relación bruta entre cada variable y la función discriminante.

Matriz de estructuras

	Función 1
Experiencia	,905
Edad	,578
TrabajosAnteriores	,372

Correlaciones dentro de grupos combinados entre las variables discriminantes y las funciones discriminantes canónicas estandarizadas
Variables ordenadas por el tamaño absoluto de la correlación dentro de la función.

Cuando existe colinealidad entre las variables independientes puede ocurrir que alguna de ellas quede fuera del análisis por no aportar información nueva. Sin embargo, conocer estas relaciones puede ayudar a interpretar mejor la función discriminante.

La matriz de estructura presenta las variables ordenadas por su grado de correlación (de mayor a menor) con la función discriminante. Este orden puede ser distinto del orden en el que aparecen en otras tablas y del orden en que han sido incluidas en el análisis.

La siguiente tabla contiene la ubicación de los centroides en la función discriminante (bruta) tal y como se muestran en la tabla de estadísticos por grupos. Esta tabla es de gran utilidad para interpretar la función discriminante.

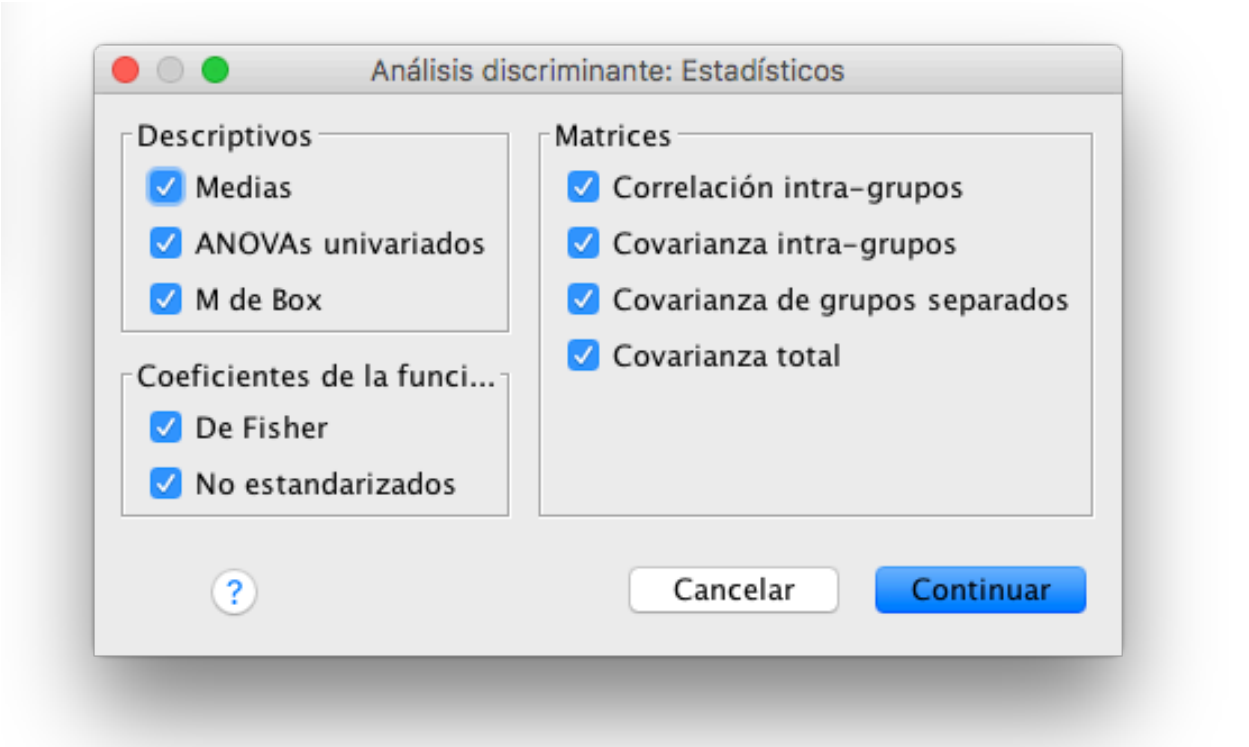
Funciones en centroides de grupo

	Función
MarchaTemprana	1
,00	,363
1,00	-,398

Las funciones discriminantes canónicas sin estandarizar se han evaluado en medias de grupos

1. Estadísticos

El subcuadro de diálogo Estadísticos permite obtener información adicional sobre algunos aspectos del análisis. Parte de esta información es descriptiva, pero también contiene estadísticos que permiten comprobar algunos de los supuestos en los que se fundamenta la técnica. Para obtener esta información:



DESCRIPTIVOS. Este apartado contiene opciones que permiten obtener información descriptiva y constrastes univariantes y multivariantes sobre las variables utilizadas en el análisis:

• **Medias**. Media, desviación típica, número de casos válidos (ponderado y no ponderado) para cada uno de los grupos y para la muestra total.

• **ANOVAs univariados.** Tabla de ANOVA con estadísticos *F* que permiten contrastar la hipótesis de igualdad de medias entre los grupos en cada variable independiente. La tabla de ANOVA incluye también el estadístico *lambda de Wilks* univariante. La información de esta tabla suele utilizarse como prueba preliminar para detectar si los grupos difieren en las variables de clasificación seleccionadas; sin embargo, debe tenerse en cuenta que una variable no significativa a nivel univariante podría aportar información discriminativa a nivel multivariante.

• **M de Box**. Prueba *M* de Box para el contraste de la hipótesis nula de igualdad de las matrices de varianzas-covarianzas poblacionales. Uno de los supuestos del análisis discriminante es que todos los grupos proceden de la misma población y, más concretamente, que las matrices de varianzas-covarianzas poblacionales correspondientes a cada grupo son iguales entre sí. El estadístico *M* carece de distribución muestral conocida, pero puede transformase en un estadístico *F* e interpretarse como tal. La tabla muestra los logaritmos de los determinantes de todas las matrices utilizadas en el cálculo del estadístico *M*. Dado que el estadístico es multivariante, la tabla permite comprobar qué grupos (cuando hay más de dos) difieren más. La tabla ofrece la prueba *M* de Box y su transformación en un estadístico *F*. El resultado de la prueba permite rechazar la hipótesis de igualdad de matrices de varianzas-covarianzas (*Sig.*=0,000 <0,05) y, por tanto, concluir que uno de los dos grupos es más variable que el otro.

Prueba de Box de la igualdad de matrices de covarianzas

Logaritmo de los determinantes

MarchaTemprana	Rango	Logaritmo del determinante
,00	3	2,133
1,00	3	,987
Dentro de grupos combinados	3	2,015

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Resultados de prueba

M de Box		26,912
F	Aprox.	4,252
	gl1	6
	gl2	28071,755
	Sig.	,000

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

MATRICES. Las opciones de este apartado permiten obtener las matrices de varianzas-covarianzas utilizadas en el análisis:

• **Correlación intra-grupos**. Muestra la matriz de correlaciones intra-grupo combinada, es decir la matriz de correlaciones entre las variables independientes estimada a partir de las correlaciones obtenidas dentro de cada grupo. Aparece en la misma ¿TABLA? Que la matriz de varianzas-covarianzas intra-grupos combinada.

