



**UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO**

FACULTAD DE CIENCIAS

**APLICACIONES DE ANÁLISIS DISCRIMINANTE MEDIANTE
CASANDRA**

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

Adriana Ramírez González



**DIRECTOR DE TESIS:
Francisco Sánchez Villarreal
2009**



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

AGRADECIMIENTOS

Agradezco a la Universidad Nacional Autónoma de México por haber formado parte importante de mi desarrollo humano y profesional, a todos mis profesores por enseñarme todo lo que ahora sé, en particular a Margarita Chávez Cano quien motivó en gran medida a la elección del tema de esta tesis y a mi profesor y director Francisco Sánchez Villareal por orientarme, motivarme y darme herramientas para la realización así como también por el interés puesto en mi tanto como alumna como de tesista.

A mi familia y a mis padres por su cariño, comprensión por haberme inculcado la perseverancia, ya que con sus consejos y apoyo.

Agradezco también a mis amigas Ofelia Martínez y Cesar Carreón ya que me acompañaron durante gran parte de la carrera y fueron un excelente equipo de estudio y un gran apoyo emocional que nunca olvidaré.

INDICE

INTRODUCCIÓN.....	6
CAPITULO I : Análisis Discriminante.....	9
CAPITULO I I: Uso de CASANDRA.....	19
CAPITULO III:DISCRIMINANTE PARA TUMORES CANCEROSOS (CASANDRA Y SPSS) Y ASIGNACIÓN DE CRÉDITO.....	37
III.1 TUMOR: MALIGNO VS BENIGNO.....	37
III.1.1 ANALISIS PARA EL CASO REDUCIDO A 7 VARIABLES.....	45
. III.2 COMPARATIVO CON SPSS.....	51
. III.3 ANALISIS DISCRIMINANTE CON CASANDRA USASNDNO LAS VARIABLES QUE SPSS CONSIDERA.....	63
. III.4 ANÁLISIS PARA ASIGNACIÓN DE CRÉDITO.....	64
CONCLUSIONES.....	71
ANEXO 1 : Análisis de Conglomerados.....	72
ANEXO 2 TABLAS.....	76
Bibliografía.....	80

INTRODUCCIÓN.

CAPITULO I : Análisis Discriminante.

CAPITULO II : Uso de CASANDRA.

CAPITULO III: DISCRIMINANTE PARA TUMORESCANCEROSOS (CASANDRA Y SPSS) Y ASIGNACIÓN DE CRÉDITO.

III.1 TUMOR: MALIGNO VS BENIGNO.

III.1.1 ANALISIS PARA EL CASO REDUCIDO A 7 VARIABLES.

III.2 COMPARATIVO CON SPSS.

III.3 ANALISIS DISCRIMINANTE CON CASANDRA USASNDO LAS VARIABLES QUE SPSS CONSIDERA..

III.4 ANÁLISIS PARA ASIGNACIÓN DE CRÉDITO

CONCLUSIONES.

ANEXO 1 : Análisis de Conglomerados.

ANEXO 2 TABLAS.

Bibliografía.

Aplicaciones de Análisis discriminante mediante
CASANDRA.

INTRODUCCIÓN

En distintos ámbitos de la vida resulta útil contar con algún tipo de clasificación, esta clasificación en algunos casos resulta obvia pero en otros casos se puede tener dificultades para poder construir una representación exacta. Los elementos que se usan para una clasificación deben estar bien definidos y ordenados para que de esta manera se pueda tener un manejo objetivo. Esta clasificación se hace más útil en el campo científico cuando por ejemplo se busca la clasificación de elementos de una tabla periódica, taxonomías de especies vegetales o animales, en la apertura de un crédito bancario o inclusive en la clasificación de enfermedades en el campo de la medicina.

La discriminación es una forma de clasificar que supone la existencia de dos o más poblaciones o grupos, en este sentido la idea es tener reglas que permitan colocar a un individuo (elemento) en algún grupo.

Otra forma de clasificar es agrupar elementos de una población de manera que los grupos sean “suficientemente” diferentes.

Con la discriminación se podría clasificar a una persona como merecedora de un crédito financiero. Predecir si una persona es buena pagadora o no mediante alguna “técnica” tiene un riesgo, porque si el individuo es moroso y se le otorga algún crédito el banco habrá perdido el crédito usado por el cliente, los gastos de cobranza y gastos de administración. Pero si no otorga el crédito el riesgo puede ser perder el cliente, pero esto no quiere decir que sea menor el “costo” de error.

El análisis de conglomerados (cluster) es una técnica estadística multivariante de clasificación que usa una serie de variables para crear clases homogéneas y bien separadas entre sí, es decir se forman grupos en los que los elementos de un grupo tienen características o atributos semejantes. El análisis discriminante por otro lado presupone la existencia de grupos y se pretende incluir nuevos elementos en uno u otro grupo, esto se logra con ciertas reglas que contemplan otras que son sistemáticas y estadísticas, como por ejemplo que los elementos tengan cierta separación o distancia.

Las primeras nociones de distancia entre grupo es la que realiza Karl Pearson, el cual propuso el “coeficiente de parecido racial”. En la India Mahalanobis formula otra idea de distancia entre grupos que llevaría su nombre “Distancia de Mahalanobis”. Fisher uso este concepto de distancia para crear una combinación lineal de variables para discriminar entre grupos.

El análisis Discriminante se ha aplicado a múltiples campos de la actividad científica pero se ha ido modificando hasta lo que actualmente se maneja. Es

importante la función lineal Discriminante ya que interpreta los efectos observados a través de un Análisis Multivariante de la Varianza (MANOVA).

El problema básico del análisis discriminante es determinar las variables que mejor contribuyen a discriminar entre los grupos y de esta manera poder colocar un individuo en algún grupo, esta inclusión de determinado individuo tendrá asociado cierto error. El Análisis Discriminante también permite reducir el número de variables con la finalidad de poder explicar las diferencias fundamentales entre los grupos. Este análisis tiene básicamente dos propósitos.

- Describir las diferencias entre grupos.
- Predecir la pertenencia a los grupos.

Para la realización de la presente tesis se comenzó con dos problemas:

El caso de individuos que presentan tumores y que han sido medidas con un cierto número de variables. Estas variables tienen la característica de ser cuantitativas. Los tumores han sido clasificados previamente como:

- Malignos
- Benignos

Las preguntas que se tienen en este caso son:

- a) Dada la medición de un “nuevo” individuo, ¿En qué grupo estará asignado?
- b) ¿Con que probabilidad esta dicho individuo en el grupo asignado? y
- c) ¿Qué “costo” se tiene al estar “bien” o “mal” clasificado?

El individuo asignado a un grupo con una probabilidad baja de pertenencia da lugar a un “costo” el cual es traducido a que si el individuo fué clasificado con tumor maligno teniendo uno benigno probablemente se le estudiará nuevamente y con más variables, esto con la finalidad de verificar lo predicho, sin embargo esto solo tendrá un “costo” monetario al haber realizado un mayor número de análisis. Caso contrario ocurre cuando se asigna un individuo al grupo de los de tumor benigno cuando en realidad tiene uno maligno, este tipo de “error” desencadenaría el hecho de no tomar medidas para controlar el cancer , minimizarlo o erradicarlo.

El caso de individuos que acuden a un banco para solicitar un crédito, se han clasificado previamente en dos grupos:

- Crédito asignado.
- Crédito denegado.

Las preguntas que se tienen son:

- a) Dada la información de las variables de un individuo “nuevo” ¿será un individuo al que se el puede otorgar crédito o será un individuo al que no se le debe otorgar crédito?
- b) ¿Con qué probabilidad esta dicho individuo en el grupo asignado? y
- c) ¿Qué “costo” se tiene al estar “bien” o “mal” clasificado?

Si el individuo fué clasificado como “buen pagador” cuando en realidad pertenece al grupo de los que no se les debe asignar línea de crédito, se traduce como un “costo” (error de asignacion) el cual se verá reflejado cuando el individuo caiga en incumplimientos de pago, es decir , el costo en este sentido irá fuertemente relacionado con la línea de crédito asignada o bien el porcentaje de ella que se utilice.

Por otro lado si la persona fué clasificada en el grupo de los que se les niega crédito cuando en realidad pertenece al grupo de los que se les debe asignar cierta línea de crédito, querrá decir que el “costo” en este caso podría ser un tanto menor, (depende la perspectiva que se tome) ya que se puede perder un cliente y todo lo que esto implica.

