

# Capítulo 23

## Análisis discriminante: El procedimiento *Discriminante*

### 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 paciente se beneficiará del tratamiento, el cliente devolverá el crédito o el candidato se adaptará a su puesto de trabajo en la medida en que *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 capaz 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 decirnos 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 suponemos que se diferencian los grupos se utilizan como variables *independientes* o variables de *clasificación* (también llamadas variables *discriminantes*). Según vere-

mos, 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 aplicable a muy diversas áreas de conocimiento. Se ha utilizado para distinguir grupos de sujetos patológicos y normales a partir de los resultados obtenidos en pruebas diagnósticas, como los parámetros hemodinámicos en el ámbito clínico médico o las pruebas psicodiagnósticas en el ámbito clínico psicológico. En el campo de los recursos humanos se aplica a la selección de personal para realizar un filtrado de los *curricula* previo a la entrevista personal. En banca se ha utilizado para atribuir riesgos crediticios y en las compañías aseguradoras para predecir la siniestralidad.

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.

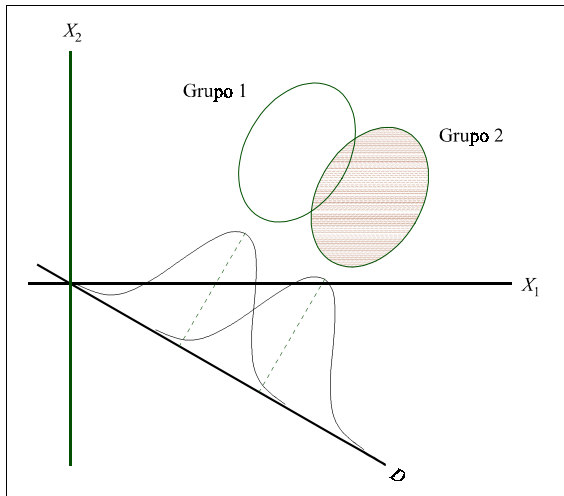
## El caso de dos grupos

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

En la figura 23.1 están representadas, en el espacio bivalente 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 figura 23.1 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 en 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).

**Figura 23.1.** Diagramas de dispersión de dos grupos en dos variables de clasificación.



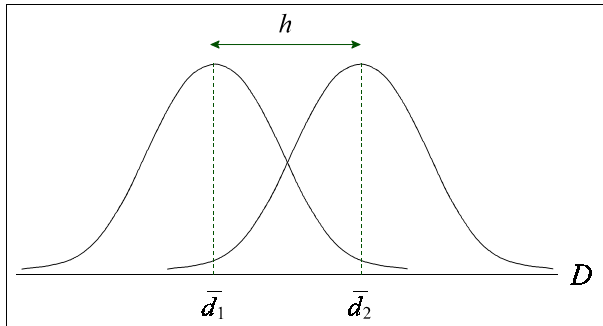
El propósito del análisis discriminante consiste en aprovechar la información contenida en las variables independientes para crear una función  $D$  combinación lineal de  $X_1$  y  $X_2$  capaz de diferenciar lo más posible a ambos grupos. La función discriminante es de la forma:

$$D = b_1 X_1 + b_2 X_2$$

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 figura 23.2 está representada sólo 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.

**Figura 23.2.** Histogramas de cada grupo y centroides representados sobre la función discriminante.



Sustituyendo en la función discriminante el valor de las medias del grupo 1 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  $\bar{d}_2$  son las medias del grupo 1 y del grupo 2 en la función  $D$ .

Como puede observarse en la figura 23.1, se desea 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 nos permitan llevar a cabo la clasificación de los sujetos.

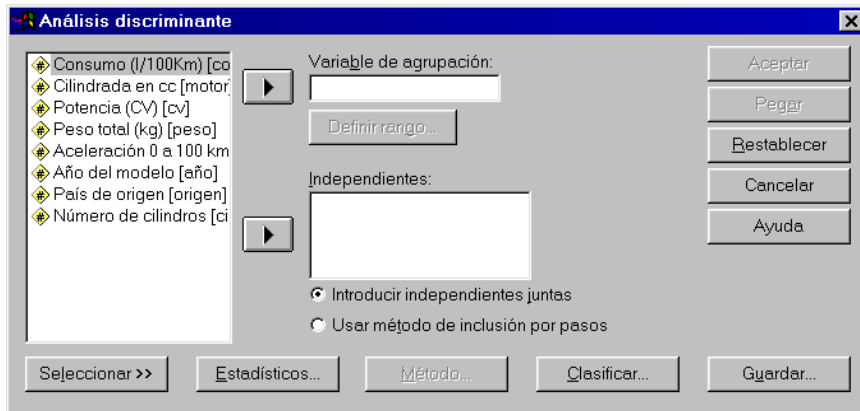
Es importante señalar que los grupos deben diferenciarse de antemano en las variables independientes. El análisis busca diferenciar los dos grupos al máximo combinando las variables independientes pero si los grupos no difieren en las variables independientes, el análisis será infructuoso: no podrá encontrar 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:

- ▣ Seleccionar la opción **Clasificar > Discriminante...** del menú **Analizar** para acceder al cuadro de diálogo *Análisis discriminante* que muestra la figura 23.3.

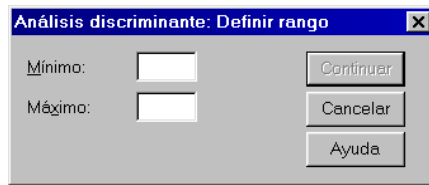
**Figura 23.3.** Cuadro de diálogo *Análisis discriminante*.



La lista de variables del archivo de datos contiene un listado con todas las variables del archivo excepto las que tienen formato de cadena. Para obtener un Análisis discriminante con las especificaciones que el programa tiene establecidas por defecto:

- ▣ Seleccionar 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.
- ▣ Seleccionar al menos una variable cuantitativa (de intervalo o razón) y trasladarla a la lista **Independientes**. Las variables independientes son aquellas en las que se desea comparar los grupos.
- ▣ Pulsar el botón **Definir rango...** para acceder al subcuadro de diálogo *Definir rango* que muestra la figura 23.4.

**Figura 23.4.** Subcuadro de diálogo *Análisis discriminante: 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). Para ello:

- ▶ Introducir el número correspondiente al límite inferior del rango en el cuadro de texto **Mínimo** y el número correspondiente al límite superior del rango en el cuadro de texto **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

### Ejemplo (Análisis discriminante)

Este ejemplo muestra cómo llevar a cabo un análisis discriminante con las especificaciones que el programa tiene establecidas por defecto. Vamos a averiguar en qué se diferencian los vehículos producidos en EE.UU. y los producidos en Europa. Para ello, utilizaremos el archivo *Coches.sav*, que se encuentra en la misma carpeta en la que ha sido instalado el SPSS. El archivo contiene información técnica (consumo, aceleración, peso, cilindrada, etc.) sobre una muestra de 406 vehículos.

Antes de iniciar el análisis hemos obtenido una representación de la dispersión de los vehículos estadounidenses y europeos en las variables *aceleración* y *peso* (figura 23.5). El archivo de datos contiene una variable llamada *origen* con tres valores: 1 = *E.UU.*, 2 = *Europa* y 3 = *Japón*. Para trabajar únicamente con los vehículos de fabricación estadounidense y europea, hay que filtrar el archivo de datos antes de obtener el diagrama de dispersión. Para ello:

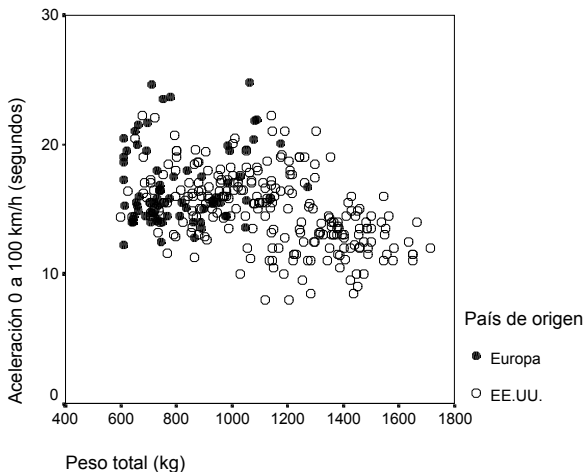
- ▶ Seleccionar la opción **Seleccionar casos...** del menú **Datos** ara acceder al cuadro de diálogo *Seleccionar casos*.
- ▶ Marcar la opción **Si se satisface la condición** y pulsar el botón **Si...** para acceder al cuadro de diálogo *Seleccionar casos: Si*.
- ▶ Establecer la condición de filtrado (por ejemplo, “origen < 3”) y pulsar el botón **Continuar**.

Aceptando estas selecciones, el archivo de datos queda filtrado dejando disponibles 306 vehículos de los 406 originales.



El diagrama de dispersión muestra que los vehículos estadounidenses tienden a situarse preferentemente en la zona de pesos altos (a la derecha), mientras que los vehículos europeos tienden a situarse más bien en la zona de pesos bajos (a la izquierda). En cuanto al eje vertical, las diferencias en aceleración parecen ser menores, si bien los vehículos con tiempos de aceleración más largos son europeos y los vehículos con tiempos de aceleración más cortos son estadounidenses.

**Figura 23.5.** Diagrama de dispersión (*peso por aceleración*) distinguiendo el país de origen.



Puesto que los casos de ambos grupos no se solapan por completo, el diagrama sugiere que existen diferencias entre ambos grupos de vehículos. Por otro lado, se aprecia cierta relación entre las variables *peso* y *aceleración*, dado que la nube de puntos adopta una forma ligeramente elipsoidal inclinada (de hecho, la correlación entre ambas variables vale  $-0,430$ ,  $p < 0,001$ ).

Si efectuamos un contraste sobre medias para comparar ambos grupos, podremos comprobar que los grupos difieren significativamente tanto en *aceleración* como en *peso*. Sin embargo, estos contrastes no tienen en cuenta la correlación existente entre las variables ni nos ayudan a clasificar los vehículos como pertenecientes a uno u otro grupo. Recordemos que el análisis discriminante no sólo permite averiguar en qué variables se diferencian los grupos sino, además, construir una función para clasificar los vehículos.

Para llevar a cabo el análisis discriminante con las especificaciones que el programa tiene establecidas por defecto:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), trasladar la variable *origen* al cuadro **Variable de agrupación** las variables *acel* (aceleración) y *peso* a la lista **Independientes**.
- ▶ Pulsar en **Definir rango...** para acceder al subcuadro de diálogo *Análisis discriminante: Definir rango* (ver figura 23.4) e introducir los valores 1 y 2 en los cuadros de texto **Mínimo** y **Máximo**, respectivamente. Pulsar el botón **Continuar**.

Aceptando las selecciones hechas, el *Visor* ofrece los resultados que muestran las tablas 23.1 a la 23.7.

La tabla 23.1 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 (en el ejemplo, 80 vehículos japoneses con el código 3 en la variable *origen*), los que son excluidos porque tienen un valor perdido en al menos una variable discriminante, y los que cumplen las dos condiciones anteriores.

**Tabla 23.1.** Tabla resumen de los casos procesados.

Casos no ponderados		N	Porcentaje
Válidos		326	80.3
Excluidos	Por pertenecer a un grupo fuera de rango	80	19.7
	Por tener valor perdido en al menos una variable discriminante	0	.0
	Por pertenecer a un grupo fuera de rango o por tener valor perdido en al menos una variable discriminante	0	.0
	Total	80	19.7
Total		406	100.0

La tabla 23.2 ofrece 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. En nuestro ejemplo, los vehículos europeos representan menos del 25% del total de vehículos analizados.

**Tabla 23.2.** Estadísticos por grupo (nº de casos válidos en cada variable).

País de origen		N válido (según lista)	
		No ponderados	Ponderados
EE.UU.	Peso total (kg)	253	253
	Aceleración 0 a 100 km/h	253	253
Europa	Peso total (kg)	73	73
	Aceleración 0 a 100 km/h	73	73
Total	Peso total (kg)	326	326
	Aceleración 0 a 100 km/h	326	326

La tabla 23.3 contiene los *autovalores* y algunos estadísticos descriptivos multivariantes. Esta tabla y la siguiente se encuentran estrechamente relacionadas y cobran mayor significado en el caso de más de dos grupos. Como veremos más adelante, 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; como en nuestro ejemplo sólo hay una función, sólo se muestra una fila. Esta única función explica el 100% de las diferencias existentes entre los sujetos de los grupos.

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 sólo. 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 co-

relación 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 autovalor obtenido en nuestro ejemplo está bastante próximo a 0 y la correlación canónica es moderada, por lo que debemos suponer que las variables discriminantes utilizadas (*peso* y *aceleración*) no permiten distinguir demasiado bien entre los dos grupos.

**Tabla 23.3.** Autovalores.

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	.294 <sup>a</sup>	100.0	100.0	.477

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

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. Wilks (1932), basándose en el principio de razón de verosimilitud generalizada (según el cual la varianza generalizada de un espacio multivariante puede ser calculada mediante el determinante de la matriz de dispersión), planteó el estadístico  $\Lambda$ , definido como:

$$\Lambda = \frac{\text{Suma de cuadrados intragrupos}}{\text{Suma de cuadrados total}} = \frac{|\mathbf{S}|}{|\mathbf{T}|}$$

donde  $\mathbf{S}$  es la matriz de varianzas-covarianzas *combinada*, calculada a partir de las matrices de varianzas-covarianzas de cada grupo, y  $\mathbf{T}$  es la matriz de varianzas-covarianzas *total*, calculada sobre todos los casos como si pertenecieran a un único grupo. Cuando los grupos se encuentren superpuestos en el espacio multidimensional, los valores del numerador y del denominador serán aproximadamente iguales y su cociente valdrá 1; a medida que los grupos se vayan separando más y más, la variabilidad *inter-grupos* irá aumentando y la variabilidad *intra-grupos* se irá haciendo comparativamente menor respecto a la variabilidad *total*, disminuyendo así el valor del cociente. 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  $\text{lambda} + \text{eta}^2 = 1$ .