• **Covarianza intra-grupos**. Matriz de varianzas-covarianzas intra-grupo combinada. Esta matriz se calcula obteniendo las matrices de sumas de cuadrados y productos cruzados de cada grupo por separado, sumando a continuación las matrices de todos los grupos y dividiendo finalmente por los grados de libertad. Es la matriz S utilizada en el cálculo de la lambda de Wilks. La matriz se ofrece junto a la de correlaciones intra-grupo en una única tabla.

Matrices dentro de grupos combinados[a]

		Edad	TrabajosAnteriores	Experiencia
Covarianza	Edad	8,804	1,936	4,091
	TrabajosAnteriores	1,936	,948	1,425
	Experiencia	4,091	1,425	4,059
Correlación	Edad	1,000	,670	,684
	TrabajosAnteriores	,670	1,000	,726
	Experiencia	,684	,726	1,000

a. La matriz de covarianzas tiene 63 grados de libertad.

• **Covarianza de grupos separados**. Matrices de varianzas-covarianzas de cada grupo. En la tabla, la matriz de cada grupo se presenta precedida de un encabezado que indica el grupo al que se refiere. Las matrices de varianza-covarianza individuales calculadas por separado para cada uno de los grupos se utilizan en ocasiones especiales para obtener una estimación de la matriz de varianzas-covarianzas intra-grupo combinada. La suma de estas matrices sólo será igual a la matriz de varianzas-covarianzas combinada cuando los tamaños de los grupos sean grandes y similares. Estas matrices aparecen en la misma tabla que la matriz de varianzas-covarianzas total.

Matrices de covarianzas[a]

Efectúe una doble pulsación para activar

MarchaTemprana				riencia
,00	Edad	8,513	1,178	4,209
	TrabajosAnteriores	1,178	,614	,963
	Experiencia	4,209	,963	4,599
1,00	Edad	9,125	2,769	3,962
	TrabajosAnteriores	2,769	1,316	1,934
	Experiencia	3,962	1,934	3,465
Total	Edad	9,100	1,997	4,488
	TrabajosAnteriores	1,997	,953	1,500
	Experiencia	4,488	1,500	4,484

a. La matriz de covarianzas total tiene 64 grados de libertad.

• **Covarianza total**. Matriz de varianzas-covarianzas total, es decir, calculada sobre todos los sujetos de la muestra como si pertenecieran a un único grupo. Aparece en la última submatriz, de la ¿TABLA?, con el encabezado *Total*. Es la matriz T utilizada en el cálculo de la *lambda de Wilks*.

• **Coeficientes no tipificados**. Coeficientes brutos de la función canónica discriminante. Son los coeficientes utilizados por el programa para calcular las puntuaciones discriminantes y la ubicación de los centroides de los grupos. La función discriminante incluye una constante correctora que consigue que las puntuaciones discriminantes tomen el valor 0 en algún punto entre los dos centroides.

Coeficientes de la función discriminante canónica

	Función 1
Edad	,046
TrabajosAnteriores	-,671
Experiencia	,639
(Constante)	-2,007

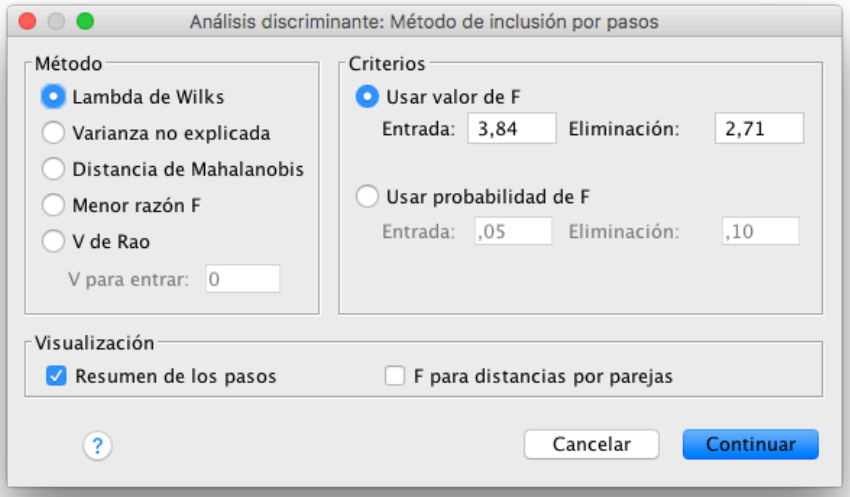
Coeficientes no estandarizados

• **Coeficientes de clasificación de Fisher**. Fisher (1936) presentó la primera aproximación a la clasificación multivariante para el caso de dos grupos. Los coeficientes propuestos por Fisher se utilizan únicamente para la clasificación. Al solicitar esta opción se obtiene una función de clasificación para cada grupo. En el caso de dos grupos, la diferencia entre ambas funciones da lugar a un vector de coeficientes proporcional a los coeficientes no tipificados de la función discriminante canónica. Para aplicar estos coeficientes, se calcula cada una de las funciones para un sujeto dado y se clasifica al sujeto en el grupo en el que la función obtiene una puntuación mayor. En la práctica, el programa no utiliza estos coeficientes para la clasificación de los sujetos.

2. Método

Las variables independientes pueden incorporarse a la función discriminante utilizando dos estrategias distintas. Por defecto, el SPSS utiliza una estrategia de inclusión forzosa de variables que permite construir la función discriminante incorporando todas las variables independientes incluidas en el análisis. Según hemos visto en los ejemplos anteriores, los únicos estadísticos que se obtienen con esta estrategia se refieren al ajuste global de la función discriminante; no se obtienen estadísticos referidos a la significación individual de cada coeficiente discriminante (como, por ejemplo, los estadísticos t del análisis de regresión múltiple).

Una manera de obtener información sobre la significación individual de cada variable en la función discriminante consiste en utilizar una estrategia de inclusión por pasos. Con esta estrategia, las variables se van incorporando a la función discriminante una a una y, de esta manera, es posible, por un lado, construir una función utilizando únicamente aquellas variables que realmente son útiles para la clasificación y, por otra, evaluar la contribución individual de cada variable al modelo discriminante. Para utilizar esta estrategia de inclusión por pasos hay que seleccionar la opción **Por pasos**.



En la estrategia de inclusión por pasos, las variables independientes van siendo incorporadas paso a paso a la función discriminante tras evaluar su grado de contribución individual a la diferenciación

entre los grupos. Las opciones de este apartado permiten seleccionar el estadístico que será utilizado como método de selección de variables:

• **Lambda de Wilks**. Cada variable independiente candidata a ser incluida en el modelo se evalúa mediante un estadístico F_{cambio} que mide el cambio que se produce en el valor de la lambda de Wilks al incorporar cada una de las variables al modelo. Obtenido el valor del estadístico F_{cambio} para cada variable, se incorpora al modelo la variable a la que le corresponde el mayor valor F_{cambio} (o, lo que es lo mismo, la que produce el mayor cambio en la lambda de Wilks):

$$F_{cambio} = \left(\frac{n-g-p}{g-1}\right)\left(\frac{1-\lambda_{p+1}/\lambda_p}{\lambda_{p+1}}\right)$$

donde *n* es el número de casos válidos, *g* es el número de grupos, λ_p es la *lambda* de Wilks que corresponde al modelo antes de incluir la variable que se está evaluando y λ_{p+1} es la *lambda* de Wilks que corresponde al modelo después de incluir esa variable. Este estadístico *F* es también conocido como *R* de Rao (ver Tatsuoka, 1971).