Las variables también son cuantitativas en todos los ejemplos aquí utilizados.

Tanto en el ejemplo de tumor como en el de crédito se tiene cierta semejanza y es que en ambos casos se tienen grupos ya establecidos.

Al tener estos problemas de clasificación con las características anteriores, se puede pensar en utilizar un método llamado:

Análisis Discriminante

De manera paralela se discutirá si los grupos están bien clasificados (asignados) y para esto se hablará brevemente del método de conglomerados, que además nos puede auxiliar para minimizar el número de variables ya que como se verá mas adelante para el caso del tumor de pecho se cuentan con demasiadas variables donde muchas de ellas resultan redundantes y al eliminarlas se puede conservar prácticamente el mismo poder de clasificación

CAPITULO I: Análisis Discriminante

ANÁLISIS DISCRIMINANTE

Se comenzará este estudio considerando el caso de dos grupos a los que llamaremos I y II y una sola variable clasificadora llamada X. El objetivo de este análisis es clasificar a cada individuo en el grupo correcto. Se hace el siguiente supuesto; la distribución y varianza de los grupos es la misma, y solo se diferencian en la media (ya que de no ser así no tendría sentido hacer un análisis discriminante). Con los supuestos anteriores, se cuenta con la gráfica de las dos distribuciones; las cuales se solapan, y este solapamiento da lugar a errores de clasificación. Para este análisis se considera que hay solapamiento, ya que de no ser así, nuestro análisis sería trivial. El análisis discriminante busca minimizar el riesgo de clasificación incorrecta al buscar la forma de diferenciar óptimamente los grupos.

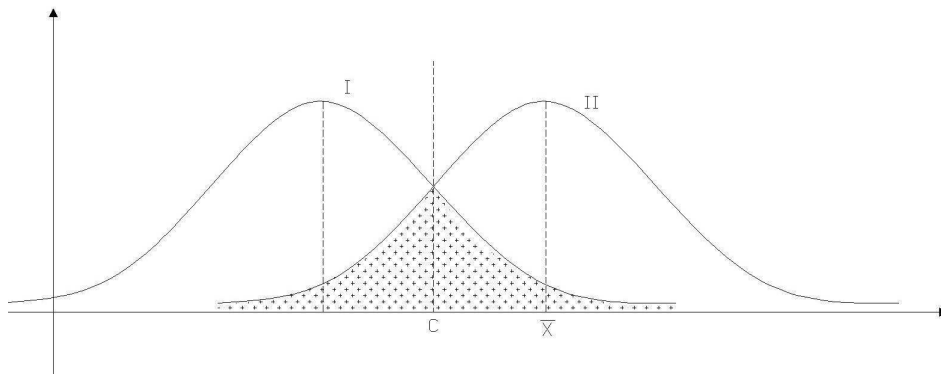


figura 1

Definamos lo siguiente:

X_I : la media del grupo I

X_{II} : la media del grupo II

$C = \frac{X_I + X_{II}}{2}$: Punto de intersección de las funciones correspondientes a los grupos I y II

Al querer decidir a que población pertenece un individuo, se cuenta con el siguiente criterio de clasificación.

Si $X_i < C$ i está en I

Si $X_i > C$ i está en II

Al escoger a la variable clasificadora de la mejor manera se cometerán un menor número de errores. Esta variable describirá "bien" a las poblaciones, pero, en realidad este escenario es utópico, ya que casi nunca se tiene una variable que describa al cien por ciento el comportamiento de una población. Un ejemplo de esto es cuando se tienen dos grupos uno de hombres y el otro de mujeres, y se considera a la variable "peso", esta puede diferenciar el grupo de hombres con el de mujeres, pero definitivamente se podría caer en muchos errores de asignación, ya que se puede tener un hombre "suficientemente" delgado, tanto que al tomar una sola variable (peso), este caiga fácilmente en el grupo de las mujeres.

Por lo anterior, para obtener una mejor clasificación, y para dar una representación gráfica, se incluirá una "segunda variable de clasificación".

Esta gráfica muestra dos elipses que tienen el mismo tamaño, pero difieren en su centro. Debajo del eje X_1 se ha representado la proyección de las distribuciones univariantes marginales de la variable X_1 , análogamente con X_2 , en ambos casos se tendrá un "alto" grado de solapamiento, pero se puede minimizar esto obteniendo una mejor "función discriminante" usando las dos variables conjuntamente. Si se dibuja una línea recta a través de los puntos donde las elipses se intersectan y luego proyectar la línea sobre un nuevo eje Z, se puede decir que el solapamiento entre las distribuciones univariadas de las dos poblaciones I y II (representadas por el área de cada una de las elipses) es menor que la que podemos obtener por cualquier línea a través de las elipses.

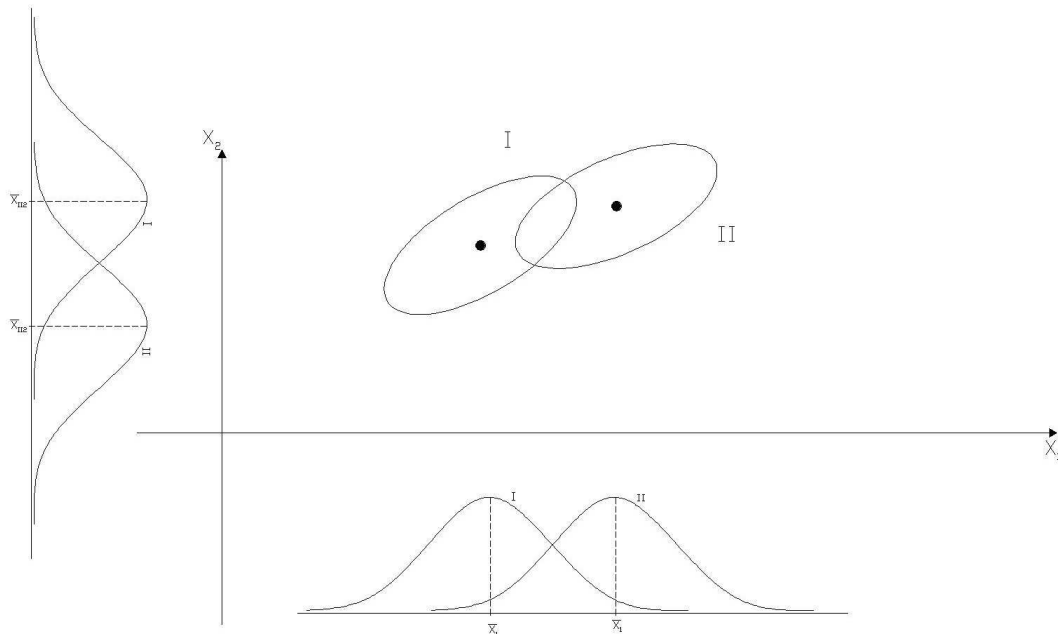


figura 2

Para encontrar una combinación lineal de las variables originales X_1 y X_2 se puede proyectar el resultado como una función discriminante.

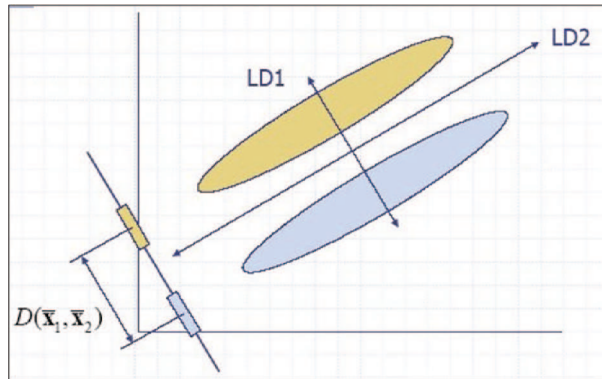


figura 3

En resumen para un problema de análisis discriminante, una combinación de variables independientes es obtenida, resultando en una serie de puntajes discriminantes para cada individuo en cada grupo.

Los puntajes discriminantes son colocados de acuerdo a la regla estadística de maximizar la varianza entre los grupos y minimizar la varianza dentro de ellos. Si la varianza entre grupos es relativamente grande con respecto a la varianza dentro de cada grupo se dice que la función discriminante separa bien a los grupos.