**Tabla 23.4.** Lambda de Wilks.

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.773	83.202	2	.000

Aunque Schatzoff (1966) obtuvo los puntos críticos exactos de la distribución de  $\Lambda$  bajo ciertas condiciones, es más frecuente utilizar una transformación de  $\Lambda$  que posee distribución aproximada conocida. Bartlett (1947) ha demostrado que el estadístico:

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

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

funciones discriminantes obtenidas con anterioridad al contraste (cuando sólo existe una función —porque sólo hay dos grupos—,  $k = 0$ ).

La gran ventaja diagnóstica del estadístico *lambda* es que, puesto que se basa en las matrices de varianzas-covarianzas, puede calcularse antes de obtener las funciones discriminantes.

En nuestro ejemplo, el valor de *lambda* es moderadamente alto (0,773), lo cual significa que existe bastante solapamiento entre los grupos. Sin embargo, el valor transformado de *lambda* (*Chi-cuadrado* = 83,202) tiene asociado, con 2 grados de libertad, un nivel crítico (*Sig.*) de 0,000, por lo que podemos rechazar la hipótesis nula de que los grupos comparados tienen promedios iguales en las dos variables discriminantes.

La tabla de *coeficientes estandarizados* (tabla 23.5) 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. Atendiendo al valor de los coeficientes estandarizados de la tabla 23.5 podemos concluir que la variable *peso* tiene mayor importancia que la variable *aceleración* a la hora de predecir el grupo de pertenencia de los vehículos.

**Tabla 23.5.** Coeficientes estandarizados de las funciones discriminantes canónicas.

	Función
	1
Peso total (kg)	.919
Aceleración 0 a 100 km/h	-.184

Para interpretar los signos de las ponderaciones resulta útil inspeccionar primero la ubicación de los centroides de cada grupo. Los centroides se muestran en la tabla 23.7. Podemos comprobar que el grupo de coches estadounidenses tiende a obtener puntuaciones positivas en la función discriminante, mientras que el grupo de vehículos europeos tiende a obtener puntuaciones negativas. Sabido esto, la función discriminante nos indica que un incremento en el *peso* (por encima de la media) hará más probable que el vehículo obtenga una puntuación positiva y, con ello, que se ajuste al patrón de los vehículos estadounidenses. Por el contrario, un *peso* por debajo de la media será característico de un vehículo europeo. En cuanto a la variable *aceleración*, un valor por encima de la media (mayor número de segundos en alcanzar los 100 km/h) hará disminuir la puntuación discriminante (dado que el signo es negativo) y será más característico de los vehículos europeos, y viceversa, una puntuación en *aceleración* por debajo de la media aumentará las posibilidades de que el vehículo sea clasificado como estadounidense. Basándonos en estos resultados, podemos afirmar que los vehículos estadounidenses tienen mayor peso y tardan menos en alcanzar los 100 km/h.

La *matriz de estructura* (tabla 23.6) contiene las correlaciones entre las variables discriminantes y la función discriminante estandarizada. Mientras que los coeficientes estandarizados muestran la contribución *net*a 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.

Cuando existe colinealidad entre las variables independientes puede ocurrir que alguna de ellas quede fuera del análisis por no aportar información nueva. Sin embargo, no por ello carece de interés conocer cómo se relaciona cada variable independiente con la función discriminante. Conocer estas relaciones puede ayudar a interpretar mejor la función discriminante.

En la tabla 23.6 podemos apreciar que la *aceleración* correlaciona alto con la función discriminante, aunque sea una variable poco importante en la función. Posiblemente, la poca importancia de esta variable en la función se deba a su relación con la variable *peso*, la cual ha capitalizado la información que comparte con la *aceleración* y la aporta de manera individual a 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.

**Tabla 23.6.** Matriz de estructura.

	Función
	1
Peso total (kg)	.985
Aceleración 0 a 100 km/h	-.513



La tabla 23.7 contiene la ubicación de los centroides en la función discriminante (bruta) tal y como se muestran en la figura 23.2. Esta tabla es de gran utilidad para interpretar la función discriminante. Podemos observar que el grupo de vehículos estadounidenses se encuentra localizado, en promedio, en las puntuaciones positivas de la función, mientras que los vehículos europeos se encuentran ubicados en las puntuaciones negativas.

Si desconocemos la procedencia de un vehículo pero tenemos información sobre su *peso* y *aceleración*, podemos calcular su *puntuación discriminante* y, a partir de ella, asignarlo al grupo de cuyo centroide se encuentre más próximo.

**Tabla 23.7.** Valores de los centroides en la función discriminante.

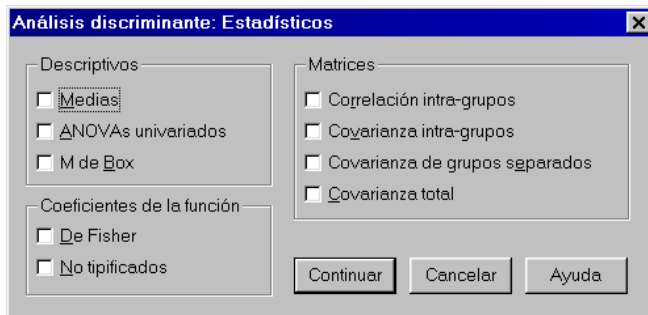
	Función
País de origen	1
EE.UU.	.290
Europa	-1.006

## 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:

- ▶ Pulsar en el botón **Estadísticos...** (ver figura 23.3) para acceder al subcuadro de diálogo *Análisis discriminante: Estadísticos* que se muestra en la figura 23.6.

**Figura 23.6.** Subcuadro de diálogo *Análisis discriminante: Estadísticos*.



**Descriptivos.** Este apartado contiene opciones que permiten obtener información descriptiva y contrastes 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 (tabla 23.8).

**Tabla 23.8.** Estadísticos descriptivos.

País de origen		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
EE.UU.	Peso total (kg)	1122.11	262.87	253	253.000
	Aceleración 0 a 100 km/h	14.93	2.80	253	253.000
Europa	Peso total (kg)	810.12	163.62	73	73.000
	Aceleración 0 a 100 km/h	16.82	3.01	73	73.000
Total	Peso total (kg)	1052.25	276.55	326	326.000
	Aceleración 0 a 100 km/h	15.35	2.95	326	326.000

- ☐ **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.

**Tabla 23.9.** Pruebas de igualdad de las medias de los grupos.

	Lambda de Wilks	F	gl1	gl2	Sig.
Peso total (kg)	.778	92.381	1	324	.000
Aceleración 0 a 100 km/h	.928	25.022	1	324	.000

- **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  $M$  de Box toma la forma:

$$M = (n - g) \log |\mathbf{S}| - \sum_{j=1}^g (n_j - 1) \log |\mathbf{S}^{(j)}|$$

donde  $\mathbf{S}$  es la matriz de varianzas-covarianzas *combinada*,  $\mathbf{S}^{(j)}$  es la matriz de varianzas-covarianzas del  $j$ -ésimo grupo,  $n$  es el número total de casos,  $n_j$  es el número de casos en el  $j$ -ésimo grupo y  $g$  es el número de grupos. El estadístico  $M$  carece de distribución muestral conocida, pero puede transformarse en un estadístico  $F$  e interpretarse como tal (muchos analistas critican el uso de este estadístico por ser demasiado sensible a pequeñas desviaciones de la normalidad multivariante y a tamaños muestrales grandes, tendiendo a ser conservador).

La tabla 23.10 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.

**Tabla 23.10.** Logaritmos de los determinantes.

País de origen	Rango	Logaritmo del determinante
EE.UU.	2	12.963
Europa	2	12.379
Intra-grupos combinada	2	12.954

La tabla 23.11 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 ( $\text{Sig.} = 0,000 < 0,05$ ) y, por tanto, concluir que uno de los dos grupos es más variable que el otro.

**Tabla 23.11.** Tabla de resultados de la prueba  $M$  de Box.

M de Box		39.135
F	Aprox.	12.906
	gl1	3
	gl2	263888.2
	Sig.	.000

**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 (ver tabla 23.12). Aparece en la misma tabla que la matriz de varianzas-covarianzas intra-grupos combinada.
- ☐ **Covarianza intra-grupos.** Matriz de varianzas-covarianzas intra-grupo *combinada* (ver tabla 23.12). 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.

**Tabla 23.12.** Matrices intra-grupo combinadas.

		Peso total (kg)	Aceleración 0 a 100 km/h
Covarianza	Peso total (kg)	59693.536	-248.818
	Aceleración 0 a 100 km/h	-248.818	8.117
Correlación	Peso total (kg)	1.000	-.357
	Aceleración 0 a 100 km/h	-.357	1.000

- ☐ **Covarianza de grupos separados.** Matrices de varianzas-covarianzas de cada grupo (ver tabla 23.13). 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.

- **Covarianza total.** Matriz de varianzas-covarianzas *total*, es decir, calculada sobre todos los sujetos de la muestra como si pertenecieran a un único grupo (ver tabla 23.13). Aparece en la última submatriz de la tabla 23.13, con el encabezado *Total*. Es la matriz **T** utilizada en el cálculo de la *lambda* de Wilks.

**Tabla 23.13.** Matrices de varianzas-covarianzas.

País de origen		Peso total (kg)	Aceleración 0 a 100 km/h
EE.UU.	Peso total (kg)	69099.515	-340.100
	Aceleración 0 a 100 km/h	-340.100	7.846
Europa	Peso total (kg)	26772.610	70.669
	Aceleración 0 a 100 km/h	70.669	9.066
Total	Peso total (kg)	76477.740	-351.030
	Aceleración 0 a 100 km/h	-351.030	8.717

**Coefficientes de la función.** Este apartado contiene opciones que permiten seleccionar algunos coeficientes adicionales utilizados en la clasificación de los casos.

- **Coefficientes 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 de la tabla 23.7. No es habitual solicitar esta tabla por dos motivos. En primer lugar, el programa calcula de manera automática las puntuaciones discriminantes. En segundo lugar, este conjunto de coeficientes depende de la variabilidad y la métrica de las variables (de manera similar a lo que sucede con los coeficientes de regresión no tipificados del análisis de regresión múltiple), lo que dificulta su interpretación. La función discriminante incluye una constante correctora que consigue que las puntuaciones discriminantes tomen el valor 0 en algún punto entre los dos centroides.

**Tabla 23.14.** Coeficientes de la función discriminante (no tipificados).

	Función
	1
Peso total (kg)	.004
Aceleración 0 a 100 km/h	-.065
(Constante)	-2.967

Coeficientes no tipificados

A modo de ejemplo, puede comprobarse que a partir de las medias de cada grupo en las variables discriminantes (ver tabla 23.8) y este conjunto de coeficientes se obtienen los centroides en la función discriminante (ver tabla 23.7):

$$\bar{d}_1 = b_0 + b_1 \bar{x}_1^{(1)} + b_2 \bar{x}_2^{(1)} = -2,967 + 0,004 \times 1.122,11 - 0,65 \times 14,93 = 0,290$$

$$\bar{d}_2 = b_0 + b_1 \bar{x}_1^{(2)} + b_2 \bar{x}_2^{(2)} = -2,967 + 0,004 \times 810,12 - 0,65 \times 16,82 = -1,006$$

- **Coefficientes 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.

**Tabla 23.15.** Funciones de clasificación de Fisher.

	País de origen	
	EE.UU.	Europa
Peso total (kg)	.030	.025
Aceleración 0 a 100 km/h	2.769	2.853
(Constante)	-38.385	-35.003

Funciones discriminantes lineales de Fisher



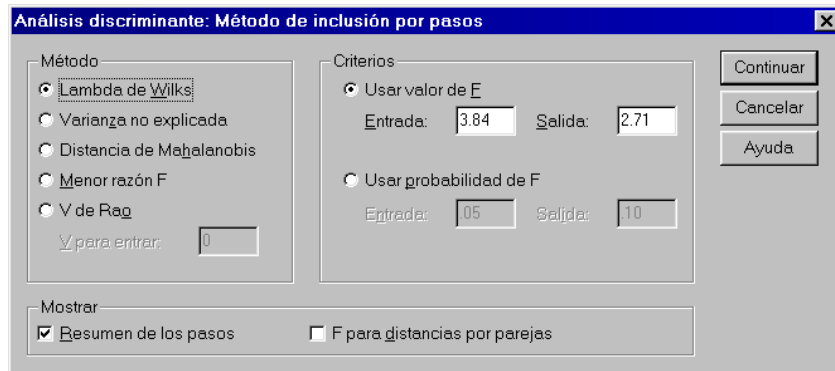
## Método

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:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), seleccionar la opción **Usar método de inclusión por pasos**.
- ▶ Pulsar en el botón **Método...** (inactivo hasta que se marca la opción **Usar método de inclusión por pasos**) para acceder al subcuadro de diálogo *Análisis discriminante: Método de inclusión por pasos* que muestra en la figura 23.7.