• **Varianza no explicada.** Utiliza como criterio de inclusión la suma de la variación entre todos los pares de grupos no explicada por las variables ya incluidas en el modelo. Se incorpora al modelo la variable que minimiza la cantidad de varianza no explicada. La cantidad de varianza explica por el modelo, R^2, es proporcional, en una constante c, a la distancia *H* de Mahalanobis:

$$R^2 = cH_{ab}^2 \text{ c}$$

Para calcular la cantidad de varianza no explicada se utiliza el estadístico *R* (Dixon, 1973):

$$R = \sum_{a=1}^{g-1} \sum_{b=a+1}^{g} \frac{4}{4+H_{ab}^2}$$

donde *g* es el número de grupos, y a y b son dos grupos cualesquiera.

• **Distancia de Mahalanobis.** Se incorpora en cada paso la variable que maximiza la distancia de Mahalanobis (1936) entre los dos grupos más próximos. La distancia multivariante entre los grupos a y b se define como:

donde *n* es el número de casos válidos, *g* es el número de grupos, $\bar{X}_i^{(a)}$ es la media del grupo *a* en la *i*-ésima variable independiente, $\bar{X}_i^{(b)}$ es la media del grupo *b* en la *i*-ésima variable independiente,

y w_{ij}^{*} es un elemento de la inversa de la matriz de varianzas-covarianzas intra-grupos. Morrison (1976).

• **Menor razón *F*.** Se incorpora en cada paso la variable que maximiza la menor razón *F* para las parejas de grupos. El estadístico *F* utilizado es la distancia de Mahalanobis ponderada por el tamaño de los grupos:

• **V de Rao.** El estadístico V de Rao (1952) es una transformación de la traza de Lawley- Hotelling (Lawley, 1938; Hotelling, 1931) que es directamente proporcional a la distancia entre los grupos. Al utilizar este criterio, la variable que se incorpora al modelo es aquella que produce un mayor incremento en el valor de *V*:

donde *p* es el número de variables en el modelo, *g* es el número de grupos, n_k es el número de casos válidos del grupo *k*, $\bar{X}_1^{(k)}$ es la media del grupo *k* en la *i*-ésima variable, $\bar{X}_i$ es la media de todos los grupos en la *i*-ésima variable, y w_{ij}^{*} es un elemento de la inversa de la matriz de varianzas-covarianzas intra-grupos.

Esta opción permite especificar el incremento mínimo que se tiene que dar en el valor de *V* para que una variable pueda ser incorporada al modelo. Para establecer ese mínimo, introducir un valor mayor que 0 en el cuadro de texto ***V para entrar***.

i. Criterios

Cualquiera que sea el método seleccionado, en la estrategia de inclusión por pasos siempre se comienza seleccionando la mejor variable independiente desde el punto de vista de la clasificación (es decir, la variable independiente en la que más se diferencian los grupos). Pero esta variable sólo es seleccionada si cumple el criterio de entrada. A continuación, se selecciona la variable independiente que, cumpliendo el criterio de entrada, más contribuye a conseguir que la función discriminante diferencie a los grupos. Etc. Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son evaluadas nuevamente para determinar si cumplen o no el criterio de salida. Si alguna variable de las ya seleccionadas cumple el criterio de salida, es expulsada del modelo.

Las opciones de este apartado permiten establecer los criterios de *entrada* y *salida* utilizados por el programa para incorporar o eliminar variables. De acuerdo con estos criterios, sólo son incluidas en el modelo aquellas variables que contribuyen a discriminar significativamente entre los grupos:

• **Usar valor de *F*.** Una variable pasa a formar parte de la función discriminante si el valor del estadístico *F* es mayor que 3,84 (valor de *entrada*). Y es expulsada de la función si el valor del estadístico *F* es menor que 2,71 (valor de *salida*). Para modificar los valores de entrada y salida es necesario seleccionar el criterio **Usar valor de *F*** (si no está ya seleccionado) e introducir los valores deseados (siempre mayores que 0) en los cuadros de texto **Entrada y Salida**. El valor de entrada debe ser mayor que el de salida.

• **Usar la probabilidad de *F***. Una variable pasa a formar parte de la función discriminante si el nivel crítico asociado al valor del estadístico *F* es menor que 0,05 (probabilidad de *entrada*). Y es expulsada de la función si ese nivel crítico es mayor que 0,10 (probabilidad de *salida*). Para modificar los valores de *entrada* y *salida* debe seleccionarse el criterio **Usar valor de *F*** (si no está ya seleccionado) e introducir los valores deseados (siempre entre 0 y 1) en los cuadros de texto Entrada y Salida. El valor de entrada debe ser menor que el de salida.

Mostrar. Las opciones de este apartado permiten obtener información detallada sobre algunos aspectos relacionados con el proceso de inclusión por pasos:

• **Resumen de los pasos.** Estadísticos para cada una de las variables después de cada paso, así como estadísticos de resumen del paso. Para omitir de los resultados la información sobre el proceso por pasos, desactive esta selección.