El análisis discriminante involucra la combinación lineal de una o más variables independientes que discriminarán mejor entre grupos definidos a priori, ¹ es decir, las funciones discriminantes se construyen como combinaciones lineales de las variables independientes de tal modo que dan lugar a la máxima separación posible entre grupos al mismo tiempo que no están correlacionados entre sí. En el caso general esta función involucra k variables "explicativas" y G grupos. El número de funciones discriminantes que se pueden obtener son q donde:

$$q = \min(k, G - 1)$$

Sea

$$D = u_1 X_1 + u_2 X_2 + \dots + u_k X_k$$

La función discriminante donde:

X_m = m -ésima variable explicativa

u_m = m -ésimo coeficiente de ponderación.

Supongamos además que se tienen

$$n = \sum_{i=1}^G n_i$$

Observaciones donde:

n_l = número de individuos perteneciente al grupo l

¹La combinación lineal para el análisis discriminante, también se conoce como función discriminante de Fisher

Sea

$$D_i = u_1 X_{1i} + u_2 X_{2i} + \dots + u_k X_{ki}$$

$i = 1, 2, \dots, n$ observaciones o individuos.

La puntuación discriminante correspondiente a la observación i -ésima es decir el valor que toma la i -ésima observación en la combinación lineal, donde:

X_{pi} = valor que toma la i -ésima observación en la p -ésima variable $p = 1, \dots, k$

La idea principal del análisis discriminante es obtener una serie de funciones lineales a partir de las variables independientes que permitan interpretar las diferencias entre los grupos y clasificar a los individuos en algún grupo definido por la variable dependiente.

La función discriminante puede expresarse mediante un producto escalar de vectores:

$$\begin{aligned} D &= u_1 X_1 + u_2 X_2 + \dots + u_k X_k \\ &= \begin{bmatrix} u_1 & u_2 & \dots & u_k \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \end{aligned}$$

De tal forma que todas las puntuaciones discriminantes quedan expresadas de la siguiente manera.

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{k1} \\ X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

Si se pretende determinar las diferencias entre medias de los grupos en la variable D podríamos recurrir a la razón F que se utiliza en el análisis de la varianza, es decir:

$$F = \frac{\text{suma de cuadrados y productos totales entre grupos}}{\text{suma de cuadrados y productos totales dentro de grupos}}$$

$$\begin{aligned} F &= \frac{SC_{\text{entre grupos}} / (g-1)}{SC_{\text{dentro de grupos}} / (n-g)} \\ &= \frac{SC_{\text{entre grupos}}}{SC_{\text{dentro de grupos}}} \frac{(n-g)}{(g-1)} \end{aligned}$$

Para analizar este cociente, definamos primeramente a la matriz² T

$$T = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \cdots & \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - \bar{X}_i)(X_{jkm} - \bar{X}_j) & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$T_{k \times k}$ da información acerca de la covariación entre cada pareja de variables

² X_{jkm} indica variable j individuo k grupo m

$$T = \begin{bmatrix} t_{11} & & & \\ & \vdots & & \\ & & t_{ij} & \\ & & \vdots & \\ & & & & t_{kk} \end{bmatrix}$$

Como t_{ij} se obtiene de la suma de productos entre las desviaciones que presentan las puntuaciones de un individuo en las variables i y j respecto a las medias alcanzadas de dichas variables en el grupo global de individuos,

$$\frac{t_{ij}}{n-1}$$

será la covarianza entre las dos variables, notese que si $i = j$

$$\frac{t_{ii}}{n-1}$$

será la varianza para la variable i

T da entonces informacion de la variabilidad total que presentan las k variables independientes.

Análogamente se puede definir una matriz W

$$W = \begin{bmatrix} & & \vdots & & \\ \cdots & \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - \bar{X}_{ik})(X_{jkm} - \bar{X}_{jk}) & \cdots & & \\ & & \vdots & & \\ & & & & \end{bmatrix}$$

$$= \begin{bmatrix} & & \vdots & & \\ \cdots & w_{ij} & \cdots & & \\ & & \vdots & & \end{bmatrix}$$

Como W tiene informacion de la variabilidad en el interior de los grupos (intra-grupos, dentro del grupo), T puede descomponerse en la variabilidad dentro de los grupos y la variabilidad entre los grupos siendo esta última variabilidad expresada con la matriz B

$$T = B + W$$

$B_{k \times k}$ entonces será la matriz de sumas de cuadrados y productos cruzados dentro de los grupos

$$\therefore B = T - W$$

Con las matrices B y W se pueden expresar las sumas de cuadrados dentro de grupos y entre grupos para la variable D_i

$$SC_{dentro\ de\ grupos} = u' B u$$

$$SC_{entre\ de\ grupos} = u' W u$$

$$\text{con } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

$$\frac{u' Bu}{u' W u}$$

Los valores de T , F y W ya los podemos obtener con los datos muestrales pero aún faltan los coeficientes u_i . Para la estimación de u_i Fisher utilizó lo siguiente:

$$\text{max } \lambda = (u' Bu)/(u' W u)$$

Se usó un cociente porque al maximizarlo se estaría diciendo que la variabilidad de F es mayor que la variabilidad en W . Dicho en otras palabras, la variabilidad entre los grupos es "muy grande" respecto a la variabilidad dentro de los grupos (si este fuera "muy grande" los grupos no estarían bien definidos, tanto que podrían empalmarse con los de otros grupos).

Con este criterio se trata de determinar el eje discriminante de forma que las distribuciones proyectadas sobre el mismo estén lo más separadas posibles entre sí (mayor variabilidad entre grupos) y, al mismo tiempo, cada una de las distribuciones (por separado) estén lo menos dispersas (menor variabilidad dentro de los grupos).

$$\text{max } \lambda = \text{max} \left[\frac{U' B U}{U' W U} \right]$$

Tomando en cuenta que λ es un escalar que podemos tomar como criterio para medir la discriminación de grados a lo largo de la dimensión especificada por el vector U , el objetivo será encontrar los coeficientes u_1, u_2, \dots, u_k que maximicen el criterio de discriminación λ . Por lo tanto hay que calcular la derivada parcial de λ respecto a cada componente del vector U , e igualar a cero.

$$\frac{\delta \lambda}{\delta U} = \frac{2[B U (U' W U) - (U' B U) (W U)]}{(U' W U)^2}$$

Dividiendo el numerador y denominador por $U' W U$ se obtiene

$$\frac{\delta \lambda}{\delta U} = \frac{2 \left[\frac{B U (U' W U) - (U' B U) (W U)}{U' W U} \right]}{\frac{(U' W U)^2}{U' W U}}$$

$$\frac{\delta \lambda}{\delta U} = \frac{2 \left[B U * \frac{(U' W U)}{U' W U} - \frac{(U' B U)}{U' W U} * (W U) \right]}{\frac{(U' W U)^2}{U' W U}}$$

$$\frac{\delta \lambda}{\delta U} = \frac{2[B U - \lambda * (W U)]}{U' W U}$$

$$\frac{2(B U - \lambda W U)}{U' W U} = 0$$

$$B U - \lambda W U = 0$$

\Rightarrow

$$(B - \lambda W) U = 0$$

Suponiendo que W no es una matriz singular, y que $|W| \neq 0$ (determinante distinto de cero), es posible calcular la matriz inversa W . Multiplicando por la izquierda por W^{-1} ambos miembros de la igualdad, se obtiene lo siguiente.

$$W^{-1}(B - \lambda W)U = 0$$

$$(W^{-1}B - \lambda I)U = 0$$

$$W^{-1}BU = \lambda IU$$

Por lo tanto para encontrar el vector U es necesario encontrar el valor característico asociado $W^{-1}B$ (la cual debe ser una matriz simétrica).

Cuando ya se tengan todos los valores característicos y por lo tanto todos los vectores asociados a cada valor característico, se elije el valor característico más grande junto con su respectivo vector para formar la primer función discriminante.

Sea λ_1 el valor característico más grande.

La primera función discriminante, como se dijo anteriormente será la función que maximizará más la variabilidad entre los grupos y que al mismo tiempo minimizará la variabilidad dentro de los grupos será entonces nuestro primer vector propio asociado al valor característico más grande, es decir λ_1 , son u_1 se forma:

$$D_2 = u_2X$$

La siguiente función discriminante será no correlacionada con la primera y tendrá las mismas características de variabilidad entre grupos y dentro de grupos, además se formará con el vector propio asociado a el segundo más grande valor propio , esto es:

$$D_2 = u_2X$$

Las siguientes funciones no deberán estar correlacionadas con las anteriores, y se irán tomando en el orden en que se fueron ordenando los valores propios, es decir, en estricto orden decreciente y respetando la no correlación de unas con otras.