**Figura 23.7.** Subcuadro de diálogo *Análisis discriminante: Método de inclusión por pasos*.



**Método.** 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  $F_{\text{cambio}}$  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  $F_{\text{cambio}}$  para cada variable, se incorpora al modelo la variable a la que le corresponde el mayor valor  $F_{\text{cambio}}$  (o, lo que es lo mismo, la que produce el mayor cambio en la *lambda* de Wilks):

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

donde  $n$  es el número de casos válidos,  $g$  es el número de grupos,  $\lambda_p$  es la *lambda* de Wilks que corresponde al modelo antes de incluir la variable que se está evaluando y  $\lambda_{p+1}$  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 explicada por el modelo,  $R^2$ , es proporcional, en una constante  $c$ , a la distancia  $H$  de Mahalanobis (ver más abajo):

$$R^2 = c H_{ab}^2$$

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:

$$H_{ab}^2 = (n - g) \sum_{i=1}^p \sum_{j=1}^p w_{ij}^* (\bar{X}_i^{(a)} - \bar{X}_i^{(b)}) (\bar{X}_j^{(a)} - \bar{X}_j^{(b)})$$

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:

$$F = \frac{(n - p - 1) n_1 n_2}{p(n - 2)(n_1 + n_2)} H_{ab}^2$$

- **$V$  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  $V$ :

$$V = (n_k - g) \sum_{i=1}^p \sum_{j=1}^p w_{ij}^* \sum_{k=1}^g (\bar{X}_i^{(k)} - \bar{X}_i) (\bar{X}_j^{(k)} - \bar{X}_j)$$

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  $k$ ,  $\bar{X}_i^{(k)}$  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**.

**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*:
  - ▣ 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*:

---

\* Superado el criterio de *significación*, una variable sólo pasa a formar parte del modelo si su *nivel de tolerancia* es mayor que el nivel establecido por defecto (este nivel es 0,001, pero puede cambiarse mediante sintaxis) y si, además, su incorporación al modelo no hace que alguna de las variables previamente seleccionadas pase a tener un nivel de tolerancia por debajo del nivel establecido por defecto. La tolerancia de una variable independiente es la proporción de varianza de esa variable que no está asociada (que no depende) del resto de variables independientes incluidas en la ecuación. Una variable con una tolerancia de, por ejemplo, 0,01 es una variable que comparte el 99 % de su varianza con el resto de variables independientes, lo cual significa que se trata de una variable redundante casi por completo.

- ☒ Seleccionar 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.

### Ejemplo (Análisis discriminante > Método)

Este ejemplo muestra cómo utilizar la estrategia de inclusión de variables por pasos y cómo interpretar los resultados obtenidos.

Aunque cada uno de los métodos disponibles puede dar lugar a una función discriminante distinta y los estadísticos que aparecen en las tablas de resultados dependen del método seleccionado, creemos que basta con estudiar uno cualquiera de los métodos disponibles para comprender cómo funciona la estrategia de inclusión por pasos. Para construir un modelo por pasos:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), trasladar las variables *consumo*, *motor* (cilindrada), *peso*, *acel* (aceleración), *cv* (potencia), *año* y *cilindr* (número de cilindros) a la lista **Independientes** y la variable *origen* al cuadro **Variable de agrupación**.
- ▶ Pulsar el botón **Definir rango...** para acceder al subcuadro de diálogo *Definir rango* y escribir el valor 1 (el código de EE.UU. en la variable *origen*) en la casilla **Mínimo** y el valor 2 (el código de Europa en la variable *origen*) en la casilla **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Seleccionar la opción **Usar método de inclusión por pasos** y pulsar el botón **Método...** para acceder al subcuadro de diálogo *Análisis discriminante: Método de inclusión por pasos* (ver figura 23.7).
- ▶ Seleccionar la opción **F para distancias por parejas** del apartado **Mostrar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las tablas 23.16 a la 23.27.

La tabla 23.16 indica que disponemos de 312 casos válidos. Se han excluido 94 casos de la muestra original de 406 casos. De estos 94 casos, 79 no pertenecen a ninguno de los grupos seleccionados (bien por que en la variable de agrupación, *origen*, tienen el código del país no seleccionado –Japón–, bien porque tienen valor perdido en esa variable); 14 casos pertenecen a uno de los dos grupos seleccionados (EE.UU. o Europa; códigos 1 y 2 en la variable *origen*) pero no disponen de información completa en todas las variables independientes; y en 1 caso se dan ambas circunstancias.

Los casos excluidos por tener algún valor perdido en las variables independientes no se utilizan para calcular la función discriminante, pero pueden ser utilizados más tarde en la fase de clasificación (ver más adelante).

En ocasiones puede resultar interesante realizar un análisis pormenorizado de los valores perdidos (por ejemplo, con el módulo *Valores perdidos*). Si la muestra contiene muchos casos con valor perdido en al menos una variable independiente, la función discriminante se construirá a partir de un número muy reducido de casos. Esto representa un serio inconveniente por dos razones. Por un lado, las estimaciones basadas en pocos casos suelen ser poco eficientes (muy variables y, por tanto, inestables: cambian mucho al utilizar muestras distintas). Por otro, si los casos con valores perdidos difieren de los casos válidos en alguna característica concreta, las estimaciones estarán sesgadas. Si, por ejemplo, los fabricantes de vehículos de más cilindrada tienen por costumbre no informar de los consumos de sus vehículos, podemos encontrarnos con que los casos de los que se dispone información son en su mayoría vehículos de bajo consumo. Los resultados obtenidos prescindiendo del consumo de los vehículos de gran cilindrada podrían ser, obviamente, muy distintos de los obtenidos si se contara con información sobre el consumo de todos los vehículos.

Siempre es, por tanto, conveniente detenerse a revisar los valores perdidos para averiguar si poseen alguna característica común. Podría ocurrir, por ejemplo, que la mayor parte de los valores perdidos se concentraran en una o dos variables; excluir esa o esas variables del análisis permitiría aumentar el número de casos válidos y, con ello, obtener estimaciones menos sesgadas y más eficientes.

**Tabla 23.16.** Resumen de los casos procesados.

Casos no ponderados		N	Porcentaje
Válidos		312	76.8
Excluidos	Por pertenecer a un grupo fuera de rango	79	19.5
	Por tener valor perdido en al menos una variable discriminante	14	3.4
	Por pertenecer a un grupo fuera de rango o por tener valor perdido en al menos una variable discriminante	1	.2
	Total	94	23.2
Total		406	100.0

La tabla 23.17 informa del número de casos válidos en cada grupo y en cada una de las variables independientes. Puesto que la exclusión de casos se realiza según *lista* (es decir, se excluyen del análisis los casos con valor perdido en al menos una variable independiente), el número de casos válido de todas las variables es el mismo en cada grupo.

**Tabla 23.17.** Estadísticos por grupo (nº de casos válidos en cada variable).

País de origen		N válido (según lista)	
		No ponderados	Ponderados
EE.UU.	Consumo (l/100Km)	244	244.000
	Cilindrada en cc	244	244.000
	Potencia (CV)	244	244.000
	Peso total (kg)	244	244.000
	Aceleración 0 a 100 km/h	244	244.000
	Año del modelo	244	244.000
	Número de cilindros	244	244.000
Europa	Consumo (l/100Km)	68	68.000
	Cilindrada en cc	68	68.000
	Potencia (CV)	68	68.000
	Peso total (kg)	68	68.000
	Aceleración 0 a 100 km/h	68	68.000
	Año del modelo	68	68.000
	Número de cilindros	68	68.000
Total	Consumo (l/100Km)	312	312.000
	Cilindrada en cc	312	312.000
	Potencia (CV)	312	312.000
	Peso total (kg)	312	312.000
	Aceleración 0 a 100 km/h	312	312.000
	Año del modelo	312	312.000
	Número de cilindros	312	312.000



La tabla de *variables introducidas/eliminadas* (tabla 23.18) muestra un resumen de todos los pasos llevados a cabo en la construcción de la función discriminante y recuerda los criterios utilizados en la selección de variables. En cada paso se informa de la variable que ha sido incorporada al modelo y, en su caso, de la variable o variables que han sido expulsadas. En nuestro ejemplo, todos los pasos llevados a cabo han sido de incorporación de variables: en el primer paso, *cilindrada*; en el segundo, *potencia*; etc. Así, hasta un total de 5 variables. En ninguno de los 5 pasos ha habido expulsión de variables. Si alguna de las variables previamente incorporadas hubiera sido expulsada en algún paso posterior, la tabla mostraría una columna adicional indicando tal circunstancia.

Las notas a pie de tabla recuerdan algunas de las opciones establecidas para el análisis: la selección de variables se ha llevado a cabo utilizando el estadístico *lambda* de Wilks global, el número máximo de pasos permitidos es 14 (valor que no se ha alcanzado puesto que sólo se han realizado 6 pasos), el valor del estadístico *F* para incorporar variables es 3,84 (criterio de *entrada*), el valor del estadístico *F* para excluir variables es 2,71 (criterio de *salida*) y, por último, en la nota *d* se informa de que se ha alcanzado alguno de los criterios de *parada* (los niveles del estadístico *F*, el criterio de tolerancia y la *V* mínima de Rao), por lo que alguna de las variables independientes inicialmente propuestas no ha sido incluida en el modelo final.

Puede observarse que el valor del estadístico *lambda* de Wilks va disminuyendo en cada paso, lo cual es síntoma de que, conforme se van incorporando variables al modelo, los grupos van estando cada vez menos solapados. En la columna *F exacta* se encuentra el valor transformado de la *lambda* de Wilks y su significación. Los valores del estadístico se refieren al estadístico *global* y no al *cambio* en el estadístico. Esta tabla siempre recoge el estadístico seleccionado en la opción **Método**, por lo que la cabecera de la columna cambiará de un análisis a otro dependiendo del estadístico elegido.

**Tabla 23.18.** Variables introducidas/eliminadas (resumen del análisis por pasos).

Paso	Introducidas	Lambda de Wilks <sup>a,b,c,d</sup>							
		Estadístico	ql1	ql2	ql3	F exacta			
						Estadístico	ql1	ql2	Sig.
1	Cilindrada en cc	.704	1	1	310.000	130.522	1	310.000	.000
2	Potencia (CV)	.660	2	1	310.000	79.582	2	309.000	.000
3	Año del modelo	.631	3	1	310.000	60.009	3	308.000	.000
4	Peso total (kg)	.620	4	1	310.000	47.075	4	307.000	.000
5	Consumo (l/100Km)	.606	5	1	310.000	39.744	5	306.000	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

- El número máximo de pasos es 14.
- La *F* parcial mínima para entrar es 3.84.
- La *F* parcial máxima para salir es 2.71.
- El nivel de *F*, la tolerancia o el VIN son insuficientes para continuar los cálculos.

La tabla 23.19 se encuentra dividida por cada uno de los pasos. En cada paso se mencionan las variables incorporadas al modelo hasta ese momento y, para cada variable, el nivel de tolerancia, el valor del estadístico  $F$  que permite valorar si la variable debe o no ser expulsada, ( $F$  para ser eliminada) y la  $\lambda$  de Wilks global que obtendríamos si se eliminara la variable del modelo.

Esta tabla permite valorar (mediante  $F$  y  $\lambda$ ) el efecto de la exclusión de cada variable y (mediante el nivel de tolerancia) el grado de colinealidad existente entre las variables independientes. Puesto que las variables utilizadas en nuestro ejemplo se encuentran muy relacionadas entre sí, la tolerancia disminuye sensiblemente en el momento en que se incorpora una nueva variable al modelo (recordemos que la tolerancia es la proporción de varianza de una variable independiente que no está explicada por el resto de variables independientes). En el paso 0 todas las variables tienen una tolerancia igual a 1, pues todavía no existen variables en el modelo. En el paso 1 permanece en ese valor de tolerancia para la primera variable pues, al estar sola, no existen variables que puedan explicar nada de ella (véase la tolerancia de la variable *cilindrada* en el paso 1). En el segundo paso, al incorporarse la variable *potencia* al modelo, la tolerancia baja a 0,212, lo cual es síntoma de que existe una alta correlación entre ambas variables (es fácil deducir que la correlación entre las dos variables es de 0,89). Sin embargo, la variable *año del modelo* no correlaciona tanto con la *potencia* y *cilindrada*: al incorporarse al modelo en el tercer paso, su tolerancia sólo baja hasta 0,799.

**Tabla 23.19.** Variables incluidas en el análisis (variables seleccionadas en cada paso).

Paso		Tolerancia	F para eliminar	Lambda de Wilks
1	Cilindrada en cc	1.000	130.522	
2	Cilindrada en cc	.212	84.720	.841
	Potencia (CV)	.212	20.452	.704
3	Cilindrada en cc	.210	88.464	.812
	Potencia (CV)	.208	14.658	.661
	Año del modelo	.799	14.110	.660
4	Cilindrada en cc	.114	71.727	.765
	Potencia (CV)	.205	12.097	.644
	Año del modelo	.767	17.124	.654
	Peso total (kg)	.179	5.591	.631
5	Cilindrada en cc	.114	71.054	.747
	Potencia (CV)	.194	15.813	.638
	Año del modelo	.509	24.242	.654
	Peso total (kg)	.127	11.612	.629
	Consumo (l/100Km)	.157	6.839	.620

La tabla 23.20 ofrece una evaluación de las variables candidatas a ser incluidas en el modelo en cada uno de los pasos. La tabla muestra, en cada paso, las variables que todavía no han sido incorporadas al modelo.