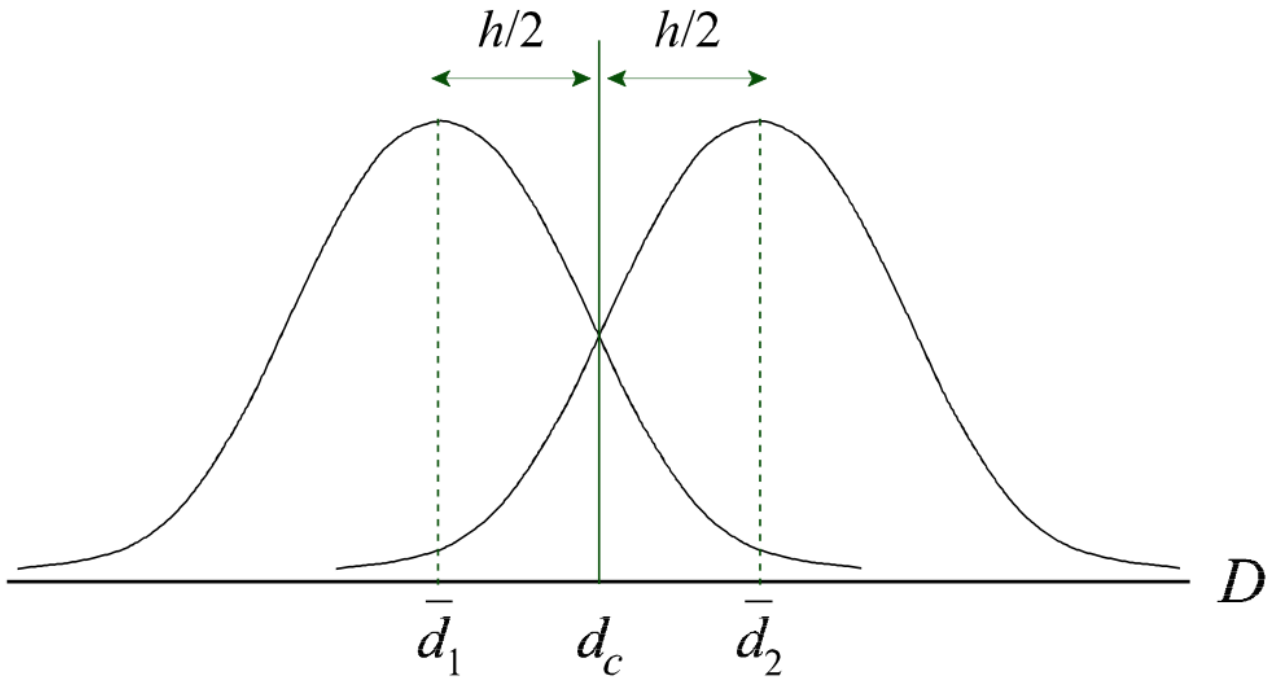
• ***F* para distancias por parejas.** Muestra una matriz de estadísticos *F* que contrasta si cada pareja de grupos difieren en la función discriminante. Se comparan todas las parejas de grupos. Esta opción es útil en el caso de más de dos grupos.

III. El problema de la clasificación

Si el objetivo consiste en averiguar en qué difieren dos grupos, con lo visto hasta ahora es más que suficiente. Sin embargo, la mayor utilidad de una función discriminante radica en su capacidad para clasificar nuevos casos. Ahora bien, la clasificación de casos es algo muy distinto de la estimación de la función discriminante. De hecho, una función perfectamente estimada puede no pasar de una pobre capacidad clasificatoria.

Una vez obtenida la función discriminante puede utilizarse para efectuar una clasificación de los mismos casos utilizados para obtener la función: esto permitirá comprobar el grado de eficacia la función desde el punto de vista de la clasificación. Si los resultados son satisfactorios, la función discriminante podrá utilizarse, en segundo lugar, para clasificar futuros casos de los que, conociendo su puntuación en las variables independientes, se desconozca el grupo al que pertenecen.

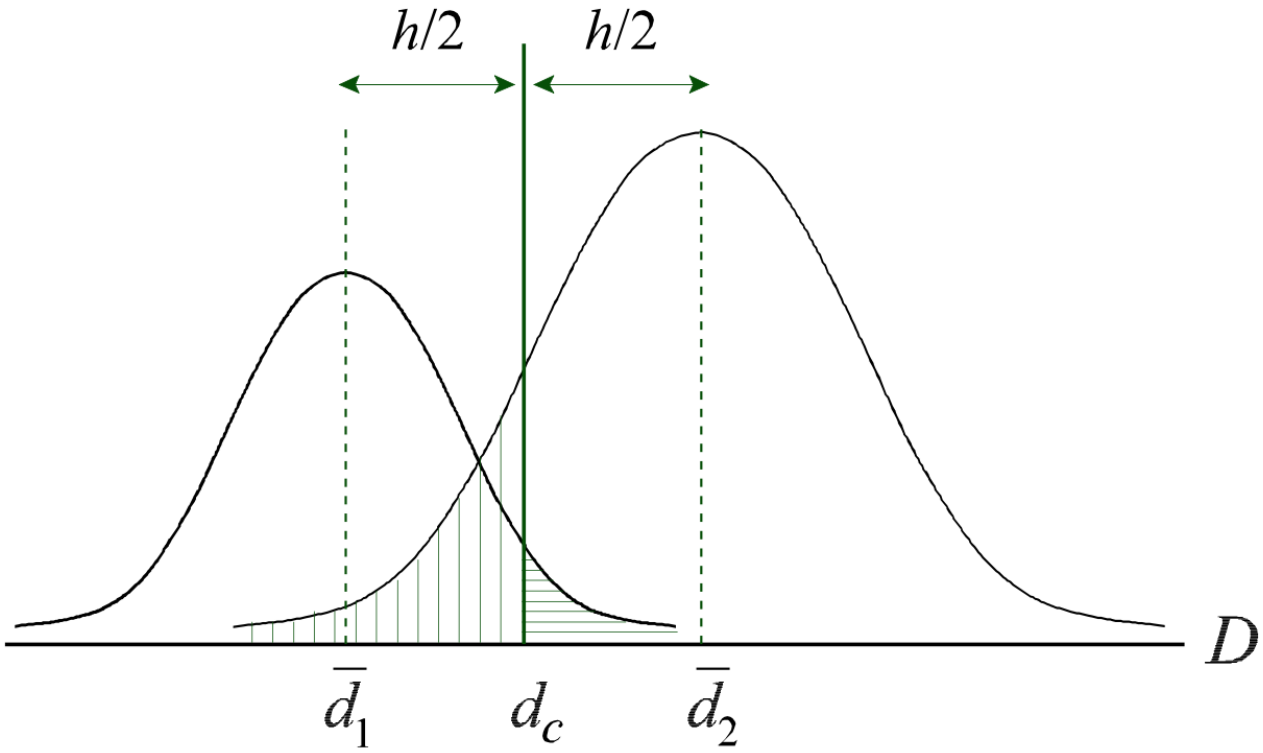
Una manera de clasificar los casos consiste en calcular la distancia existente entre los centroides de ambos grupos y situar un punto de corte d_c equidistante de ambos centroides (ver figura inferior). A partir de ese momento, los casos cuyas puntuaciones discriminantes sean mayores que el punto de corte d_c serán asignados al grupo superior y los casos cuyas puntuaciones discriminantes sean menores que el punto de corte d_c serán asignados al grupo inferior.



Utilización de un punto de corte equidistante de ambos centroides ($n_1 = n_2$).