De esta manera se puede estar seguro que se obtendrán q autovalores no nulos y q autovectores asociados, es decir, q funciones discriminantes.

Con lo anterior que resuelto el problema de discriminación, pero no se puede dar una interpretación directa de los coeficiente debido a que las soluciones no se han calculado bajo ninguna restricción relativa al origen y la métrica del espacio discriminante, es decir, no se han estandarizado los coeficientes. Los coeficientes estandarizados tienen la siguiente forma:

$$u_i = v_i \sqrt{(n - g)}$$

$$u_0 = \sum_{i=1}^p v_i \bar{x}_i$$

Gracias a esta transformación, se puede trasladar el origen de cada eje discriminante para hacerlo coincidir con el centroide global. De esta manera los valores de la función discriminante para los casos tendrá una media cero y una desviación típica igual a uno.

Retomando lo ya dicho, el número de funciones que se va a obtener es $q = \min(k, G - 1)$, pero esto no quiere decir que necesariamente se deban tomar todas las q funciones. Es posible que baste considerar un número menor a q , por ejemplo $\alpha = q - \gamma$ con $q < \gamma$ de tal manera que existan α funciones discriminantes "significativas". Las γ funciones restantes tendrán una fuerte capacidad discriminante. Estas α funciones tendrán varianzas uno y son incorrelacionadas entre si, es decir que $u_i W u_j = \delta_{ij}$, $j = 1, \dots, r$ se obtienen como soluciones de r

vectores propios de $W^{-1}F$ asociados a los r mayores valores propios de esta matriz $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ (los valores propios miden el poder discriminante de la i -ésima función discriminante de forma que si $\lambda_i \approx 0$ la función discriminante no tendrá ningún poder discriminante). A las funciones $D_i = u_i Y$ se les llama funciones discriminantes canónicas o de Fisher.

Existe un método que ayuda a elegir las "mejores" funciones discriminantes y a eliminar aquellas funciones que no aportan suficiente información. Este procedimiento inferencial que sirve para determinar la significación de la función discriminante se llama Lambda de Wilks (Λ), con esto se puede determinar la probabilidad de que se obtenga a partir de una muestra un grado de discriminación similar al observado, bajo la hipótesis nula de que no hay diferencias entre los grupos debidas a varias funciones discriminantes, toma en cuenta las diferencias entre grupos debidas a varias funciones discriminantes dentro de los grupos respecto a la desviación total sin considerar grupos.

Al calcular la lambda de Wilks para la primera función, se mide la discriminación que las variables permiten entre los grupos y se decide si esta discriminación es significativa. En caso de serlo, se extrae la primera función y se examina la discriminación residual que permanece en el sistema, con la finalidad de extraer una segunda función discriminante. Este proceso se repite hasta encontrar que la información restante acerca de las diferencias entre grupos no es significativa y por lo tanto no sería necesario encontrar una nueva función discriminante.

La Lambda de Wilks para diferencias multivariadas tiene la siguiente expresión:

$$\Lambda = \frac{|T|}{|W|}$$

Este estadístico permite comprobar la significación de las diferencias entre los centroides de K grupos. Al tomar la inversa se obtiene:

$$\begin{aligned} \frac{1}{\Lambda} &= \frac{|W|}{|T|} = |W^{-1}T| \\ \Rightarrow |W^{-1}(W+B)| &= |I+W^{-1}B| \end{aligned}$$

Con lo cual se obtiene que el determinante de esta matriz es:

$$(1 + \lambda) * (1 + \lambda_2) * \dots * (1 + \lambda_q)$$

$$\Lambda = \left[\frac{1}{1+\lambda_1} \right] \left[\frac{1}{1+\lambda_2} \right] \dots \left[\frac{1}{1+\lambda_q} \right]$$

Con esta nueva expresión para Λ , se puede ver la relación que hay con los valores propios de la función discriminante. Debido a que Λ fue calculada como una medida inversa, cuando los valores de Lambda se aproximan a 0 significará una alta discriminación, mientras que valores próximos a 1 indicarán escasa discriminación.

Si se excluye la discriminación debida a la primera función, el valor de lambda será:

$$\Lambda = \left[\frac{1}{1+\lambda_2} \right] \left[\frac{1}{1+\lambda_3} \right] \dots \left[\frac{1}{1+\lambda_q} \right]$$

Sin pérdida de generalidad, al haber extraído las primeras r funciones discriminantes, el valor de Lambda será:

$$\Lambda = \prod_{q=r+1}^{\min(p,g-1)} \frac{1}{(1 + \lambda_q)}$$

Cuando al llegar a la discriminación $r+1$ se encuentra que ya no es significativo, entonces se concluye que se habrán de tomar las r primeras funciones discriminantes, las cuales serán las que mejor explican las diferencias entre grupos. Es decir, serán necesarios solamente r dimensiones para representar las diferencias entre grupos.

Este estadístico tiene una distribución Lambda de Wilks con $p, g - 1$ y $n - g$ grados de libertad.

Observación:

$$\lambda_i = \sum n_g (\bar{d}_g^i - \bar{d}_g)^2 \quad i =, \dots, q$$

en donde d_g^i con $g=1, \dots, G$ son la puntuaciones medias de las i -ésima función discriminante de los G grados. d_g es la puntuación media total.

El número de funciones significativas se determina considerando el estadístico propuesto por Bartlett que consiste en usar un contraste de hipótesis secuencial, El proceso comienza con $i = 0$. En el paso $i = r + 1$ del algoritmo la hipótesis nula a contrastar es:

$$H_0 : \lambda_{r+1} = \dots = \lambda_{\min(p, g-1)} = 0$$

y el estadístico de contraste viene dado por:

$$(n - 1 - \frac{p+g}{2}) \sum_{j=r+1}^{\min(p, g-1)} \ln(1 + \lambda_j)$$

el cual tiene una distribución χ^2 con $(p-r)(g-r-1)$ grados de libertad si H_0 es verdad.

El p -valor asociado al contraste es:

$$P \left[X_{(p-r)(g-r-1)}^2 \geq T_{obs} \right]$$

T_{obs} es el valor observado de T . El contraste para el primer valor de r para el cual la hipótesis nula H_0 se acepta.

Cuando ya se tienen las funciones discriminantes, el siguiente paso es analizar cuales son las variables a considerar. Se debe tomar en cuenta que para esto se analizan las variables que ya han sido estandarizadas. Las condiciones para la selección de las variables se basan en la tolerancia de las variables y en las estadísticas multivariantes parciales F con las cuales se garantiza que el incremento de discriminación debido a al variable supera un nivel fijo.

Las variables que se tomarán en cuenta no tienen que ser correlacionadas linealmente ya que de ser así se tendrá redundancia. Para evitar considerar variables que esten correlacionadas se usa la tolerancia. La tolerancia de una variable no seleccionada es $1 - R$ donde R es la correlación multiplicada entre esa variable y todas las variables ya incluidas, cuando han sido obtenidas a partir de la matriz de correlación intragrupos. Si R tiende a 1 querrá decir que existe un alto grado de correlación entre esa variable y las otras, en otras palabras, la variable que se pretendería incluir es combinación lineal de una o más variables de las ya incluidas, por lo tanto si $1 - R$ tiende a 0 no se deberá incluir la variable en la función discriminante.

Además de ser no correlacionadas las variables de la función Discriminante se debe tomar en cuenta dos estadísticas más: el estadístico F de entrada y el estadístico F de salida que servirá para comprobar que todas las variables seleccionadas son adecuadas y que en el proceso de ir introduciendo variables aporten la misma contribución a la separación de los grupos.

El cálculo de F (de entrada) representa el incremento producido en la incorporación de una variable respecto al total de discriminación alcanzado por las variables ya introducidas es, es por esto que una F pequeña implica el no seleccionar la variable. El cálculo de F cuando ya se tienen S variables seleccionadas es:

$$F = \left[\frac{n - g - s}{g - 1} \right] \left[\frac{1 - \Lambda_{s+1}/\Lambda_s}{\Lambda_{s+1}/\Lambda_s} \right]$$

Λ_s =valor de Lambda de Wilks antes de añadir la variable

Λ_{s+1} =Lambda de Wilks incluyendo la nueva variable.