**Tabla 23.20.** Variables no incluidas en el análisis (variables no seleccionadas en cada paso)

Paso		Tolerancia	Tolerancia mínima	F para introducir	Lambda de Wilks
0	Consumo (l/100Km)	1.000	1.000	66.617	.823
	Cilindrada en cc	1.000	1.000	130.522	.704
	Potencia (CV)	1.000	1.000	58.615	.841
	Peso total (kg)	1.000	1.000	85.178	.784
	Aceleración 0 a 100 km/h	1.000	1.000	22.220	.933
	Año del modelo	1.000	1.000	.016	1.000
	Número de cilindros	1.000	1.000	107.273	.743
1	Consumo (l/100Km)	.302	.302	4.448	.694
	Potencia (CV)	.212	.212	20.452	.660
	Peso total (kg)	.188	.188	4.257	.694
	Aceleración 0 a 100 km/h	.717	.717	1.809	.700
	Año del modelo	.814	.814	19.892	.661
	Número de cilindros	.138	.138	.308	.703
2	Consumo (l/100Km)	.277	.174	.672	.659
	Peso total (kg)	.186	.119	2.648	.654
	Aceleración 0 a 100 km/h	.506	.149	1.637	.657
	Año del modelo	.799	.208	14.110	.631
	Número de cilindros	.137	.083	1.033	.658
3	Consumo (l/100Km)	.221	.172	.893	.629
	Peso total (kg)	.179	.114	5.591	.620
	Aceleración 0 a 100 km/h	.502	.149	2.502	.626
	Número de cilindros	.137	.082	1.120	.629
4	Consumo (l/100Km)	.157	.114	6.839	.606
	Aceleración 0 a 100 km/h	.382	.110	.235	.619
	Número de cilindros	.135	.067	.564	.619
5	Aceleración 0 a 100 km/h	.380	.107	.460	.605
	Número de cilindros	.132	.066	1.285	.604

Antes de iniciar la construcción del modelo (paso 0) la tolerancia de todas las variables es la máxima posible y, puesto que las variables están siendo evaluadas individualmente, la  $F$  para entrar en el modelo ( $F$  para introducir) coincide con el valor de la  $F$  univariante que se obtendría al marcar la opción **ANOVAs univariantes** del cuadro de diálogo *Análisis discriminante: Estadísticos* (ver figura 23.6). Además, para cada variable ya incorporada al modelo, el valor de la  $F$  para salir en un determinado paso ( $F$  para eliminar de la tabla 23.19) coincide con el valor de la  $F$  para entrar en el paso anterior ( $F$  para introducir de la tabla 23.20).

En cuanto a la tolerancia de las variables, la tabla incluye dos columnas: la primera (*Tolerancia*) ofrece, en cada paso, la tolerancia que tendría cada variable si fuera incorporada al modelo en el siguiente paso; la segunda columna (*Tolerancia mínima*) ofrece la tolerancia correspondiente a la variable (de las ya incluidas en el modelo) cuya tolerancia más se vería afectada por la incorporación de la nueva variable (es decir, la tolerancia de la variable cuya tolerancia pasaría a ser la más pequeña de todas). En el paso 3, por ejemplo, la variable *peso* es la mejor candidata para ser incorporada al modelo en el siguiente paso (es la que haría menor el valor de la *lambda* de Wilks). Al ser incorporada en el paso 4 (ver tabla 23.19), su tolerancia dentro de la ecuación es el valor informado antes de entrar (en la tabla 23.20). Y la variable a la que más afecta su inclusión es a la *cilindrada*: en el paso 3 (tabla 23.20) se indica que la tolerancia mínima pasará a ser 0,084, y en el paso 4 (tabla 23.19) se confirma ese valor.

La tabla 23.21 siempre muestra el estadístico *lambda* de Wilks global para el modelo generado en cada paso, independientemente de que se haya optado por otro estadístico como método de selección de variables. Según sabemos ya, este estadístico permite valorar el grado de diferenciación entre los grupos tomando como referencia las variables independientes incluidas en cada paso. En este caso, la información de la tabla 23.21 coincide exactamente con la de la tabla 23.18.

**Tabla 23.21.** Lambda de Wilks.

Paso	Número de variables	Lambda	gl1	gl2	gl3	F exacta			
						Estadístico	gl1	gl2	Sig.
1	1	.704	1	1	310	130.522	1	310.000	6.569E-21
2	2	.660	2	1	310	79.582	2	309.000	.000
3	3	.631	3	1	310	60.009	3	308.000	.000
4	4	.620	4	1	310	47.075	4	307.000	.000
5	5	.606	5	1	310	39.744	5	306.000	.000

La tabla 23.22 ofrece estadísticos  $F$  que permiten contrastar la hipótesis de igualdad de medias entre cada dos grupos. Esta tabla tiene mayor sentido cuando el análisis busca discriminar entre más de dos grupos, pues permite averiguar qué grupos difieren de qué otros (recordemos que la  $\lambda$  de Wilks hace una valoración global del grado de diferenciación entre los grupos). Puesto que en nuestro ejemplo estamos utilizando sólo dos grupos, los valores de esta tabla coinciden con los de la tabla 23.21.

**Tabla 23.22.** Estadístico  $F$  para la comparación de los grupos por pares.

Origen	Paso		Europa
EE.UU.	1	F	130.522
		Sig.	.000
	2	F	79.582
		Sig.	.000
	3	F	60.009
		Sig.	.000
	4	F	47.075
		Sig.	.000
	5	F	39.744
		Sig.	.000

En la tabla 23.23 podemos apreciar que el *autovalor* ha aumentado respecto al caso de dos variables (ver tabla 23.3). También ha aumentado considerablemente la correlación canónica. Técnicamente, el autovalor es proporcional a la dispersión obtenida en la dirección del mayor autovector de la nube de puntos multivariante. Si el autovalor aumenta es porque la nube de puntos multivariante aumenta su dispersión y es posible distinguir mejor los grupos.

**Tabla 23.23.** Autovalores.

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	.649	100.0	100.0	.627

La tabla 23.24 muestra el valor de la  $\lambda$  de Wilks para el modelo final. Su significación se evalúa mediante una transformación *chi-cuadrado* (ver tabla 23.4).

**Tabla 23.24.** Lambda de Wilks global.

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.606	153.879	5	.000

La tabla 23.25 ofrece la matriz de *coeficientes estandarizados*. Estos coeficientes permiten valorar la contribución *net*a de cada variable a la función discriminante. Así, la variable que más contribuye a diferenciar los grupos es la *cilindrada* (un análisis factorial de las variables independientes nos diría que las variables correlacionan mucho entre sí, por lo que no es de extrañar que la variable *cilindrada* pueda estar capitalizando las relación entre de la función discriminante y el resto de variables independientes).

Así pues, a mayor *cilindrada*, mayor puntuación en la función discriminante y, en consecuencia, mayor tendencia a que el vehículo sea clasificado como estadounidense (ver la tabla de los centroides 23.27). La variable *potencia*, sin embargo, presenta un coeficiente negativo. Esto quiere decir que para vehículos con iguales puntuaciones en las restantes variables, los que tienen mayor *potencia* tendrán una puntuación menor en la función discriminante y, consecuentemente, será más fácil que sean clasificados como vehículos europeos. De manera similar podemos interpretar que, manteniendo constantes el resto de variables, los vehículos estadounidenses presentan, comparativamente, un mayor consumo y menor peso y antigüedad.

**Tabla 23.25.** Coeficientes estandarizados de las funciones discriminantes canónicas.

	Función
	1
Consumo (l/100Km)	.595
Cilindrada en cc	2.052
Potencia (CV)	-.801
Peso total (kg)	-.856
Año del modelo	.605

En la *matriz de estructura* (tabla 23.26) se encuentran los coeficientes de correlación *brutos* entre cada variable y la función discriminante. Puede observarse que el signo de los coeficientes correspondientes a las variables *potencia* y *peso* han cambiado (respecto al que tenían en la matriz de coeficientes estandarizados). Este cambio de signo es consecuencia del alto grado de colinealidad entre las variables. En nuestro ejemplo, el valor de los coeficientes de correlación indican que la función discriminante distingue básicamente entre vehículos grandes (los estadounidenses) y vehículos pequeños (los europeos). Aunque no han sido utilizadas para construir la función discriminante, la tabla también informa de las correlaciones existentes entre la función discriminante y las variables *aceleración* y *número de cilindros*.

**Tabla 23.26.** Matriz de estructura.

	Función
	1
Cilindrada en cc	.805
Número de cilindros <sup>a</sup>	.768
Peso total (kg)	.650
Consumo (l/100Km)	.575
Potencia (CV)	.540
Aceleración 0 a 100 km/h <sup>a</sup>	-.294
Año del modelo	-.009

a. Variable no utilizada en el análisis.

La tabla de centroides (23.27) muestra que los vehículos estadounidenses obtienen, en términos generales, mayores puntuaciones que los vehículos europeos.

**Tabla 23.27.** Valores de los centroides en la función discriminante.

	Función
Origen	1
EE.UU.	.424
Europa	-1.522

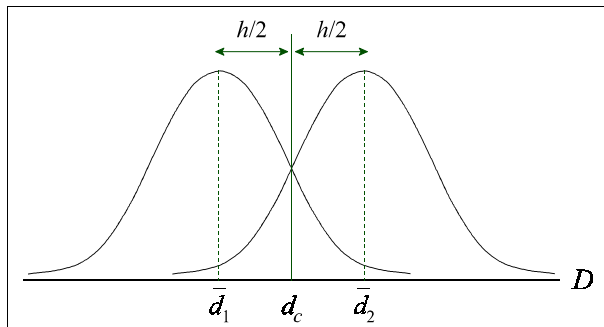
## El problema de la clasificación

En los apartados precedentes hemos estudiado, básicamente, cómo construir o estimar la función discriminante. Si nuestro 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 podemos utilizarla, en primer lugar, 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 23.8). 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.

**Figura 23.8.** Utilización de un punto de corte equidistante de ambos centroides ( $n_1 = n_2$ ).



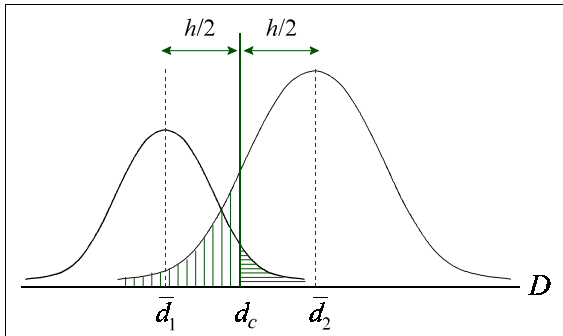
Esta regla de clasificación tiene un serio inconveniente: sólo permite distinguir entre dos grupos y es difícilmente aplicable al caso de más de dos grupos. Además, no tiene en cuenta que los



grupos pueden tener distinto tamaño. Si ambos grupos son de igual tamaño, la situación real será muy similar a la descrita en la figura 23.8. Pero si, por el contrario, los tamaños muestrales son muy desiguales, la situación real será más parecida a la que muestra la figura 23.9. 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:

$$\bar{d}_c = \frac{n_1 \bar{d}_1 + n_2 \bar{d}_2}{n_1 + n_2}$$

**Figura 23.9.** 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.

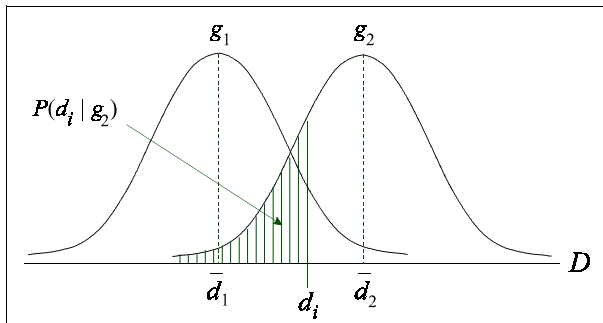
Es frecuente que, aunque los tamaños de los grupos sean intrínsecamente diferentes, se desee compensar estadísticamente esa desigualdad a la hora de clasificar a los sujetos. Esta situación es muy frecuente en el ámbito clínico cuando se comparan sujetos normales con sujetos enfermos. Si podemos estimar la proporción de casos que, en la población, pertenece a cada

uno 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 ejemplo, una sintomatología diagnóstica puede ser más frecuente en un grupo patológico que un grupo normal.

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**:  $P(D > d_i | G = g_k)$  o, simplemente,  $P(d_i | g_k)$ . La probabilidad *condicional* de una puntuación discriminante puede calcularse mediante tablas de probabilidad asintótica o a partir de los cuantiles observados (ver figura 23.10).

**Figura 23.10.** Probabilidad *condicional* de la puntuación discriminante  $d_i$  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 corres-

ponden 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 discriminante, 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:

$$P(g_k | d_i) = \frac{P(d_i | g_k) P(g_k)}{\sum_{k=1}^g P(d_i | g_k) P(g_k)}$$

El sumatorio del denominador posee tantos términos como grupos (no hay límite en el número de grupos). Con esta regla de clasificación, los casos nuevos son clasificados en el grupo al que corresponde mayor probabilidad *a posteriori*.

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*.

## Selección de las opciones de clasificación

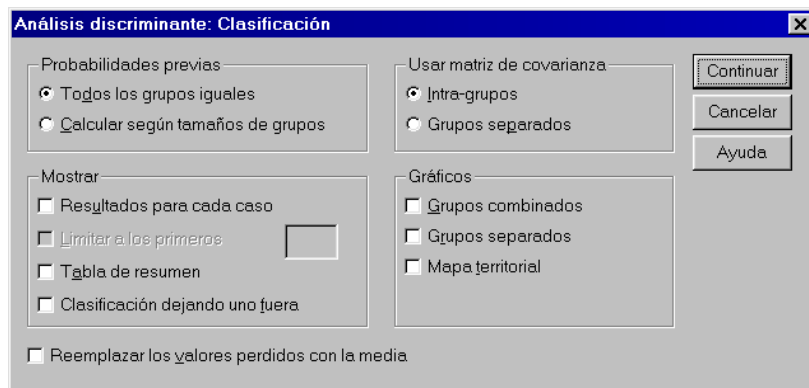
Las opciones de clasificación no afectan a la función discriminante; sólo influyen en el resultado de la clasificación de los casos.