Esta regla de clasificación tiene un serio inconveniente: sólo permite distinguir entre dos grupos y es difícilmente aplicable al caso de más de dos grupos. Además, no tiene en cuenta que los grupos pueden tener distinto tamaño. Si ambos grupos son de igual tamaño, la situación real será muy similar

a la descrita en la figura anterior. Pero si, por el contrario, los tamaños muestrales son muy desiguales, la situación real será más parecida a la que muestra la figura siguiente. En esta figura puede verse con claridad que, si utilizamos el punto de corte d_c como punto de clasificación, la proporción de casos mal clasificados en el grupo de menor tamaño (zona rayada horizontalmente) será mucho menor que en el grupo de mayor tamaño (zona rayada verticalmente). Por tanto, con tamaños desiguales es preferible utilizar una regla de clasificación que desplace el punto de corte hacia el centroide del grupo de menor tamaño buscando igualar los errores de clasificación. Para calcular este punto de corte podemos utilizar una distancia ponderada:



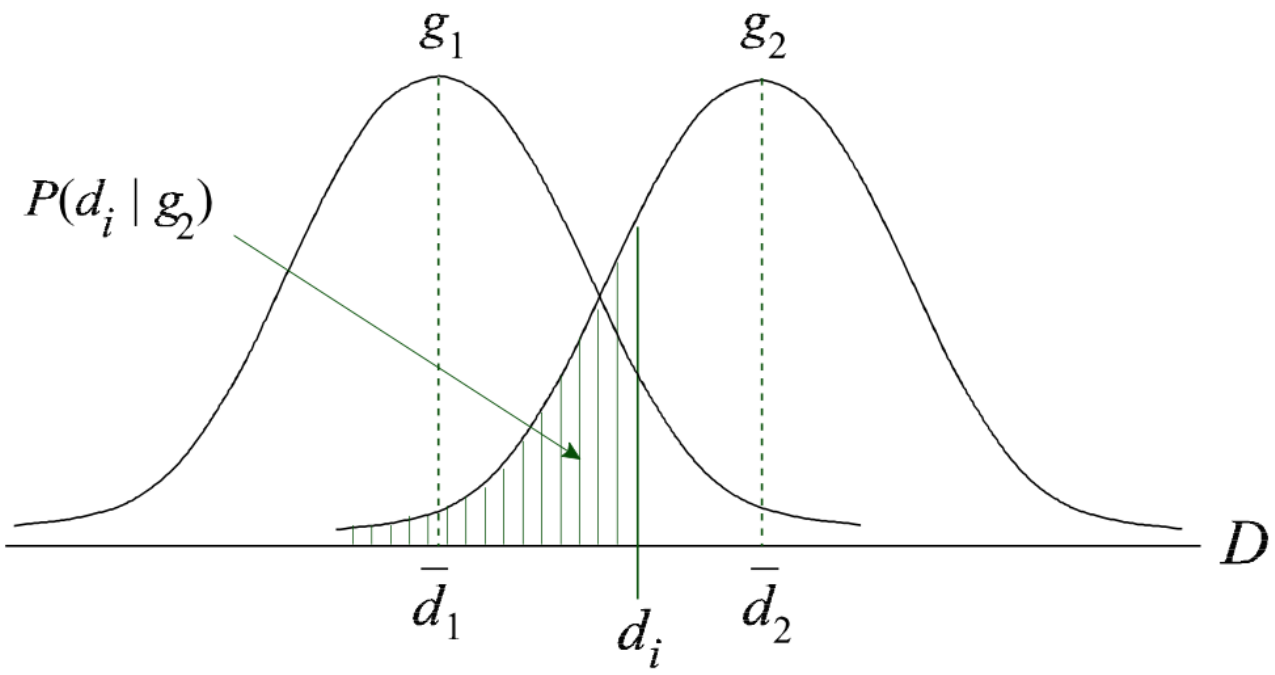
Utilización de un punto de corte equidistante de ambos centroides ($n_1 \neq n_2$).

Fukunaga y Kessell (1973) y Glick (1978) han propuesto una regla de clasificación basada en la teoría bayesiana. Esta otra regla permite incorporar fácilmente la información relativa al tamaño de los grupos y, además, es extensible al caso de más de dos grupos.

Es frecuente que, aunque los tamaños de los grupos sean intrínsecamente diferentes, se desee compensar estadísticamente esa desigualdad a la hora de clasificar a los sujetos. Esta situación es muy frecuente en el ámbito clínico cuando se comparan sujetos normales con sujetos enfermos. Si podemos estimar la proporción de casos que, en la población, pertenece a cada o de los grupos, tendremos una probabilidad a priori: $P(g_k)$. Estas probabilidades a priori pueden estimarse a partir de la muestra (si se ha realizado un muestreo aleatorio), o recurriendo directamente a datos poblacionales previos (si se tienen).

Las probabilidades a priori ofrecen alguna información sobre la representatividad de los casos, pero no ofrecen información concreta sobre un caso en particular. Además, las probabilidades a priori no tienen en cuenta que las probabilidades de aparición de las variables independientes en cada grupo pueden no ser simétricas.

Por supuesto, siempre es posible aprovechar la información adicional que proporciona saber a qué grupo pertenece cada caso. Si asumimos que las puntuaciones discriminantes se distribuyen normalmente, podemos calcular la probabilidad asociada a un caso (es decir, la probabilidad que queda por encima o por debajo de ese caso) en cada uno de los grupos utilizados en el análisis. Esto es lo que se conoce como probabilidad condicional: $P(D > d_1|G = g_k$ o, simplemente, $P(d_1|g_k)$. La probabilidad condicional de una puntuación discriminante puede calcularse mediante tablas de probabilidad asintótica o a partir de los cuantiles observados.



Probabilidad condicional de la puntuación discriminante d_1, en el grupo 2.

Una puntuación discriminante tiene asociadas tantas probabilidades condicionales como grupos hay en el análisis. Esas probabilidades condicionales indican cómo es de probable una puntuación concreta en cada uno de los grupos. Pero sólo son útiles cuando se conoce a qué grupo pertenece un caso. Cuando se desea clasificar un caso nuevo (del que, obviamente se desconoce a qué grupo pertenece), es necesario comparar las probabilidades condicionales que le corresponden en cada uno de los grupos del análisis. Por ello, para clasificar un caso nuevo, es más apropiado utilizar las probabilidades a posteriori, es decir, las probabilidades de pertenecer a cada uno de los grupos, dado que a ese caso le corresponde una determinada puntuación discriminate, es decir: $P(G = g_k|D = d_i)$ o, simplemente, $P(g_k|d_i)$. Estas probabilidades a posteriori se obtienen utilizando el teorema de Bayes.

Aunque en la estimación de las probabilidades a priori es habitual utilizar los tamaños de los grupos, la aplicación del teorema de Bayes permite manipular esas probabilidades y asignarles un valor arbitrario (para reflejar mejor la composición de la población, para compensar el coste de una clasificación errónea, etc.). La manipulación de las probabilidades a priori hace que se desplace el punto de clasificación. Si se asigna igual probabilidad a priori a todos los grupos, el punto de corte

para la clasificación será equidistante de todos ellos; si se aumenta la probabilidad a priori de un grupo, el punto de corte para la clasificación se alejará de su centroide.

Una forma más de determinar el punto de corte óptimo para la clasificación consiste en la curva COR (curva característica del receptor ideal), disponible como procedimiento adicional dentro del propio SPSS.