Este estadístico se distribuye según F con $(g - 1)$ y $(n - s - g + 1)$ grados de libertad y ayuda para determinar la significación producida en la discriminación, lo anterior es más confiable cuando las poblaciones no son "tan pequeñas".

Los centroides de cada grupo ayudan a entender el comportamiento de los grupos. Las puntuaciones discriminantes de los centroides son calculadas sustituyendo en las variables por sus valores medios. Para obtener una buena interpretación es conveniente considerar a la función discriminante que tenga mayor capacidad para separar grupos.

Para clasificar a los individuos en un grupo u otro se puede utilizar el criterio de Bayes. Este método contempla el hecho de tener información a priori de la probabilidad de pertenencia a un grupo determinado, pero en caso de que no se cuente con esa información se puede considerar que la probabilidad de pertenencia al grupo i (\prod_i) es .5. Considerando el caso general de G grupos, el teorema de Bayes dice que la probabilidad de pertenencia a un grupo. dado un puntaje discriminante es igual a:

$$P[g | D] = \frac{\prod_i * P[D | g]}{\sum \prod_i * P[D | i]}$$

Cuando se comparan todas las probabilidades de pertenencia, es claro suponer que el individuo se asignará al grupo para el cual haya presentado la probabilidad más alta.

CAPITULO II :Uso de Casandra

Antes de comenzar con el análisis discriminante de Tumor de Pecho o con el de la clasificación de los individuos a los que se les asignarán cierta línea de crédito, en este capítulo se presentará un ejemplo que tiene dos grupos en la variable sexo, dos variables y 27 elementos ,algunos de los cuales están en el grupo 0 (mujeres) y otros en el grupo 1(hombres), las variables predictoras corresponden a mediciones antropométricas:estatura, peso, longitud del pie, longitud del brazo, longitud de la espalda en hombros, circunferencia del craneo y longitud de la pierna. El propósito es ejemplificar tanto el uso de CASANDRA como el Análisis Discriminante.

obs	sexo	estatura	peso	pie	brazo	espalda	cráneo	pierna
20	0	152	45	34	66	40	55	38
19	0	156	52	36	67	36	56	41
1	0	159	49	36	68	42	57	40
23	0	155	53	36	67	43	56	38
10	0	158	50	36	68.5	44	57	41
13	0	158	43	36	68	43	55	39
11	0	156	65	36	68	46	58	41
27	0	168	56	37.5	70.5	48	60	40
4	0	167	52	37	73	41.5	58	44
5	0	164	51	36	71	44.5	54	40
25	0	170	70	38	73	45	56	43
3	0	172	65	38	75	48	58	44
6	0	161	67	38	71	44	56	42
18	0	162	68	39	72	44	59	42
7	0	168	48	39	72.5	41	54.5	43
2	1	164	62	39	73	44	55	44
9	1	183	74	41	79	47.5	59.5	47
26	1	170	67	40	77	46.5	58	44.5
22	1	173	69	41	74	48	56	44
14	1	178	74	42	75	50	59	45
16	1	182	91	41	83	53	59	43
12	1	173	64	40	79	48	56.5	47
8	1	181	74	43	74	50	60	47

figura 4

Una vez abierto el “Sistema Estadístico CASANDRA” el primer paso es abrir alguna de las tablas para así comenzar el “Análisis”, esta tabla puede ser obtenida de las tablas de ejemplos que ya trae CASANDRA; o bien, la tabla puede importarse de Excel o archivos pdf. Cuando la tabla esté en la pantalla de inicio hay que verificar que los grupos estén definidos en alguna columna, ya que posteriormente CASANDRA preguntará por la variable criterio , la cual tendrá que ser de naturaleza cualitativa.



figura 5

Para empezar a hacer un “Análisis de Clasificación” (análisis discriminante) es necesario dirigirse a menú-Inferencia-Análisis Multivariado-Análisis de Discriminante clásico y dar un “clic” para que despliegue una pantalla en la que se debe indicar las variables a usar.

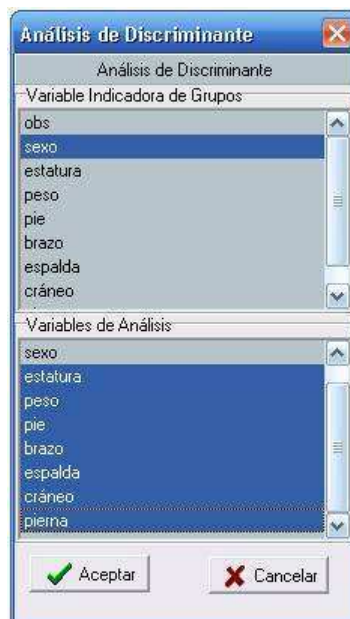
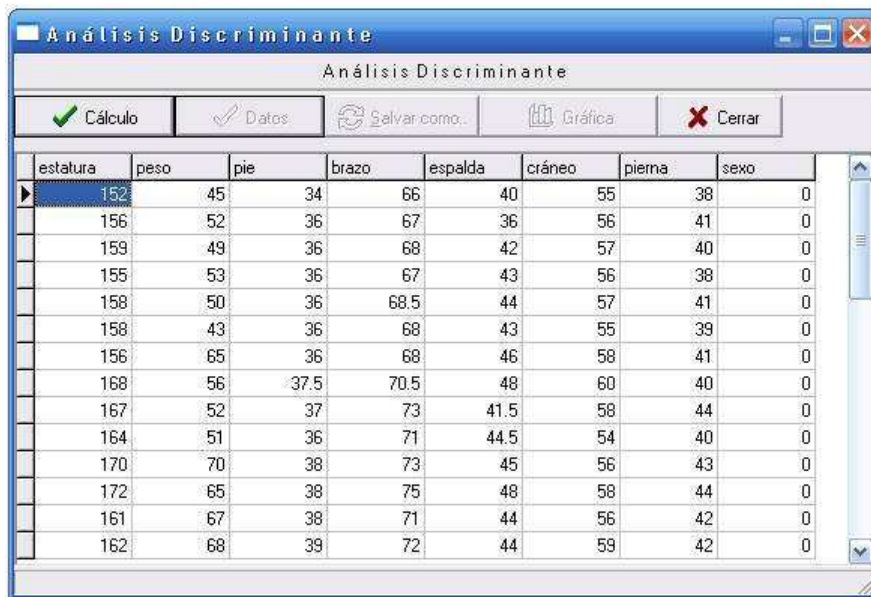


figura 6

Dentro de la ventana “Análisis de Discriminante” se selecciona la variable indicadora de Grupos, es decir la variable que identifica la pertenencia a algún grupo (la cual es de naturaleza cualitativa). Además es necesario seleccionar las variables de análisis, las cuales son llamadas también variables predictoras. Una vez hecho esto se selecciona el boton de “Aceptar” , con lo cual se abre la ventana siguiente.



Análisis Discriminante

✓ Cálculo ✓ Datos Salvar como... Gráfica ✗ Cerrar

estatura	peso	pie	brazo	espalda	cráneo	pierna	sexo
152	45	34	66	40	55	38	0
156	52	36	67	36	56	41	0
159	49	36	68	42	57	40	0
155	53	36	67	43	56	38	0
158	50	36	68.5	44	57	41	0
158	43	36	68	43	55	39	0
156	65	36	68	46	58	41	0
168	56	37.5	70.5	48	60	40	0
167	52	37	73	41.5	58	44	0
164	51	36	71	44.5	54	40	0
170	70	38	73	45	56	43	0
172	65	38	75	48	58	44	0
161	67	38	71	44	56	42	0
162	68	39	72	44	59	42	0

figura 7

En esta pantalla aparecen primero las variables predictoras; que en este caso son dos y al final CASANDRA coloca a la variable criterio. Al dar “clic” en “Cálculo” comenzará a hacer el cómputo de los datos, y aparecerá la siguiente ventana :



Análisis Discriminante

✓ Cálculo ✓ Datos Salvar como... Gráfica ✗ Cerrar

Análisis Discriminante

Manova de Grupos

Estadísticas Básicas

Variable	No. Observaciones	Media Aritmética	V
estatura	27	168.777778	1
peso	27	63.888889	1
pie	27	38.981481	8

figura 8

Este Sistema Estadístico tiene la particularidad de que todo este cálculo lo puede exportar hacia Excel. Las siguientes tablas muestran lo exportado a Excel.