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:

- En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), pulsar en el botón **Clasificar...** para acceder al subcuadro de diálogo *Análisis discriminante: Clasificación* que muestra la figura 23.11.

**Figura 23.11.** 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. Mediante sintaxis, es posible asignar a cada grupo probabilidades *a priori* personalizadas.

**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  $g \times g$  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.



### Ejemplo (Análisis discriminante > Clasificar)

Este ejemplo muestra cómo clasificar casos y cómo interpretar los resultados de la clasificación. Continuamos utilizando dos grupos (vehículos estadounidenses y europeos: códigos 1 y 2 de la variable *origen*) y las variables independientes que han resultado significativas en el ejemplo anterior: *consumo*, *motor* (cilindrada), *cv* (potencia), *peso* y *año*. Para solicitar los resultados de la clasificación:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), trasladar las variables *consumo*, *motor* (cilindrada), *cv* (potencia), *peso* y *año* a la lista **Independientes** y la variable *origen* al cuadro **Variable de agrupación**.
- ▶ Pulsar en **Definir rango...** para acceder al subcuadro de diálogo *Análisis discriminante: Definir rango* (ver figura 23.4), e introducir el valor 1 (código para EE.UU. en la variable *origen*) en el cuadro de texto **Mínimo** y el valor 2 (código de Europa en la variable *origen*) en el cuadro de texto **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Pulsar en el botón **Clasificar...** para acceder al subcuadro de diálogo *Análisis discriminante: Clasificación* (ver figura 23.11) y seleccionar las opciones **Resultados para cada caso**, **Tabla de resumen**, **Clasificación dejando uno fuera** del apartado **Mostrar**; y las opciones **Grupos combinados** y **Grupos separados** del apartado **Gráficos**. Seleccionar también la opción **Reemplazar los valores perdidos con la media**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece, entre otros, los resultados que muestran las tablas 23.28 y 23.29, y las figuras 23.12 a la 23.15

La tabla 23.28 se ha modificado sustancialmente llevando los casos a la dimensión de las capas y transponiendo después las filas y las columnas para reducir las dimensiones de la tabla y economizar espacio. La tabla contiene toda la información necesaria para valorar la clasificación del caso 54. La penúltima columna incluye la clasificación *original* o estándar y la última columna ofrece la clasificación resultante de la *validación cruzada*.

El vehículo en cuestión (el caso número 54) pertenece (*grupo real*) al grupo 1 (estadounidense), pero ha sido clasificado (*grupo pronosticado*) en el grupo 2 (europeo). Recordemos que el centroide del grupo estadounidense vale 0,424 y el del grupo europeo -1,522. La puntuación del caso es -1,033 (= *puntuación discriminante*). El caso se encuentra entre ambos centroides, pero más próximo al centroide de los vehículos europeos.

$P(D>d \mid G=g)$  es la *probabilidad condicional*: la probabilidad de obtener una puntuación como la obtenida o más extrema (en la dirección de la cola en la que se encuentra el caso), dentro del grupo *pronosticado* (en este caso, dentro del grupo europeo). Esta probabilidad indica el grado de *rareza* del caso dentro del grupo en el que ha sido clasificado. La probabilidad condicional del caso 54 se vale 0,625, lo que indica que ese caso se encuentra próximo al centroide del grupo *pronosticado* y no debe ser considerado un caso atípico dentro de él

**Tabla 23.28.** Estadísticos de clasificación por caso.

Número de caso: 54

				Original	Validación cruzada <sup>a</sup>
Grupo real				1	1
Grupo mayor				2**	2**
	Grupo pronosticado				
	$P(D>d \mid G=g)$	p		.625	.524
		gl		1	5
	$P(G=g \mid D=d)$			.720	.733
	Distancia de Mahalanobis al cuadrado hasta el centroide			.239	4.180
Segundo grupo mayor					
	Grupo			1	1
	$P(G=g \mid D=d)$			.280	.267
	Distancia de Mahalanobis al cuadrado hasta el centroide			2.123	6.195
Puntuaciones discriminantes Función 1				-1.033	

Para los datos originales, la distancia de Mahalanobis al cuadrado se basa en las funciones canónicas.

Para los datos validados mediante validación cruzada, la distancia de Mahalanobis al cuadrado se basa en las observaciones.

\*\*.. Caso mal clasificado

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

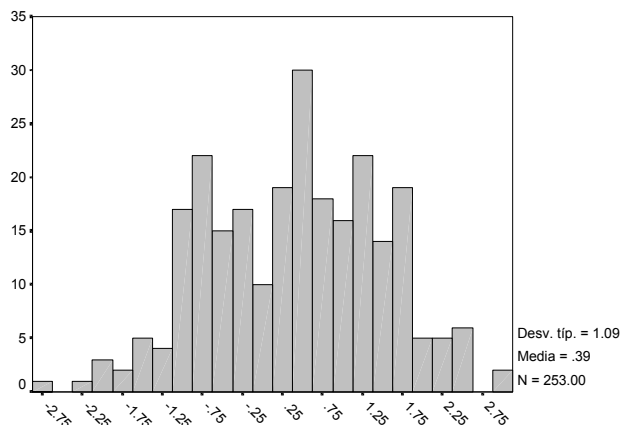
A partir de estas probabilidades *condicionales* y de las probabilidades *previas* de cada grupo (en nuestro ejemplo, 0,5 para ambos), se calculan las probabilidades *a posteriori* de cada grupo:  $P(G=g | D=d)$ . Dada una puntuación discriminante de  $-0,907$ , la probabilidad *a posteriori* de pertenecer al grupo europeo vale  $0,720$ . Y la probabilidad *a posteriori* de pertenecer al grupo estadounidense es la complementaria:  $1-0,720 = 0,280$ . Consecuentemente, el caso ha sido asignado al grupo 2, que es al que corresponde mayor probabilidad *a posteriori*.

La tabla también informa de la distancia del caso a cada uno de los centroides. Esta distancia se calcula a partir de las puntuaciones en las variables independientes originales. El caso se encuentra más próximo al grupo europeo (0,239) que al estadounidense (2,123).

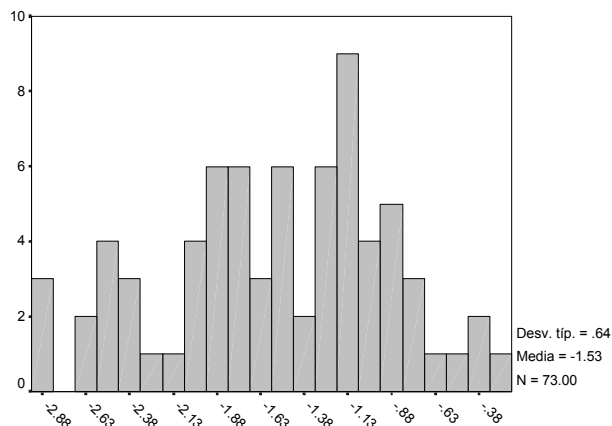
En la columna correspondiente a la *validación cruzada* podemos comprobar que el caso es clasificado también en el grupo europeo. Pero también podemos comprobar que la distancia al centroide del grupo es muy grande (4,180) en comparación con la distancia original, por lo que podemos pesar que se trata de un caso bastante extremo. No obstante, la distancia al centroide de ese grupo es menor que la distancia al centroide del otro grupo (6,195). Y la probabilidad *a posteriori* del grupo pronosticado (*grupo mayor* =  $0,733$ ) es también es mayor que la del otro grupo (*segundo grupo mayor* =  $0,267$ ).

Las figuras 23.12 y 23.13 muestran los *histogramas de las puntuaciones discriminantes*. El primero de ellos contiene los vehículos pertenecientes al grupo estadounidense. El segundo, los pertenecientes al grupo europeo.

**Figura 23.12.** Histograma de las puntuaciones discriminantes. Vehículos estadounidenses.



**Figura 23.13.** Histograma de las puntuaciones discriminantes. Vehículos europeos.



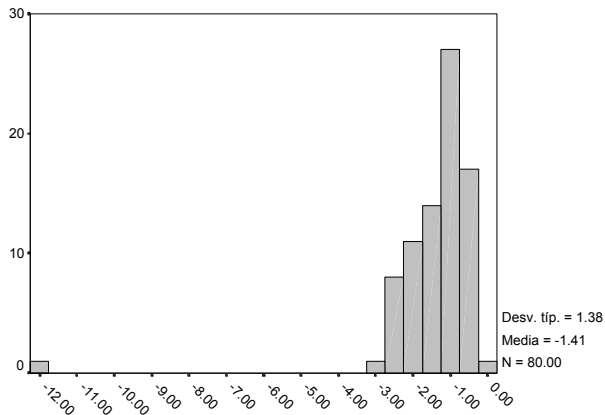
Estos histogramas permiten formarse una idea aproximada tanto de la forma de la distribución como del grado de dispersión de los vehículos dentro de su propio grupo, todo ello tomando como base sus puntuaciones en la función discriminante, o lo que es lo mismo, tomando como base sus puntuaciones en el 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 figura 23.14 muestra el histograma de las puntuaciones discriminantes de los vehículos que no tienen código de grupo o que tienen un código de grupo que se encuentra fuera del rango especificado en el análisis (es decir, los vehículos del archivo de datos que no pertenecen a ninguno de los dos grupos incluidos en el análisis).

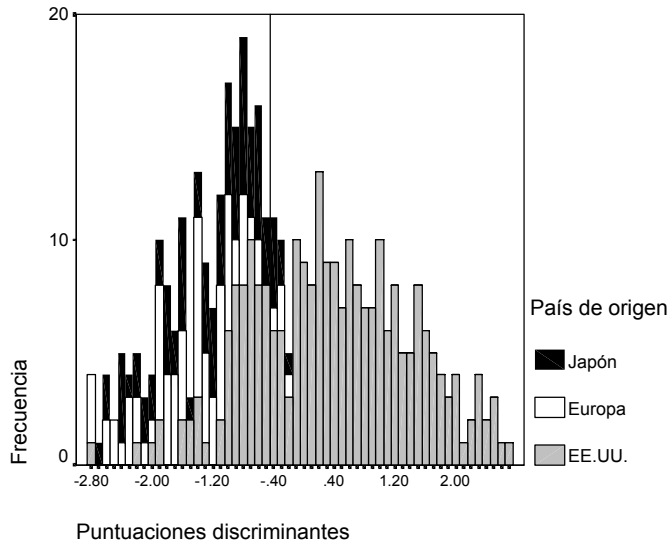
En este histograma podemos apreciar que, si obtuviéramos las puntuaciones discriminantes que les asigna la función, el grupo de vehículos no incluidos en el análisis se encontraría en la zona negativa de la función discriminante; también podemos apreciar en el histograma que existe un caso extremo.

**Figura 23.14.** Histograma de las puntuaciones discriminantes. Vehículos no incluidos en el análisis.



En el histograma de la figura 23.15 están representados los vehículos de *todos los grupos*. El gráfico incluye una línea vertical sobre el *punto de corte* que se está utilizando para la clasificación (aproximadamente -0,55). Para conocer cuál es el punto exacto se han guardado dos variables en el archivo de datos (ver, más abajo, el apartado *Guardar*): las puntuaciones discriminantes y el grupo pronosticado para cada caso. Después se ha ordenado el archivo tomando como criterio de ordenación las puntuaciones discriminantes: el *punto de corte* corresponde al valor de la función discriminante en el momento en que los casos dejan de ser clasificados en un grupo y pasan a ser clasificados en el otro.

**Figura 23.15.** Histograma de las puntuaciones discriminantes. Todos los vehículos del archivo.



Este gráfico no se encuentra disponible en el procedimiento **Discriminante** de la versión actual del programa; se ha creado mediante la opción **Barras > Apiladas** del menú **Gráficos**. Aunque este gráfico ha sido descartado de las versiones más recientes del SPSS, pensamos que es muy ilustrativo comparar la situación relativa de todos los grupos de manera simultánea, y por esta razón lo ofrecemos. En lugar del histograma con todos los grupos apilados, el *Visor* emite una advertencia indicando que ya no se ofrece tal histograma.

La tabla 23.29 muestra los *resultados de la clasificación* (la *matriz de confusión*). Esta tabla es en sí misma un procedimiento de validación de la función, pues resume la capacidad predictiva de la función discriminante. Los vehículos estadounidenses son correctamente clasificados en el 76,3 % de los casos y los vehículos europeos en el 94,5%. En total, la función consigue clasificar correctamente al 80,4 % de los casos. Si no existen datos previos acerca de la eficacia clasificatoria de otros métodos, lo apropiado es comparar estos porcentajes con la clasificación correcta esperable por azar. En nuestro ejemplo, puesto que sólo hay dos grupos de vehículos, la expectativa de clasificación correcta por azar es del 50 %.

La tabla 23.29 también incluye información sobre los casos *desagrupados* (es decir, los casos que no pertenecen a ninguno de los dos grupos utilizados en el análisis). Los resultados obtenidos indican que estos casos serían clasificados mayoritariamente (90,1 %) como vehículos europeos.

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.