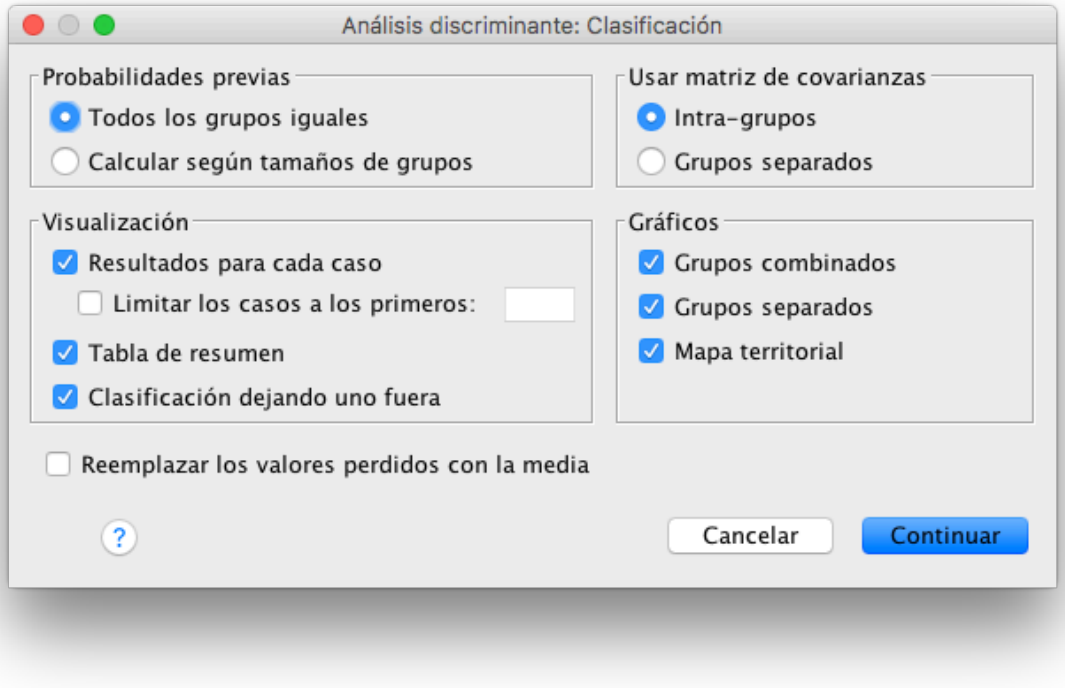
Ninguno de los procedimientos mencionados valora el coste de la clasificación errónea de los sujetos: todos ellos asumen igual coste para los aciertos y los errores en todos los grupos. Si existen costes diferenciales para cada tipo de acierto y para cada tipo de error, será necesario establecer el punto de corte mediante otro tipo de procedimientos más característicos de la *Teoría de la toma de decisiones.*

1. Selección de las opciones de clasificación

Las opciones de clasificación no afectan a la función discriminante; sólo influyen en el resultado de la clasificación de los casos.

El proceso de clasificación asigna o pronostica un grupo a todos los casos utilizados en la estimación de la función discriminante y a todos los casos que, aun no perteneciendo a ninguno de los grupos utilizados (es decir, aun teniendo valor perdido en la variable de agrupación), poseen información completa en las variables independientes. También es posible, opcionalmente, clasificar los casos con información incompleta (es decir, con valor perdido en alguna de las variables independientes).

Para clasificar los casos utilizando la función discriminante hay que acudir al cuadro de Análisis discriminante y pulsar en el botón Clasificar... para acceder al subcuadro de diálogo Análisis discriminante: Clasificación.



Probabilidades previas. Las opciones de este apartado permiten controlar el valor que adoptarán las probabilidades previas o probabilidades *a priori*:

• **Todos los grupos iguales.** Se asigna la misma probabilidad a todos los grupos. Si el análisis discrimina entre *k* grupos, la probabilidad *a priori* asignada a cada grupo vale *1/k*. Con esta opción el tamaño de los grupos no influya en la clasificación.

• **Calcular según el tamaño de los grupos.** La probabilidad a priori que se asigna a cada grupo es proporcional a su tamaño. Siendo *N* el tamaño de la muestra y n_g el tamaño de un grupo cualquiera, la probabilidad a priori asignada a ese grupo es n_g/N. Con esta opción, si un caso posee una puntuación discriminante equidistante de los centroides de dos grupos, el caso es clasificado en el grupo de mayor tamaño.

Mostrar. Estas opciones permiten decidir qué aspectos de la clasificación deseamos que muestre el *Visor de resultados:*

• **Resultados para cada caso.** Muestra un listado de los casos del archivo de datos con el resultado de la clasificación. Esta información incluye, para cada caso: el número del caso en el archivo de datos, el número de variables independientes en las que tiene valor perdido y el grupo al que de hecho pertenece (grupo *nominal*). Además, para el grupo pronosticado con mayor probabilidad: el grupo asignado (marcado con dos asteriscos si difiere del nominal), la probabilidad condicional de obtener una puntuación discriminante como la obtenida o mayor en ese grupo, $P(d_i|g_k)$ la probabilidad a posteriori de ese grupo, $P(g_k|d_i)$, y la distancia de Mahalanobis del caso al centroide de ese grupo. Y para el grupo *pronosticado con la segunda mayor probabilidad:* el grupo asignado, la probabilidad a posteriori de ese grupo, $P(g_k|d_i)$, y la distancia de Mahalanobis al centroide de ese grupo. Por último, el listado ofrece las puntuaciones discriminantes en cada una de las funciones discriminantes obtenidas.

• **Limitar a los primeros *n*.** Permite limitar el listado con los detalles de la clasificación a los primeros *n* casos del archivo. Esta selección sólo afecta a la tabla de resultados para cada caso.

• **Tabla de resumen.** Muestra una tabla de clasificación de tamaño *gxg* con el grupo nominal en las filas y el grupo pronosticado en las columnas. La tabla ofrece las frecuencias absolutas, los porcentajes de fila y el porcentaje total de clasificaciones correctas. Esta tabla se denomina también matriz de confusión. En la diagonal principal de la matriz se encuentran las clasificaciones correctas.

• **Clasificación dejando uno fuera**. Ofrece una validación cruzada para comprobar la capacidad predictiva de la función discriminante. Para ello, el SPSS genera tantas funciones discriminantes como casos válidos tiene el análisis; cada una de esas funciones se obtiene eliminando un caso; después, cada caso es clasificado utilizando la función discriminante en la que no ha intervenido. La tabla de clasificación incluye una segunda matriz de confusión con el resultado de la clasificación siguiendo esta estrategia.

Usar matriz de covarianza. La clasificación siempre se basa en las funciones discriminantes. Pero esta clasificación puede realizarse a partir de matrices de varianzas-covarianzas distintas y el resultado de la clasificación puede ser diferente con cada estrategia.

• **Intra-grupos**. La probabilidad a posteriori de un caso en un grupo dado se calcula a partir de la matriz de varianzas-covarianzas combinada de las variables discriminantes. Por tanto, no se tiene en cuenta la distinta variabilidad de las puntuaciones discriminantes dentro de cada grupo.

• **Grupos separados.** La probabilidad a posteriori de un caso en un grupo determinado se calcula utilizando la matriz de varianzas-covarianzas de las *funciones* discriminantes en ese grupo. De esta manera se tiene en cuenta la diferente variabilidad de los grupos en las funciones discriminantes. Seleccionando esta opción, el *Visor* muestra la matriz de varianzas-covarianzas de las funciones discriminantes para cada grupo.