En la siguiente tabla se presentan las estadísticas básicas de las variables predictoras para el total de observaciones. A continuación se emite el reporte de estadísticas básicas definido por los grupos que definen las variables indicadoras.

En la siguiente tabla se hace un análisis tanto de la media como de la varianza Total, esto quiere decir que se toma en cuenta todas las observaciones.

Estadísticas Básicas				
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muest. n-1
estatura	27	168.78	103.95	10.20
peso	27	63.89	163.87	12.80
pie	27	38.98	8.20	2.86
brazo	27	73.46	24.58	4.96
espalda	27	45.85	16.17	4.02
cráneo	27	57.24	3.39	1.84
pierna	27	43.09	9.96	3.16

figura 9

En las siguientes dos tablas se contemplan por separado a los grupos, al grupo que tiene como variable indicadora “0” y a el grupo que tiene la variable indicadora igual a “1”, en cada una de estas se hace el mismo análisis que en la tabla de Estadísticas Básicas, pero tomando en cuenta a los individuos involucrados en los grupos “0” y “1”.

Estadísticas Básicas				
Variable Indicadora	0			
Casos	15			
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muest. n-1
estatura	15	161.73	37.64	6.13
peso	15	55.60	80.40	8.97
pie	15	36.83	1.92	1.38
brazo	15	70.03	7.41	2.72
espalda	15	43.33	9.42	3.07
cráneo	15	56.63	2.95	1.72
pierna	15	41.07	3.78	1.94

figura 10

Las estadísticas básicas de ambos grupos presentan una idea de el tipo del perfil de cada grupo. Por ejemplo que las mujeres son más bajas en promedio que los hombres, 161.73 cm contra 177.58 de los hombres, lo cual indica que existe mayor variabilidad en el peso de la personas que en su estatura. En los dos grupos hay alta varianza muestral y desviación estandar para la variable peso.

Estadísticas Básicas				
Variable Indicada	1			
Casos	12			
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1
estatura	12	177.58	45.54	6.75
peso	12	74.25	74.20	8.61
pie	12	41.67	2.79	1.67
brazo	12	77.75	12.57	3.55
espalda	12	49.00	6.77	2.60
cráneo	12	58.00	3.14	1.77
pierna	12	45.63	6.14	2.48

figura 11

La matriz (B) es la suma de cuadrados y productos cruzados de las desviaciones entre la media de cada grupo y la media global.

Matriz de Suma de Cuadrados Entre Grupos (B)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	1,674.82	1,970.68	510.72	815.39	598.78	144.41	481.66
peso	1,970.68	2,318.82	600.94	959.44	704.56	169.92	566.75
pie	510.72	600.94	155.74	248.65	182.59	44.04	146.88
brazo	815.39	959.44	248.65	396.98	291.52	70.31	234.50
espalda	598.78	704.56	182.59	291.52	214.07	51.63	172.20
cráneo	144.41	169.92	44.04	70.31	51.63	12.45	41.53
pierna	481.66	566.75	146.88	234.50	172.20	41.53	138.52

figura 12

La matriz (W) es la matriz de suma de cuadrados y desviaciones entre cada dato y la media de su grupo.

Matriz de Suma de Cuadrados Dentro de Grupos (W)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	1,027.85	846.65	193.67	376.88	296.33	142.53	224.39
peso	846.65	1,941.85	209.00	393.95	419.00	209.80	188.03
pie	193.67	209.00	57.50	65.83	56.33	31.08	53.17
brazo	376.88	393.95	65.83	241.98	123.83	42.43	76.84
espalda	296.33	419.00	56.33	123.83	206.33	69.08	35.92
cráneo	142.53	209.80	31.08	42.43	69.08	75.73	42.37
pierna	224.39	188.03	53.17	76.84	35.92	42.37	120.50

figura 13

Las matrices T , B y W se relacionan:

$$T = B + W$$

$$W = W_0 + W_1$$

donde W_0 y W_1 son matrices de suma de cuadrados y productos cruzados de el grupo de las mujeres y hombres respectivamente.

Matriz de Suma de Cuadrados Total (T)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	2,702.67	2,817.33	704.39	1,192.28	895.11	286.94	706.06
peso	2,817.33	4,260.67	809.94	1,353.39	1,123.56	379.72	754.78
pie	704.39	809.94	213.24	314.48	238.93	75.12	200.05
brazo	1,192.28	1,353.39	314.48	638.96	415.35	112.74	311.34
espalda	895.11	1,123.56	238.93	415.35	420.41	120.71	208.12
cráneo	286.94	379.72	75.12	112.74	120.71	88.19	83.90
pierna	706.06	754.78	200.05	311.34	208.12	83.90	259.02

figura 14

Lambda de Wilks	0.2286632
Grupos	2
Variables	7
Casos	27
Grados de Libertad L1	7
Grados de Libertad L2	19
F	9.15595
Probabilidad Asociada	0.0001503

figura 15

Se ha obtenido que la Lambda que tiende más hacia el cero que hacia el uno entonces se podría decir que los grupos no están solapados o que se pueden diferenciar claramente.

Análisis de Varianza			
las Variables que intervienen en Manova			
Variable	Suma de Cuadrados		Total
	Entre Grupos	Dentro de Grupos	
estatura	1,674.82	1,027.85	2,702.67
peso	2,318.82	1,941.85	4,260.67
pie	155.74	57.50	213.24
brazo	396.98	241.98	638.96
espalda	214.07	206.33	420.41
cráneo	12.45	75.73	88.19
pierna	138.52	120.50	259.02

Cuadrados Medios		Grados de Libertad	
Entre Grupos	Dentro de Grupos	Entre Grupos	Dentro de Grupos
1,674.82	41.11	1	25
2,318.82	77.67	1	25
155.74	2.30	1	25
396.98	9.68	1	25
214.07	8.25	1	25
12.45	3.03	1	25
138.52	4.82	1	25

figura 16

Análisis de Varianza las Variables que intervienen en Manova		
Estadística	Probabilidad	Lamda de
F	Asociada	Wilks
40.736	0.000017	0.380
29.853	0.000060	0.456
67.713	0.000002	0.270
41.013	0.000016	0.379
25.938	0.000107	0.491
4.110	0.050734	0.859
28.740	0.000070	0.465

figura 17

Se tienen aquí resultados univariantes de la varianza de la varianza para las dos variables. Se puede obtener la significancia para cada variable ya que se tiene la estadística F y sus grados de libertad, y muchos se puede observar que ambas variables son y significantes y por tanto explican bien la función discriminante.

Cálculo de Matriz de Varianzas y Covarianzas Ponderada							
=====							
Matriz de Productos Cruzados							
Grupo: 0							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	526.93	309.40	89.33	219.13	132.83	38.53	126.27
peso	309.40	1,125.60	103.00	181.20	201.50	100.80	130.40
pie	89.33	103.00	26.83	42.83	20.83	10.58	27.67
brazo	219.13	181.20	42.83	103.73	55.83	16.43	63.47
espalda	132.83	201.50	20.83	55.83	131.83	35.83	15.67
cráneo	38.53	100.80	10.58	16.43	35.83	41.23	13.37
pierna	126.27	130.40	27.67	63.47	15.67	13.37	52.93

figura 18

En el cálculo de la matriz de varianzas y covarianzas Ponderadas, se puede observar que para el grupo 0 (grupo mujeres), existe una fuerte variabilidad con respecto a la variable estatura, la variable peso y variable brazo, pero al parecer existe en este grupo una variabilidad en cuanto a la variable craneo o pie, lo cual

indica que los individuos pertenecientes a este grupo son más parecidos en la medida de su cráneo o de su pie.

Matriz de Covarianza							
Grupo: 0							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	37.64	22.10	6.38	15.65	9.49	2.75	9.02
peso	22.10	80.40	7.36	12.94	14.39	7.20	9.31
pie	6.38	7.36	1.92	3.06	1.49	0.76	1.98
brazo	15.65	12.94	3.06	7.41	3.99	1.17	4.53
espalda	9.49	14.39	1.49	3.99	9.42	2.56	1.12
cráneo	2.75	7.20	0.76	1.17	2.56	2.95	0.95
pierna	9.02	9.31	1.98	4.53	1.12	0.95	3.78

figura 19

Para el caso de el grupo 1 (hombres) se observa también una fuerte variabilidad en la estatura pero también la viaribilidad es grande en el peso , pero no así en la piernas o en la medida de su cráneo.