**Tabla 23.29.** Resultados de la clasificación (sin tener en cuenta el tamaño de los grupos).

			Grupo de pertenencia pronosticado <sup>a,b</sup>		Total
			EE.UU.	Europa	
Original	Recuento	EE.UU.	193	60	253
		Europa	4	69	73
		Casos desagrupados	7	73	80
	%	EE.UU.	76.3	23.7	100.0
		Europa	5.5	94.5	100.0
		Casos desagrupados	8.8	91.3	100.0
Validación cruzada <sup>c</sup>	Recuento	EE.UU.	192	61	253
		Europa	4	69	73
	%	EE.UU.	75.9	24.1	100.0
		Europa	5.5	94.5	100.0

a. Clasificados correctamente el 80.4% de los casos agrupados originales.

b. Clasificados correctamente el 80.1% de los casos agrupados validados mediante validación cruzada.

c. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

La tabla de clasificación muestra que el número de casos mal clasificados es mayor en el grupo más grande (es decir, en el grupo de vehículos estadounidenses). Esto se debe a que el *punto de corte* para la clasificación se ha establecido a partir de *probabilidades previas idénticas para ambos grupos*. Para disminuir el porcentaje de clasificación incorrecta de ese grupo podemos seleccionar, en el subcuadro de diálogo *Análisis discriminante: Clasificación* (figura 23.11), la opción **Calcular según tamaños de grupos** del apartado **Probabilidades previas**.

Al repetir el análisis utilizando esta otra estrategia obtenemos una tabla de *probabilidades previas* (tabla 23.30) que contiene las probabilidades *a priori* asignadas a cada grupo. Estas probabilidades, según hemos señalado ya, únicamente reflejan el tamaño relativo de cada grupo. En el cálculo de estas probabilidades no intervienen los casos con valor perdido (aunque se haya marcado la opción **Reemplazar valores perdidos con la media**).

**Tabla 23.30.** Probabilidades previas.

País de origen	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
EE.UU.	.782	244	244.000
Europa	.218	68	68.000
Total	1.000	312	312.000



La tabla de clasificación (tabla 23.31) permite comprobar que la nueva regla de clasificación arroja resultados distintos de los obtenidos anteriormente. Ahora, el porcentaje de clasificación correcta de los vehículos estadounidenses (los vehículos del grupo más grande) ha subido del 76,3 % al 95,3 %. Como contrapartida, el porcentaje de vehículos correctamente clasificados en el grupo europeo (el grupo más pequeño) ha bajado del 94,5 % al 63,0 %. En conjunto, el porcentaje de casos correctamente clasificados ha pasado del 80,4 % al 88,0 %, lo que representa una ganancia nada despreciable: la nueva regla clasifica correctamente 25 vehículos más (287 en lugar de 262). Por otro lado, también los vehículos no agrupados (es decir, los que no pertenecen a ninguno de los dos grupos incluidos en el análisis) se han visto afectados por la nueva regla de clasificación: ahora se distribuyen de manera homogénea entre los dos grupos.

**Tabla 23.31.** Resultados de la clasificación (teniendo en cuenta el tamaño de los grupos).

			Grupo de pertenencia pronosticado <sup>a,b</sup>		Total
			EE.UU.	Europa	
Original	Recuento	EE.UU.	241	12	253
		Europa	27	46	73
		Casos desagrupados	44	36	80
	%	EE.UU.	95.3	4.7	100.0
		Europa	37.0	63.0	100.0
		Casos desagrupados	55.0	45.0	100.0
Validación cruzada <sup>c</sup>	Recuento	EE.UU.	239	14	253
		Europa	28	45	73
	%	EE.UU.	94.5	5.5	100.0
		Europa	38.4	61.6	100.0

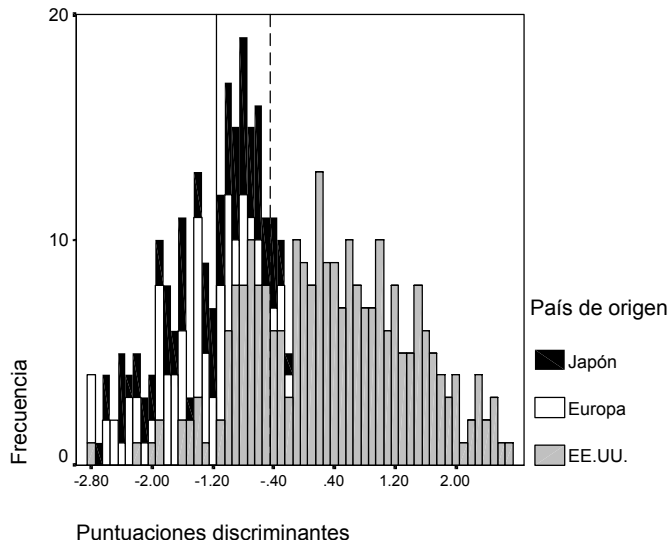
a. Clasificados correctamente el 88.0% de los casos agrupados originales.

b. Clasificados correctamente el 87.1% de los casos agrupados validados mediante validación cruzada.

c. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

El histograma conjunto de la figura 23.16 muestra los dos puntos de corte utilizados. El trazo discontinuo corresponde al punto de corte  $(-0,55)$  resultante de aplicar el criterio que atribuye igual probabilidad *previa* a los dos grupos; y el trazo continuo corresponde al punto de corte  $(-1,20)$  obtenido con el criterio que atribuye mayor probabilidad al grupo más grande. La figura muestra con claridad que el punto de corte correspondiente al segundo criterio se ha desplazado hacia la izquierda, alejándose del centroide del grupo más grande e invadiendo el territorio del grupo más pequeño. De hecho, el vehículo número 54, que inicialmente había sido clasificado en el grupo europeo (a pesar de ser un vehículo estadounidense), con la regla de clasificación basada en el segundo criterio ha sido clasificado en el grupo estadounidense. (A pesar de este cambio en la clasificación del vehículo 54, conviene señalar que su puntuación discriminante no ha cambiado; sólo ha cambiado el criterio de clasificación).

**Figura 23.16.** Histograma de las puntuaciones discriminantes. Todos los vehículos del archivo.



La rareza del caso 54 en el grupo al que realmente pertenece (grupo 1) ha aumentado respecto a la obtenida en la clasificación anterior (grupo 2). Su *probabilidad condicional* vale ahora 0,145, lo que significa que se encuentra más alejado del centro de su propio grupo de lo que se encontraba del grupo en el que fue clasificado anteriormente.

**Tabla 23.32.** Estadísticos para la clasificación del caso 54 con el nuevo criterio.

Número de caso: 54

			Original	Validación cruzada <sup>a</sup>
Grupo real			1	1
Grupo mayor	Grupo pronosticado		1	1
	$P(D>d \mid G=g)$	p	.145	.288
		gl	1	5
	$P(G=g \mid D=d)$		.583	.567
	Distancia de Mahalanobis al cuadrado hasta el centroide		2.123	6.195
Segundo grupo mayor	Grupo		2	2
	$P(G=g \mid D=d)$		.417	.433
	Distancia de Mahalanobis al cuadrado hasta el centroide		.239	4.180
Puntuaciones discriminantes	Función 1		-1.033	

Para los datos originales, la distancia de Mahalanobis al cuadrado se basa en las funciones canónicas.

Para los datos validados mediante validación cruzada, la distancia de Mahalanobis al cuadrado se basa en las observaciones.

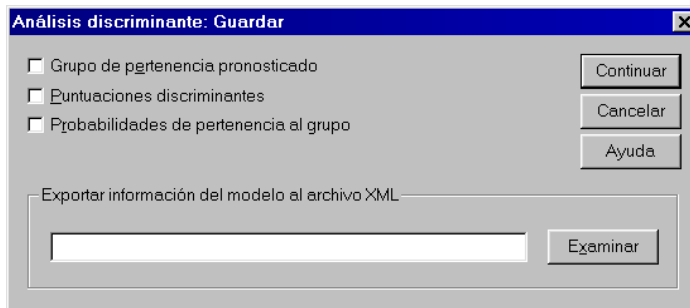
- a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

## 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 (cálculo de la curva COR, etc.). Para crear estas nuevas variables:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), pulsar en el botón **Guardar...** para acceder al subcuadro de diálogo *Análisis discriminante: Guardar* que muestra la figura 23.17.

**Figura 23.17.** 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.

**Exportar información del modelo al archivo XML.** Permite exportar la información del modelo a un archivo en formato *XML*. Los programas *SmartScore* y las próximas versiones de *WhatIf?* pueden utilizar este archivo para crear distintos escenarios de clasificación. Pulsando en el botón **Examinar** puede seleccionarse la carpeta en la que se desea ubicar el archivo, el nombre del mismo y su formato.

## Seleccionar

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:

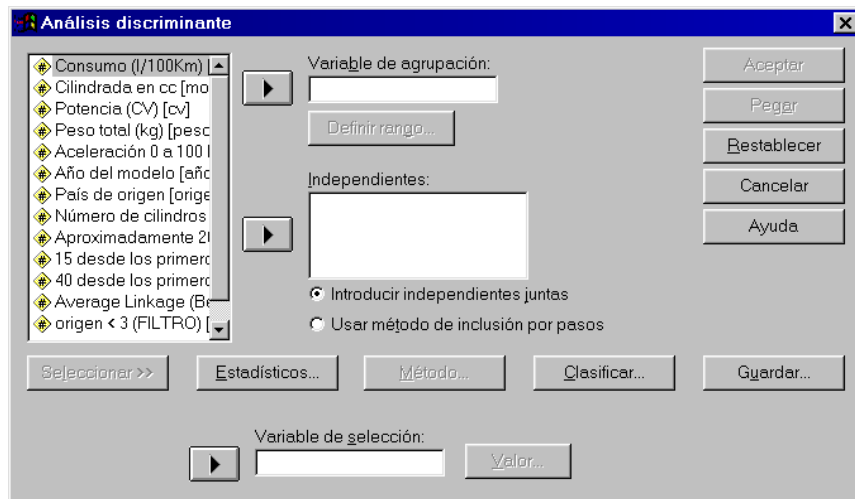
- 1) Seleccionar, de la muestra original, un subconjunto aleatorio de casos (*muestra de validación*);
- 2) Estimar la función discriminante con los casos restantes (*muestra de entrenamiento*);
- 3) Utilizar esa función para clasificar los casos de la muestra de *validación*.

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:

- En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), pulsar el botón **Seleccionar>>** para expandir el cuadro de diálogo y hacer que tome el aspecto que muestra la figura 23.18.

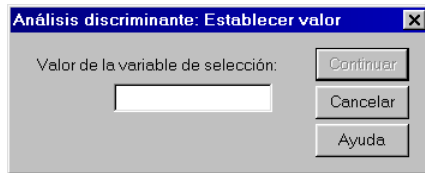
**Figura 23.18:** Cuadro de diálogo *Análisis discriminante* (expandido con la opción *Seleccionar*).



- Trasladar la variable de *selección* (la variable que identifica a los casos de la muestra de *validación* y a los de la muestra de *entrenamiento*) y trasladarla al cuadro **Variable de selección**.

- Pulsar en el botón **Valor...** para acceder al subcuadro de diálogo *Análisis discriminante: Establecer valor* que se muestra en la figura 23.19.

**Figura 23.19.** Subcuadro de diálogo Establecer valor.



- Introducir en el cuadro **Valor de la variable de selección** el valor que identifica a los casos que serán incluidos en el análisis. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

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).



### Ejemplo (Análisis discriminante > Seleccionar)

Este ejemplo ilustra el proceso de validación cruzada, es decir, explica cómo seleccionar una *muestra de entrenamiento* y cómo clasificar los casos de la *muestra de validación*. Comenzaremos creando una variable de *selección* (la variable que distingue entre la muestra de *validación* y la de *entrenamiento*); continuaremos estimando la función discriminante a partir de la muestra de *entrenamiento*; y terminaremos utilizando la función discriminante obtenida para clasificar los casos de la muestra de *validación*. Para crear la variable de *selección* podemos utilizar la opción **Seleccionar casos** del menú **Datos**:

- ▶ En la ventana del *Editor de datos*, seleccionar la opción **Seleccionar casos** del menú **Datos** para acceder al cuadro de diálogo *Seleccionar casos*.
- ▶ Marcar la opción **Muestra aleatoria de casos** y pulsar en el botón **Muestra...** para acceder al subcuadro de diálogo *Seleccionar casos: Muestra aleatoria*.
- ▶ Seleccionar la opción **Aproximadamente k % de todos los casos** e introducir el valor 50 en el correspondiente cuadro de texto (para seleccionar una muestra aleatoria de aproximadamente el 50 % de los casos).
- ▶ Pulsar en el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, El *Editor de datos* genera una variable llamada *filter\_\$* que contiene el valor *uno* para los casos seleccionados (aproximadamente el 50% de los casos del archivo) y el valor *cero* para los no seleccionados. El archivo de datos queda filtrado con los casos seleccionados. Para desactivar el filtro:

- ▶ En la ventana del *Editor de datos*, seleccionar la opción **Seleccionar casos** del menú **Datos** para acceder al cuadro de diálogo *Seleccionar casos*.
- ▶ Marcar la opción **Todos los casos** y pulsar el botón **Aceptar** para desactivar cualquier filtro que se encuentre activo.