Gráficos. Estas opciones permiten decidir cómo serán representados los casos en las funciones discriminantes. El tipo de gráfico ofrecido depende del número de funciones estimadas:

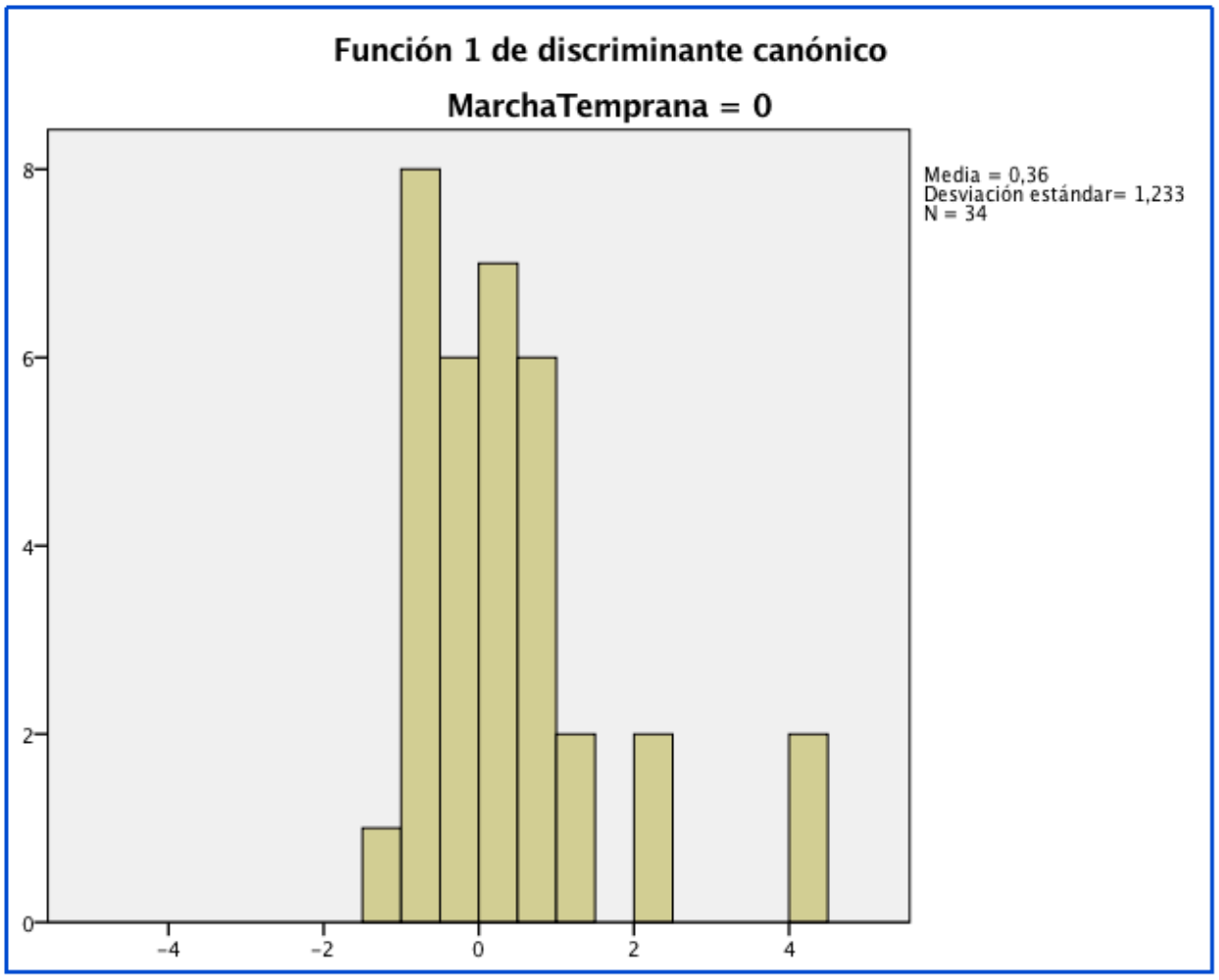
• **Grupos combinados.** Muestra un diagrama de dispersión de todos los casos en el plano definido por las dos primeras funciones discriminantes. Cuando sólo existe una función discriminante, este gráfico se omite y aparece una advertencia indicando tal circunstancia.

• **Grupos separados.** En el caso de dos grupos (una sola función discriminante), esta opción ofrece el histograma de cada grupo en la función discriminante (incluyendo los casos con valor perdido en la variable de agrupación). En el caso de más de dos grupos (más de una función discriminante), ofrece un diagrama de dispersión de cada grupo en el plano definido por las dos primeras funciones discriminantes.

• **Mapa territorial.** En el caso de más de dos grupos (más de una función discriminante), muestra la ubicación de los centroides en el plano definido por las dos primeras funciones discriminantes, así como las fronteras territoriales utilizadas en la clasificación. Las fronteras varían dependiendo de las probabilidades *a priori* seleccionadas.

• **Reemplazar los valores perdidos con la media.** Sustituye los valores perdidos de las variables independientes por sus medias aritméticas. Estas medias se calculan a partir de los casos válidos en cada variable. Los casos cuyo valor perdido es sustituido intervienen en la clasificación.

Los histogramas permiten formarse una idea aproximada tanto de la forma de la distribución como del grado de dispersión de los vehículos dentro de su propio grupo, todo ello tomando como base sus puntuaciones en la función discriminante, o lo que es lo mismo, tomando como base sus puntuaciones en el conjunto de variables independientes incluidas en el análisis.



Las leyendas de los gráficos ofrecen información descriptiva (media, desviación típica y número de casos) útil para la interpretación.

La *validación cruzada* (la clasificación de cada caso tras dejarlo fuera del cálculo de la función discriminante) arroja resultados similares a los de la clasificación original.

Resultados de clasificación[a,c]

			Pertenencia a grupos pronosticada		
		MarchaTemprana	,00	1,00	Total
Original	Recuento	,00	19	15	34
		1,00	4	27	31
	%	,00	55,9	44,1	100,0
		1,00	12,9	87,1	100,0
Validación cruzada[b]	Recuento	,00	19	15	34
		1,00	8	23	31
	%	,00	55,9	44,1	100,0
		1,00	25,8	74,2	100,0

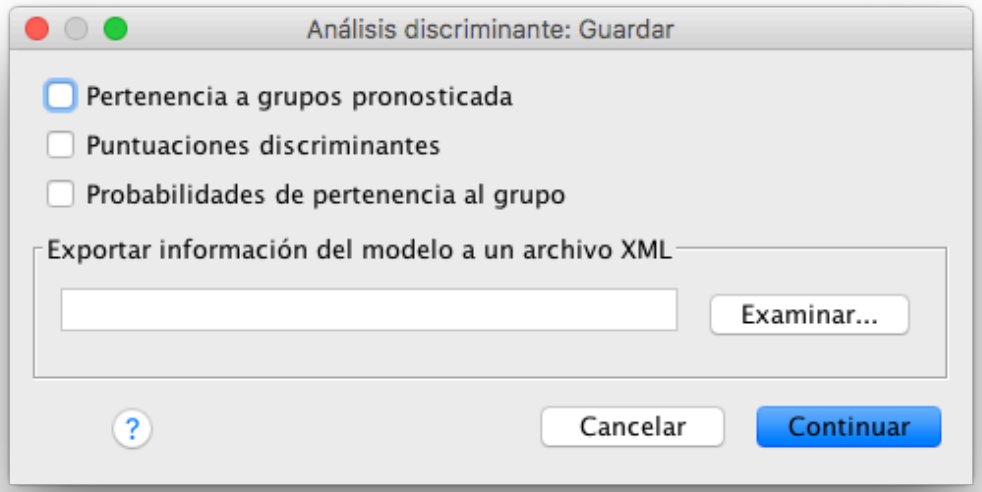
a. 70,8% de casos agrupados originales clasificados correctamente.

b. La validación cruzada se ha realizado sólo para aquellos casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas de todos los casos distintos a dicho caso.

c. 64,6% de casos agrupados validados de forma cruzada clasificados correctamente.