Matriz de Productos Cruzados							
Grupo: 1							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	500.92	537.25	104.33	157.75	163.50	104.00	98.13
peso	537.25	816.25	106.00	212.75	217.50	109.00	57.63
pie	104.33	106.00	30.67	23.00	35.50	20.50	25.50
brazo	157.75	212.75	23.00	138.25	68.00	26.00	13.38
espalda	163.50	217.50	35.50	68.00	74.50	33.25	20.25
cráneo	104.00	109.00	20.50	26.00	33.25	34.50	29.00
pierna	98.13	57.63	25.50	13.38	20.25	29.00	67.56

figura 20

Matriz de Covarianza							
Grupo: 1							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	45.54	48.84	9.48	14.34	14.86	9.45	8.92
peso	48.84	74.20	9.64	19.34	19.77	9.91	5.24
pie	9.48	9.64	2.79	2.09	3.23	1.86	2.32
brazo	14.34	19.34	2.09	12.57	6.18	2.36	1.22
espalda	14.86	19.77	3.23	6.18	6.77	3.02	1.84
cráneo	9.45	9.91	1.86	2.36	3.02	3.14	2.64
pierna	8.92	5.24	2.32	1.22	1.84	2.64	6.14

figura 21

La Matriz de Varianzas y Covarianzas (W) también llamada matriz de suma de cuadrados y productos cruzados residual, se observa que las variables estatura y peso tienen gran variabilidad.

Matriz de Varianzas y Covarianzas (W)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	41.11	33.87	7.75	15.08	11.85	5.70	8.98
peso	33.87	77.67	8.36	15.76	16.76	8.39	7.52
pie	7.75	8.36	2.30	2.63	2.25	1.24	2.13
brazo	15.08	15.76	2.63	9.68	4.95	1.70	3.07
espalda	11.85	16.76	2.25	4.95	8.25	2.76	1.44
cráneo	5.70	8.39	1.24	1.70	2.76	3.03	1.69
pierna	8.98	7.52	2.13	3.07	1.44	1.69	4.82

figura 22

Matriz de Varianzas y Covarianzas Inversa (W-1)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	0.1389	0.0129	-0.2342	-0.1065	-0.0741	-0.0323	-0.0741
peso	0.0129	0.0325	-0.0752	-0.0282	-0.0353	-0.0352	-0.0008
pie	-0.2342	-0.0752	1.4974	0.1275	0.0191	0.0677	-0.2179
brazo	-0.1065	-0.0282	0.1275	0.2820	-0.0245	0.1032	-0.0228
espalda	-0.0741	-0.0353	0.0191	-0.0245	0.3302	-0.1433	0.1522
cráneo	-0.0323	-0.0352	0.0677	0.1032	-0.1433	0.6208	-0.1562
pierna	-0.0741	-0.0008	-0.2179	-0.0228	0.1522	-0.1562	0.4669

figura 23

Matriz de Varianzas y Covarianzas (W) x Su Inversa (W-1)							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	1	0	0	0	0	0	0
peso	0	1	0	0	0	0	0
pie	0	0	1	0	0	0	0
brazo	0	0	0	1	0	0	0
espalda	0	0	0	0	1	0	0
cráneo	0	0	0	0	0	1	0
pierna	0	0	0	0	0	0	1

figura 24

En el primer capítulo se vió que la función discriminante F_0 será la que va a separar mejor a los dos grupos. En esta función también se puede notar que la variable cráneo y pie tienen un fuerte peso o puntaje discriminante. De hecho se podría hacer la pregunta acerca de lo que sucedería si se hace un análisis discriminante solo con estas dos variables y comparar de esta manera la consistencia de clasificación.

Funciones Discriminantes		
Variable	F 0	F 1
estatura	-0.9895	-1.3030
peso	-4.3981	-4.4190
pie	17.7293	20.0421
brazo	9.5030	9.9804
espalda	-2.5148	-2.0750
cráneo	25.0769	24.3570
pierna	-4.7245	-4.3653
Constante	-1,016.1794	-1,082.5278

figura 25

Funciones Discriminantes Evaluadas y Probabilidades de Asignación						
=====						
Grupo original	Grupo Asignado	F			Probabilidad (F - 0)	Probabilidad (F - 1)
		Máxima	F - 0	F - 1		
0	0	964.61	964.61	951.46	0.999998	0.000002
0	0	995.78	995.78	984.94	0.999980	0.000020
0	0	1,030.23	1,030.23	1,020.54	0.999938	0.000062
0	0	988.95	988.95	980.40	0.999807	0.000193
0	0	1,021.82	1,021.82	1,013.90	0.999635	0.000365
0	0	1,009.66	1,009.66	1,001.93	0.999559	0.000441
0	0	973.12	973.12	965.44	0.999538	0.000462
0	0	1,101.03	1,101.03	1,093.52	0.999453	0.000547
0	0	1,081.80	1,081.80	1,074.74	0.999141	0.000859
0	0	963.47	963.47	956.87	0.998645	0.001355
0	0	963.16	963.16	959.72	0.969005	0.030995
0	0	1,040.06	1,040.06	1,037.29	0.941108	0.058892
0	0	973.49	973.49	971.18	0.909876	0.090124
0	0	1,070.57	1,070.57	1,068.55	0.882444	0.117556
0	0	1,047.32	1,047.32	1,046.36	0.723356	0.276644
1	1	996.28	994.72	996.28	0.173981	0.826019
1	1	1,107.71	1,105.49	1,107.71	0.098122	0.901878
1	1	1,092.03	1,089.12	1,092.03	0.051457	0.948543
1	1	1,019.74	1,015.01	1,019.74	0.008730	0.991270
1	1	1,085.71	1,080.78	1,085.71	0.007176	0.992824
1	1	1,067.68	1,062.26	1,067.68	0.004373	0.995627
1	1	1,070.78	1,065.15	1,070.78	0.003581	0.996419
1	1	1,107.49	1,101.67	1,107.49	0.002949	0.997051
1	1	1,102.47	1,095.62	1,102.47	0.001058	0.998942
1	1	1,034.31	1,026.29	1,034.31	0.000331	0.999669
1	1	1,130.79	1,119.55	1,130.79	0.000013	0.999987
1	1	1,155.85	1,141.97	1,155.85	0.000001	0.999999

figura 26

Consistencia de Clasificación			
Grupo Original	Grupo Asignado		Suma
	F - 0	F - 1	
0	15	0	15
1	0	12	12
Suma	15	12	27
Porcentaje de Consistencia	100%		

figura 27

En la tabla de Consistencia de Clasificación se presenta el grupo actual versus sus miembros asignados a los grupos “0” y “1” representados por $F - 0$ y $F - 1$ respectivamente.

Esta tabla presenta resultados hipotéticos de un análisis de clasificación en donde 27 fueron clasificados como hombres y mujeres, todo está basado en sus puntajes discriminantes de las variables predictoras. Leyendo esta tabla de derecha a izquierda siguiendo el renglón del grupo 0, al final, el número 15 indica que existen 15 individuos que pertenecen al grupo 0 (análogamente con el siguiente renglón). Si se toma en cuenta la columna $F - 0$, al final el número 15 indica el número de individuos “finalmente” asignados a el grupo 0. Se puede identificar en esta tabla que ningún individuo fue clasificado erróneamente, la explicación de esto se vio en la columna de probabilidades ($F - 0$) de la tabla anterior de probabilidades de asignación.