Puesto que la variable *filter\_\$* permanece en el archivo de datos aunque desactivemos el filtrado de casos, ya disponemos de la variable de *selección* (es decir, de una variable en la que la mitad de los casos de la muestra tiene el valor *cero* y la otra mitad el valor *uno*). Para comenzar el proceso de *validación cruzada*:

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), trasladar las variables *consumo*, *motor* (cilindrada), *cv* (potencia), *peso* y *año* a la lista **Independientes**.
- ▶ Trasladar la variable *origen* al cuadro **Variable de agrupación**.
- ▶ Pulsar en **Definir rango...** para acceder al subcuadro de diálogo *Análisis discriminante: Definir rango* (ver figura 23.4).
- ▶ Introducir el valor 1 (código de EE.UU. en la variable *origen*) en el cuadro de texto **Mínimo** y el valor 2 (código de Europa en la variable *origen*) en el cuadro de texto **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Pulsar en el botón **Clasificar...** para acceder al subcuadro de diálogo *Análisis discriminante: Clasificación* (ver figura 23.11).
- ▶ Seleccionar la opción **Calcular según tamaños de grupos** del apartado **Probabilidades previas**, y la opción **Tabla resumen** del apartado **Mostrar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Pulsar el botón **Seleccionar>>** para expandir el cuadro de diálogo *Análisis discriminante* (ver figura 23.18).
- ▶ Trasladar la variable *filter\_\$* recién creada al cuadro **Variable de selección**.
- ▶ Pulsar en el botón **Valor...** para acceder al subcuadro de diálogo *Análisis discriminante: Establecer valor* (ver figura 23.19).
- ▶ Introducir el valor 1 en el cuadro de texto **Valor de la variable de selección**. Pulsar en el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestra la tabla 23.33. Por supuesto, en la *fase de estimación* del análisis intervienen únicamente los casos con valor *uno* en la variable de *selección* (es decir, aproximadamente el 50 % de los casos de la muestra), mientras que la *fase de clasificación* afecta tanto a los casos de la muestra de *entrenamiento* como a los de la muestra de *validación*.

La tabla 23.33 contiene las *matrices de confusión* correspondientes a los casos seleccionados (la muestra de *entrenamiento*) y a los no seleccionados (la muestra de *validación*). En la muestra de *entrenamiento* se obtiene una tasa de acierto del 89,5 % y, en la de *validación*, del 87,4 %. Podemos esperar, por tanto, que la función discriminante obtenida clasifique correctamente al 87,4 % de los futuros casos nuevos que se intenten clasificar.

**Tabla 23.33.** Tabla de clasificación (función obtenida con la muestra de entrenamiento).

				Grupo de pertenencia pronosticado <sup>a,b</sup>		Total
				EE.UU.	Europa	
Casos seleccionados	Original	Recuento	País de origen			
			EE.UU.	113	7	120
			Europa	9	24	33
			Casos desagrupados	32	21	53
		%	EE.UU.	94.2	5.8	100.0
			Europa	27.3	72.7	100.0
Casos no seleccionados	Original	Recuento	EE.UU.	119	5	124
			Europa	15	20	35
			Casos desagrupados	11	16	27
		%	EE.UU.	96.0	4.0	100.0
			Europa	42.9	57.1	100.0
			Casos desagrupados	40.7	59.3	100.0

a. Clasificados correctamente el 89.5% de los casos agrupados originales seleccionados.

b. Clasificados correctamente el 87.4% de casos agrupados originales no seleccionados.

Podemos repetir el proceso de validación cruzada ejecutando de nuevo el análisis sobre los casos con código *cero* en la variable de *selección*. Para ello, basta con cambiar el valor de la variable de *selección* en el subcuadro de diálogo *Análisis discriminante: Establecer valor* (ver figura 23.19). Sin embargo, esto puede hacerse también con el proceso *Ejecutar casos no seleccionados*. Para ejecutar este proceso:

- ▶ En el panel izquierdo (esquema) del *Visor*, pulsar sobre el icono de *libro cerrado* del título *Notas*. (Con ello queda seleccionada tabla de *Notas*, condición necesaria para poder ejecutar el proceso *Ejecutar casos no seleccionados* con el procedimiento *Análisis discriminante*).
- ▶ Seleccionar la opción **Ejecutar proceso...** del menú **Utilidades** para acceder al cuadro de diálogo *Ejecutar proceso*.
- ▶ Seleccionar el proceso *Ejecutar casos no seleccionados.sbs* que se encuentra en la carpeta *Scripts* que cuelga de la carpeta en la que está instalado el SPSS.
- ▶ Pulsar en el botón **Ejecutar** para ejecutar el proceso y, con él, el nuevo análisis.

Al ejecutar este proceso, el SPSS repite el análisis previo conmutando la muestra de *validación* por la muestra de *entrenamiento*. Y el *Visor* ofrece, entre otras cosas, el resultado de la clasificación que muestra la tabla 23.34.

Por tanto, la tabla 23.34 contiene el resultado de clasificación tras intercambiar las muestras de *entrenamiento* y *validación* del primer análisis. Podemos comprobar que el porcentaje de clasificación correcta en la nueva muestra de *entrenamiento* es del 86,8 %, y del 87,6 % en la nueva muestra de *validación*. Basándonos en estos resultados, podemos concluir que, si utilizamos cualquiera de las dos funciones obtenidas para clasificar nuevos casos, podemos esperar que el porcentaje de clasificación correcta se encuentre en torno al 87,5 %.

**Tabla 23.34.** Tabla de clasificación (función obtenida con la muestra de validación).

				Grupo de pertenencia pronosticado <sup>a,b</sup>		Total
				EE.UU.	Europa	
Casos seleccionados	Original	Recuento	Pais de origen			
			EE.UU.	117	7	124
			Europa	14	21	35
			Casos desagrupados	13	14	27
		%	EE.UU.	94.4	5.6	100.0
			Europa	40.0	60.0	100.0
Casos no seleccionados	Original	Recuento	EE.UU.	111	9	120
			Europa	10	23	33
			Casos desagrupados	34	19	53
		%	EE.UU.	92.5	7.5	100.0
			Europa	30.3	69.7	100.0
			Casos desagrupados	64.2	35.8	100.0

a. Clasificados correctamente el 86.8% de los casos agrupados originales seleccionados.

b. Clasificados correctamente el 87.6% de casos agrupados originales no seleccionados.

## El caso de más de dos grupos

Aunque hasta ahora hemos basado toda nuestra exposición del análisis discriminante en la clasificación de casos en dos grupos, lo cierto es que la 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.

Además, las funciones resultantes son ortogonales o 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.

### Ejemplo (Análisis discriminante con tres grupos)

Este ejemplo muestra cómo llevar a cabo un análisis discriminante con tres grupos utilizando un método de estimación *por pasos*. Seguimos utilizando las variables del archivo *Coches.sav* (pero hemos filtrado el caso 35 para facilitar la lectura de los gráficos):

- ▶ En el cuadro de diálogo *Análisis discriminante* (ver figura 23.3), trasladar las variables *consumo*, *motor* (cilindrada), *cv* (potencia), *peso*, *acel* (aceleración), *año* y *cilindr* (número de cilindros) a la lista **Independientes**. Trasladar la variable *origen* al cuadro **Variable de agrupación**.
- ▶ Pulsar en **Definir rango...** para acceder al subcuadro de diálogo *Análisis discriminante: Definir rango* (ver figura 23.4). Introducir el valor 1 (código de EE.UU. en la variable *origen*) en el cuadro de texto **Mínimo** y el valor 3 (código de Japón en la variable *origen*) en el cuadro de texto **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Seleccionar la opción **Usar método de inclusión por pasos** y pulsar en el botón **Método...** para acceder al subcuadro de diálogo *Análisis discriminante: Método de inclusión por pasos* (ver figura 23.7). Seleccionar la opción **F para distancias por parejas** del apartado **Mostrar**. Pulsar en el botón **Continuar** para volver al cuadro de diálogo principal.
- ▶ Pulsar en el botón **Clasificar...** para acceder al subcuadro de diálogo *Análisis discriminante: Clasificación* (ver figura 23.11). Seleccionar las opciones **Resultados para cada caso** y **Tabla resumen** del apartado **Mostrar**.
- ▶ Seleccionar las opciones **Grupos combinados**, **Grupos separados** y **Mapa territorial** del apartado **Gráficos**.
- ▶ Seleccionar la opción **Reemplazar los valores perdidos con la media**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece, entre otros, los resultados que se muestran a continuación (algunas tablas han sido modificadas para economizar espacio).

La tabla 23.35 muestra el número de casos válidos de cada grupo. Puede observarse que, ahora, el grupo de vehículos japoneses se incluye como un grupo más.

**Tabla 23.35.** Tamaños muestrales de cada grupo.

N válido (según lista) No ponderados

País de origen	Consumo (l/100Km)	Cilindrada en cc	Potencia (CV)	Peso total (kg)	Aceleración 0 a 100 km/h	Año del modelo	Número de cilindros
EE.UU.	244	244	244	244	244	244	244
Europa	68	68	68	68	68	68	68
Japón	79	79	79	79	79	79	79
Total	391	391	391	391	391	391	391

La tabla 20.36 contiene las variables independientes incluidas en el modelo en el último paso. La tabla muestra que han quedado fuera del modelo las variables *aceleración* y *cilindr* (número de cilindros).

**Tabla 20.36.** Variables incluidas en el modelo (último paso).

Paso: 5

	Tolerancia	F para eliminar	Lambda de Wilks
Cilindrada en cc	.131	47.381	.587
Potencia (CV)	.204	19.658	.519
Año del modelo	.521	11.780	.500
Peso total (kg)	.134	8.939	.493
Consumo (l/100Km)	.169	4.003	.480



La tabla 23.37 muestra cómo, a medida que se van incorporando nuevas variables al modelo en cada paso, los valores de la *lambda* de Wilks global y del estadístico *F* asociado a ella van disminuyendo.

**Tabla 23.37:** *Lambda* de Wilks.

Paso	Número de variables	Lambda	gl1	gl2	gl3	F exacta			
						Estadístico	gl1	gl2	Sig.
1	1	.570	1	2	388	146.218	2	388.000	.000
2	2	.518	2	2	388	75.407	4	774.000	.000
3	3	.500	3	2	388	53.292	6	772.000	.000
4	4	.480	4	2	388	42.610	8	770.000	.000
5	5	.471	5	2	388	35.149	10	768.000	.000

La tabla 23.38 ofrece las comparaciones entre pares de grupos. Los valores del estadístico *F* no coinciden con los del estadístico *F* asociado a la *lambda* de Wilks global. La tabla muestra todas las comparaciones posibles entre cada dos grupos y el estadístico *F* y su significación para esa comparación. Vemos que en el primer paso (al incluir la variable *cilindrada* en el modelo) se consigue distinguir significativamente a los vehículos estadounidenses de los europeos y los japoneses, pero no se consigue discriminar a los japoneses de los europeos. Hasta el paso 3 (momento en el que se incorpora al modelo la variable *año del modelo*), no se consigue diferenciar a estos dos grupos.

**Tabla 23.38.** Comparaciones entre grupos por pares.

País de origen			Paso				
			1	2	3	4	5
EE.UU.	Europa	F	160.560	97.082	69.993	54.299	45.756
		Sig.	.000	.000	.000	.000	.000
	Japón	F	198.782	124.100	82.545	62.926	51.265
		Sig.	.000	.000	.000	.000	.000
Europa	Japón	F	.280	.381	3.517	6.646	5.541
		Sig.	.597	.684	.015	.000	.000

Los *autovalores* (tabla 23.39) de las dos funciones que componen el modelo son muy desiguales. La primera función explica el 93,2 % de la variabilidad disponible en los datos, mientras que la segunda función sólo explica el 6,8%. De manera similar, la correlación *canónica* de la primera función es alta (0,704), mientras que la de la segunda función es más bien baja (0,259).

**Tabla 23.39.** Autovalores.

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	.982 <sup>a</sup>	93.2	93.2	.704
2	.072 <sup>a</sup>	6.8	100.0	.259

a. Se han empleado las 2 primeras funciones discriminantes canónicas en el análisis.

La *lambda* de Wilks de la tabla 23.40 contrasta de manera jerárquica la significación de las dos funciones obtenidas. En la primera línea (1 a la 2) se contrasta la hipótesis nula de que el modelo completo (ambas funciones discriminantes tomadas juntas) no permite distinguir las medias de los grupos. Puesto que el valor de la *lambda* de Wilks (que coincide con el valor de la *lambda* del último paso de construcción del modelo; ver tabla 23.37) tiene asociado un nivel crítico (*Sig.* = 0,000) menor que 0,05, podemos concluir que el modelo permite distinguir significativamente entre los grupos.

En la segunda línea (2) se contrasta si las medias de los grupos son iguales en la segunda función discriminante. La *lambda* de Wilks toma un valor muy próximo a 1, pero el nivel crítico (*Sig.* = 0,000) es menor que 0,05, por lo que podemos concluir que la segunda función permite discriminar entre, al menos, dos de los grupos.

Para saber entre qué grupos permite distinguir cada función, debemos basarnos en las comparaciones por pares que ofrece la tabla 23.38.

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.

**Tabla 23.40.** *Lambda* de Wilks. Contraste de las funciones del modelo.

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 2	.471	290.917	10	.000
2	.933	26.886	4	.000

La tabla 23.41 muestra la *ubicación de los centroides* en cada una de las funciones discriminantes. La primera función distingue fundamentalmente a los vehículos estadounidenses (cuyo centroide está ubicado en la parte positiva) de los vehículos europeos y japoneses (cuyos centroides se encuentran en la parte negativa).

En la segunda función, el centroide de los vehículos japoneses se sitúa en la parte positiva, mientras que el de los vehículos europeos se sitúa en la parte negativa; el de los vehículos estadounidenses queda en la parte central. Dado que la primera función ha conseguido explicar el máximo de las diferencias existentes entre los vehículos estadounidenses y el resto, es lógico que la segunda función discrimine precisamente entre los dos grupos que han quedado más próximos en la primera.