2. Guardar

Las opciones del cuadro de diálogo *Guardar* permiten guardar (crear) en el archivo de datos variables nuevas con información sobre algunos aspectos del análisis. Esta opción es útil para distintos fines, como por ejemplo, para utilizarla en otros procedimientos. Para crear estas nuevas variables debe seleccionarse en el cuadro de diálogo Análisis discriminante el botón **Guardar...** para acceder al subcuadro de diálogo *Análisis discriminante: Guardar.*



• **Grupo de pertenencia pronosticado.** Crea una variable categórica con códigos 1, 2, ..., que indican el grupo en el que ha sido clasificado cada caso (grupo pronosticado). El grupo pronosticado para cada caso depende de las selecciones hechas en el proceso de clasificación.

• **Puntuaciones discriminantes.** Crea tantas variables como funciones discriminantes se hayan estimado. Cada variable contiene las puntuaciones discriminantes de cada función. Las variables se crean en el orden en que se han extraído las funciones, es decir, en el orden definido por el tamaño de los autovalores. Las puntuaciones discriminantes no se ven afectadas por las selecciones realizadas en el proceso de clasificación.

• **Probabilidades de pertenencia al grupo.** Crea tantas variables como grupos se hayan incluido en el análisis. Cada variable contiene las probabilidades *a posteriori* de cada caso en un grupo. Las variables se crean en el orden definido por los códigos asignados a los grupos.

3. Validaciones cruzadas

Un problema habitual de los modelos estadísticos es que el modelo estimado siempre se ajusta lo más perfectamente posible a los datos de la muestra concreta utilizada. Esto, obviamente, constituye un pequeño inconveniente, pues la estructura de la muestra puede presentar ligeras divergencias respecto de la estructura real de la población. Para evitar este efecto de sobreajuste muestral puede llevarse a cabo una *validación cruzada*, que consiste en:

1) Seleccionar, de la muestra original, un subconjunto aleatorio de casos (*muestra de validación*);

2) Estimar la función discriminate con los casos restantes (*muestra de entrenamiento*);

3) Utilizar esa función para clasificar los casos de la *muestra de validación*.

La validación cruzada consiste, por tanto, en clasificar casos con una función que no incluye información sobre ellos. La validación cruzada puede llevarse a cabo una sola vez o repetirse varias veces. Si la muestra original es grande, podría bastar un solo intento utilizando una muestra de *validación* del 10% al 20% de los casos. Con muestras pequeñas, puede dividirse la muestra total en 10 submuestras y repetir el proceso de validación 10 veces, excluyendo cada vez una de las submuestras.

Para llevar a cabo una validación cruzada debe crearse primero una variable (la variable de *selección*) que distinga entre los casos que serán utilizados como muestra de *entrenamiento* y los que serán utilizados como muestra de *validación*. Para seleccionar los casos utilizados en el análisis debe seleccionarse en el cuadro de diálogo *Análisis discriminante* el botón **Seleccionar**>> para expandir el cuadro de diálogo.

Si se desea repetir el proceso con otra *muestra de entrenamiento*, se deberá especificar un nuevo valor para la variable de *selección.* (También es posible repetir el análisis utilizando la muestra de *entrenamiento* como muestra de *validación* y la muestra de *validación* como muestra de entrenamiento mediante el proceso *Ejecutar casos no seleccionados* que se encuentra en la carpeta de procesos del programa).

IV. El caso de más de dos grupos

Ésta técnica puede utilizarse para efectuar clasificaciones en más de dos grupos. No obstante, cuando se dispone de más de dos grupos de clasificación, la interpretación de los resultados cambia ligeramente.

Con más de dos grupos es posible obtener más de una función discriminante. En concreto, es posible obtener tantas como número de grupos menos uno (a no ser que el número de variables independientes sea menor que el número de grupos, en cuyo caso el número de posibles funciones discriminantes será igual al número de variables menos uno).

Las funciones discriminantes se extraen de manera *jerárquica*, de tal forma que la primera función explica el máximo posible de las diferencias entre los grupos, la segunda función explica el máximo de las diferencias todavía no explicadas, y así sucesivamente hasta alcanzar el 100% de las diferencias existentes. Esto se consigue haciendo que la primera función obtenga el mayor cociente entre las sumas de cuadrados inter-grupos e intra-grupos. La segunda, el siguiente mayor cociente entre ambas sumas de cuadrados. Etc.

Las funciones resultantes son independientes entre sí. En el caso de tres grupos, por ejemplo, el efecto final de esta independencia es que la primera función intenta discriminar lo mejor posible entre dos de los grupos y, la segunda, entre los dos grupos que aún se encuentren más próximos.

Podría ocurrir que la segunda función no resultase significativa, en cuyo caso habría que valorar la contribución de esa función al modelo (en términos de proporción de varianza explicada) y considerar la posibilidad de utilizar únicamente la primera función.

En este caso, además, es posible representar un mapa territorial, que representa el *territorio* (espacio) que corresponde a cada uno de los grupos en el plano definido por las dos funciones discriminantes: la primera función en el eje de abscisas y la segunda función en el eje de ordenadas.

Los centroides de cada grupo están representados por asteriscos. Para representar los centroides se utilizan las coordenadas de la tabla de centroides. Observando la ubicación de los centroides se aprecia que la primera función posee mayor capacidad discriminativa que la segunda, pues los centroides se dispersan o alejan más en la dirección horizontal que en la vertical.

Las secuencias de números que aparecen dividiendo el plano en territorios son los límites o fronteras impuestos por la regla de clasificación. Los números (1, 2, ...) identifican el grupo al que corresponde cada territorio. Conviene tener en cuenta que, puesto que la regla de clasificación cambia al cambiar las probabilidades previas, si se cambian esas probabilidades también cambiarán las fronteras de los territorios (el efecto concreto es que las fronteras se alejan del centroide del grupo al que se le asigna mayor probabilidad).

Para conocer el grupo pronosticado de un caso cualquiera (es decir, el grupo en el que será clasificado), basta con representar en el mapa territorial el punto definido por sus puntuaciones discriminantes en ambas funciones. El grupo pronosticado es aquel al que corresponde el territorio en el que queda ubicado el punto.