**Tabla 23.41.** Valor de los centroides en las funciones discriminantes.

País de origen	Función	
	1	2
EE.UU.	.766	.001
Europa	-1.265	-.472
Japón	-1.278	.403

La matriz de *coeficientes estandarizados* (tabla 23.42) contiene ahora dos columnas, una para cada función discriminante. Las funciones se encuentran siempre ordenadas en correspondencia con los autovalores de la tabla 23.39, siendo la primera función la de mayor capacidad discriminativa. Los coeficientes estandarizados de la primera función no difieren sustancialmente de los obtenidos en la función estimada en el caso de dos grupos (ver tabla 23.25). Esta primera función discrimina, fundamentalmente, entre vehículos de gran *cilindrada* y vehículos más optimizados en *potencia*. Puesto que el único centroide positivo en esta primera función (ver tabla 23.41) es el de los coches estadounidenses, podemos interpretar que los vehículos de *gran cilindrada* y *poca potencia* tienden a ser clasificados como estadounidenses. Y lo mismo vale decir de los vehículos que *consumen más* y que tienen *menor peso* y *antigüedad*.

La segunda función atribuye la mayor ponderación al *peso* del vehículo. Puesto que ahora el centroide de los vehículos europeos es negativo y el de los japoneses positivo, podemos interpretar que los coches con *mayor peso* tenderán a ser clasificados como europeos. Mientras que los vehículos con mayor *cilindrada*, los *menos antiguos* (año del modelo más alto), los *más potentes* y los que *más consumen* tenderán a ser clasificados como vehículos japoneses.

**Tabla 23.42.** Coeficientes estandarizados de las funciones discriminantes canónicas.

	Función	
	1	2
Consumo (l/100Km)	.456	.511
Cilindrada en cc	1.697	1.134
Potencia (CV)	-.925	.670
Peso total (kg)	-.296	-2.070
Año del modelo	.305	.983

La *matriz de estructura* (tabla 23.43) ofrece los coeficientes de correlación entre las variables independientes y las puntuaciones discriminantes de cada función. El coeficiente más alto de cada variable aparece marcado con un asterisco que indica cuál es la función con la que más correlaciona esa variable (lo que no significa que sea ésa la función en la que más discrimina la variable). Si existe alta colinealidad (alta relación entre las variables independientes), los coeficientes de esta tabla puede ser muy distintos de los coeficientes estandarizados, como de hecho sucede. En nuestro ejemplo, la primera función correlaciona con la *cilindrada*, el *peso*, el *consumo* y la *potencia*; la segunda función correlaciona con el *año del modelo*.

**Tabla 23.43.** Matriz de estructura.

	Función	
	1	2
Cilindrada en cc	.876*	-.087
Número de cilindros <sup>a</sup>	.834*	-.104
Peso total (kg)	.762*	-.348
Consumo (l/100Km)	.669*	-.308
Potencia (CV)	.565*	-.016
Aceleración 0 a 100 km/h <sup>a</sup>	-.251*	-.242
Año del modelo	-.138	.557*

\*. Mayor correlación absoluta entre cada variable y cualquier función discriminante.

a. Esta variable no se emplea en el análisis.

Hasta aquí hemos discutido el proceso de construcción o estimación del modelo. Para valorar la capacidad predictiva del modelo estimado debemos prestar atención a los resultados de la clasificación.

La tabla 23.44 ofrece las probabilidades *previas* o *a priori*. Estas probabilidades indican que se ha dado la misma importancia relativa a todos los grupos: 0,333 (a pesar de que los vehículos estadounidenses constituyen más del 60% de la muestra). Enseguida veremos qué ocurre si utilizamos probabilidades previas basadas en los tamaños de los grupos.

**Tabla 23.44.** Probabilidades previas (probabilidades *a priori* utilizadas en la clasificación).

País de origen	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
EE.UU.	.333	244	244.000
Europa	.333	68	68.000
Japón	.333	79	79.000
Total	1.000	391	391.000

La figura 23.20 muestra el *mapa territorial*. Un mapa territorial 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.

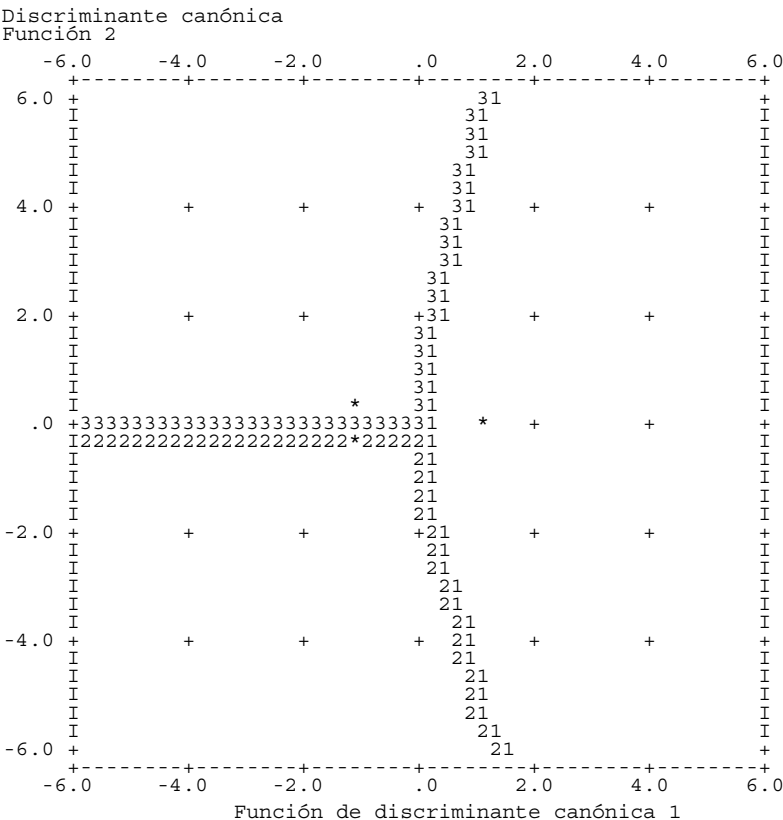
Los centroides de cada grupo están representados por asteriscos. Para representar los centroides se utilizan las coordenadas de la tabla de centroides (ver tabla 23.41). Observando la ubicación de los centroides en la figura 23.20 se aprecia claramente que la primera función posee mayor capacidad discriminativa 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.

Prestando atención a la disposición de los tres territorios sobre el mapa, resulta fácil anticipar que los vehículos con puntuaciones altas en la primera función discriminante serán clasificados en el grupo estadounidense (grupo 1), mientras que los vehículos con puntuaciones próximas a cero o negativas en esa función serán clasificados en el grupo europeo (grupo 2) o japonés (grupo 3). En este segundo caso, si la puntuación del vehículo en la segunda función discriminante es positiva será clasificado en el grupo japonés, mientras que si la puntuación en esa función es negativa será clasificado en el grupo europeo.

Figura 23.20. Mapa territorial definido por las dos funciones discriminantes.



Símbolos usados en el mapa territorial

Símbolo	Grupo	Etiqueta
-----	-----	-----
1	1	EE.UU.
2	2	Europa
3	3	Japón
*		Indica un centroide de grupo



La tabla 23.45 ofrece los resultados de la clasificación para el caso 54. Su grupo real o nominal es el estadounidense (grupo 1), pero ha sido clasificado en el grupo europeo (grupo 2). Su puntuación discriminante en la primera función (−0,474) hace que sea clasificado como vehículo no perteneciente al grupo 1, es decir como vehículo europeo o japonés (ver centroides en la tabla 23.41); y su puntuación discriminante en la segunda función (−1,078) hace que sea clasificado como vehículo europeo (recordemos que la primera función permite discriminar entre vehículos estadounidenses y vehículos europeos-japoneses; y la segunda entre vehículos europeos y japoneses).

La probabilidad *condicional*,  $P(D>d \mid G=g)$ , del caso 54 vale 0,609, lo que permite afirmar que se trata de un vehículo bastante *centrado* en el grupo en el que ha sido clasificado (grupo 2). Echando un vistazo a sus características podemos comprobar que se trata, básicamente, de un vehículo cuya *cilindrada*, *peso*, *potencia* y *año* están claramente por debajo de la media (ver, más arriba, el comentario de la tabla 23.42).

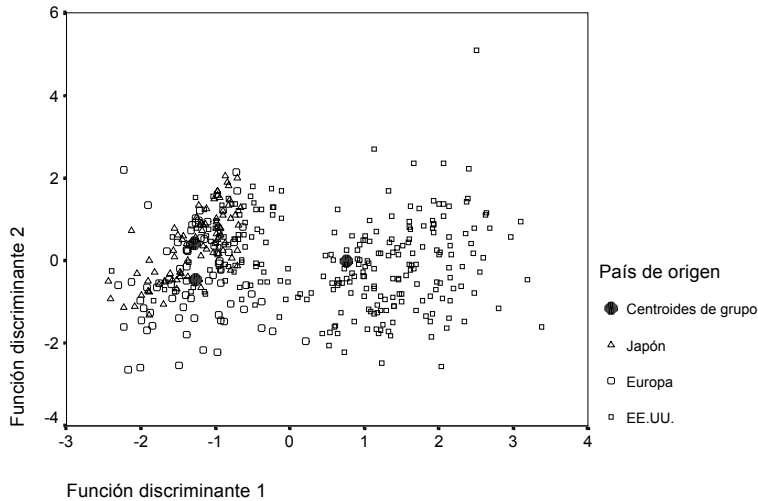
Tabla 23.45. Resultado de la clasificación (caso número 54).

Número de caso: 54			Original
Grupo real			1
Grupo mayor	Grupo pronosticado		2**
	P(D>d   G=g)	p	.609
		gl	2
	P(G=g   D=d)		.549
	Distancia de Mahalanobis al cuadrado hasta el centroide		.993
Segundo grupo mayor	Grupo		1
	P(G=g   D=d)		.234
	Distancia de Mahalanobis al cuadrado hasta el centroide		2.702
Puntuaciones discriminantes	Función 1		-.474
	Función 2		-1.078

\*\* . Caso mal clasificado

La figura 23.21 muestra el diagrama de dispersión de todos los casos utilizados en el análisis sobre el plano definido por las dos funciones discriminantes. Los casos están identificados por el país de origen de los vehículos. La mayor utilidad de este gráfico radica en la posibilidad de identificar casos atípicos difíciles de clasificar.

**Figura 23.21.** Diagrama de dispersión de los tres grupos en las dos funciones discriminantes.



En este ejemplo, el diagrama de dispersión también nos ofrece pistas sobre la conveniencia de aumentar la probabilidad *a priori* del grupo estadounidense para mejorar la clasificación, pues la primera función discriminante (el eje horizontal) parece distinguir fácilmente a los vehículos japoneses y europeos de los estadounidenses.

Por último, la *matriz de confusión* de la tabla 23.46 ofrece los resultados de la clasificación. La tabla indica que se ha clasificado correctamente el 67,4 % de los vehículos, lo cual, comparado con el 33% esperable en una clasificación completamente al azar, puede interpretarse como una mejora considerable.

Los errores de clasificación no se distribuyen de manera simétrica. En el grupo de vehículos estadounidenses se consigue el porcentaje más alto de clasificación correcta, 68,8 %, frente a un porcentaje del 61,6% en el grupo europeo y del 68,4 % en el grupo japonés. (Esta circunstancia resulta especialmente llamativa pues, a pesar de que la regla de clasificación se basa en probabilidades *a priori* iguales para todos los grupos, el porcentaje de clasificación correcta más alto se da precisamente en el grupo de mayor tamaño).

Basándonos en los porcentajes de clasificación correcta de cada grupo podemos afirmar que los vehículos estadounidenses se confunden, mayoritariamente, con los vehículos japoneses; y que los vehículos europeos y japoneses no se confunden con los estadounidenses, sino entre sí.

**Tabla 23.46.** Resultados de la clasificación (probabilidades previas iguales).

			Grupo de pertenencia pronosticado <sup>a</sup>			Total
			EE.UU.	Europa	Japón	
Original	Recuento	EE.UU.	174	29	50	253
		Europa	1	45	27	73
		Japón	0	25	54	79
	%	EE.UU.	68.8	11.5	19.8	100.0
		Europa	1.4	61.6	37.0	100.0
		Japón	.0	31.6	68.4	100.0

a. Clasificados correctamente el 67.4% de los casos agrupados originales.

Vamos a repetir el análisis con las probabilidades *a priori* calculadas a partir del tamaño de los grupos. Para ello,

- ▣ Repetir el análisis marcando la opción **Calcular según tamaños de grupos** del apartado **Probabilidades previas** en el cuadro de diálogo *Análisis discriminante: Clasificación* (ver figura 23. 11).

Procediendo de esta manera, la *matriz de confusión* ofrece los resultados que muestra la tabla 23.47. El porcentaje de clasificación correcta ha subido del 67,4 % al 73,8 %. Al variar los territorios con la nueva regla de clasificación, ha aumentado el porcentaje de clasificación correcta de los vehículos más numerosos (los estadounidenses), pero algunos de los vehículos europeos y japoneses se confunden con los vehículos estadounidenses y la tasa de clasificación correcta del grupo europeo se ha reducido considerablemente. Probablemente las probabilidades previas podrían ser mejor calibradas y ello nos permitiría obtener mejores resultados en la clasificación.

**Tabla 23.47.** Resultados de clasificación (probabilidades previas basadas en los tamaños de los grupos).

			Grupo de pertenencia pronosticado <sup>a</sup>			Total
			EE.UU.	Europa	Japón	
Original	Recuento	EE.UU.	211	12	30	253
		Europa	9	36	28	73
		Japón	4	23	52	79
	%	EE.UU.	83.4	4.7	11.9	100.0
		Europa	12.3	49.3	38.4	100.0
		Japón	5.1	29.1	65.8	100.0

a. Clasificados correctamente el 73.8% de los casos agrupados originales.