Matriz Producto W-1 x B							
Variable	estatura	peso	pie	brazo	espalda	cráneo	pierna
estatura	-1.3250	-1.5591	-0.4041	-0.6451	-0.4737	-0.1142	-0.3811
peso	-0.0883	-0.1039	-0.0269	-0.0430	-0.0316	-0.0076	-0.0254
pie	9.7754	11.5023	2.9809	4.7592	3.4949	0.8429	2.8113
brazo	2.0177	2.3742	0.6153	0.9823	0.7214	0.1740	0.5803
espalda	1.8587	2.1871	0.5668	0.9049	0.6645	0.1603	0.5346
cráneo	-3.0424	-3.5799	-0.9278	-1.4812	-1.0877	-0.2623	-0.8750
pierna	1.5183	1.7865	0.4630	0.7392	0.5428	0.1309	0.4367

figura 28

Eigen Valores y Vectores del producto W-1 x B							
Eigen Valores	Eigen Vector V(1)	Eigen Vector V(2)	Eigen Vector V(3)	Eigen Vector V(4)	Eigen Vector V(5)	Eigen Vector V(6)	Eigen Vector V(7)
3.373258	0.5840	0.1137	-0.5388	0.5043	-0.2292	-0.1010	0.1969
0	0.6871	-0.0265	0.4388	-0.4958	-0.2750	-0.0403	0.1073
-7E-06	0.1781	0.1875	-0.0325	0.0669	-0.1429	0.3335	-0.8922
0	0.2843	-0.3567	-0.4466	-0.3756	0.6566	-0.0171	-0.1417
-5E-06	0.2088	-0.5080	0.5100	0.5853	0.2953	-0.0111	-0.0912
0	0.0504	-0.0083	-0.0108	0.0178	0.0756	0.9333	0.3468
-1E-06	0.1679	0.7522	0.2376	0.1069	0.5722	-0.0745	0.0715

figura 29

La primer función explica el 100% de la varianza entre grupos, esto se corrobora con la correlación canónica y con el estadístico λ . El valor característico que maximiza es 3.373258, este auto valor alto indica que las variables discriminantes permiten distinguir bien entre los grupos y tiene una correlación canónica de .878258.

La correlación canónica de cada función con los grupos manifiesta la mayor asociación que presenta la primera de las funciones obtenidas, esto es, .878258 y la segunda tiene una correlación canónica de .0000163 lo que implica que prácticamente no es relevante considerar la segunda función. De hecho para este caso simple está analizando discriminantes para dos grupos, el número máximo de funciones discriminantes es uno. Como se tiene una correlación alta implica que las variables discriminantes permiten diferenciar entre los grupos.

Varianza Explicada			
Eigen Valores	% de Varianza Explicada	Varianza Acumulada	Correlación Canónica
3.373258	100.00%	100.00%	0.878258
0	0.00%	100.00%	0.000163

figura 30

Esta matriz muestra la composición de las variables para cada individuo de los dos grupos. Las variables que se presentan, ya son las estandarizadas. Se puede observar en esta matriz que los individuos del grupo 0 se obtienen valores negativos para la primer variable, es decir para la variable “mujeres”. Los individuos que pertenecen al grupo dos presentan en su mayoría valores positivos en su segunda variable (“hombres”). Los elementos del grupo 1 tienen la tendencia a pertenecer a dicho grupo si tienen valores positivos en la primer variable (“mujeres”).

Coefficientes de Funciones Canónicas Estandarizadas		
Variable	FCE 1	FCE 2
estatura	0.0182	0.0035
peso	0.0156	-0.0006
pie	0.0235	0.0247
brazo	0.0183	-0.0229
espalda	0.0145	-0.0354
cráneo	0.0058	-0.0010
pierna	0.0153	0.0685

figura 31

Los Coeficientes de Funciones Canónicas Estandarizadas identifican las variables con las diferencias más grandes entre los grupos y obtiene un coeficiente de ponderación para cada variable para reflejar estas diferencias. La variable pie (estandarizada) refleja un mayor peso en la primer función, también se observa que la variable cráneo es la variable que menos peso tiene.

Matriz de Valores de las Variables Estandarizadas (Z)							
Grupo	estatura	peso	pie	brazo	espalda	cráneo	pierna
0	-1.6456	-1.4756	-1.7394	-1.5054	-1.4553	-1.2167	-1.6135
0	-1.2533	-0.9287	-1.0411	-1.3037	-2.4500	-0.6737	-0.6630
0	-0.9590	-1.1631	-1.0411	-1.1020	-0.9579	-0.1307	-0.9798
0	-1.3514	-0.8506	-1.0411	-1.3037	-0.7092	-0.6737	-1.6135
0	-1.0571	-1.0850	-1.0411	-1.0011	-0.4605	-0.1307	-0.6630
0	-1.0571	-1.6318	-1.0411	-1.1020	-0.7092	-1.2167	-1.2966
0	-1.2533	0.0868	-1.0411	-1.1020	0.0368	0.4123	-0.6630
0	-0.0763	-0.6163	-0.5173	-0.5977	0.5342	1.4982	-0.9798
0	-0.1744	-0.9287	-0.6919	-0.0934	-1.0822	0.4123	0.2875
0	-0.4686	-1.0068	-1.0411	-0.4968	-0.3362	-1.7597	-0.9798
0	0.1199	0.4774	-0.3427	-0.0934	-0.2118	-0.6737	-0.0293
0	0.3160	0.0868	-0.3427	0.3101	0.5342	0.4123	0.2875
0	-0.7629	0.2430	-0.3427	-0.4968	-0.4605	-0.6737	-0.3462
0	-0.6648	0.3211	0.0065	-0.2951	-0.4605	0.9553	-0.3462
0	-0.0763	-1.2412	0.0065	-0.1942	-1.2066	-1.4882	-0.0293
1	-0.4686	-0.1476	0.0065	-0.0934	-0.4605	-1.2167	0.2875
1	1.3949	0.7899	0.7048	1.1169	0.4099	1.2267	1.2380
1	0.1199	0.2430	0.3556	0.7135	0.1612	0.4123	0.4459
1	0.4141	0.3993	0.7048	0.1083	0.5342	-0.6737	0.2875
1	0.9045	0.7899	1.0540	0.3101	1.0316	0.9553	0.6043
1	1.2969	2.1179	0.7048	1.9238	1.7776	0.9553	-0.0293
1	0.4141	0.0087	0.3556	1.1169	0.5342	-0.4022	1.2380
1	1.1988	0.7899	1.4032	0.1083	1.0316	1.4982	1.2380
1	0.7084	0.7117	1.0540	0.9152	0.5342	0.4123	0.6043
1	1.1988	1.2586	1.4032	0.5118	0.7829	-0.1307	0.9211
1	1.1988	0.9461	1.4032	1.9238	1.2803	-0.1307	-0.0293
1	1.9834	1.8054	2.1016	1.7221	1.7776	2.0412	2.8221

figura 32

La matriz de valores las variables estandarizadas muestra los puntajes estandarizados para cada uno de los individuos, con estos puntajes y los valores de los Coeficientes de Funciones Canónicas Estandarizadas es formada la función discriminante estandarizada.

Centroides de las Variables Estandarizadas por Grupo							
Grupo	estatura	peso	pie	brazo	espalda	cráneo	pierna
0	-0.6909	-0.6475	-0.7501	-0.6918	-0.6263	-0.3298	-0.6419
1	0.8637	0.8094	0.9376	0.8648	0.7829	0.4123	0.8023

figura 33

Es importante analizar las funciones con coeficientes estandarizados, ya que localiza a cada individuo en coordenadas que ya han sido transformadas, y que también se conocen como funciones canónicas estandarizadas.

Se promedian las puntuaciones discriminantes para todos los individuos dentro de un grupo particular, obteniendo la media del grupo es decir, el centroide, el

cual da la idea de localización en el espacio, para los centroides también es útil encontrar su puntaje discriminante ya que mientras más próximo este el puntaje de un individuo al puntaje de algún centroide será más probable la pertenencia a ese grupo.

Para las funciones canónicas estandarizadas tenemos un punto de corte igual a cero, este punto de corte es aquel punto que parte a los grupos y sirve para poder tomar la decisión de clasificar a un determinado individuo en un grupo o en otro.

Valores de las Funciones Canónicas Estandarizadas		
Grupo	FCE 1	FCE 2
0	-0.1742	-0.0714
0	-0.1352	0.0421
0	-0.1099	-0.0363
0	-0.1250	-0.0850
0	-0.0965	-0.0349
0	-0.1265	-0.0659
0	-0.0733	-0.0521
0	-0.0326	-0.0865
0	-0.0446	0.0426
0	-0.0878	-0.0690
0	-0.0076	-0.0001
0	0.0193	-0.0141
0	-0.0431	-0.0067
0	-0.0188	-0.0040
0	-0.0507	0.0472
1	-0.0217	0.0379
1	0.1067	0.0655
1	0.0389	0.0172
1	0.0406	0.0176
1	0.0890	0.0257
1	0.1393	-0.0892
1	0.0608	0.0510
1	0.1117	0.0829
1	0.0849	0.0293
1	0.1085	0.0620
1	0.1221	-0.0529
1	0.2259	0.1470

figura 34

Una vez evaluadas la funciones canónicas estandarizadas con los valores de los centroides de las variables estandarizadas se obtiene para cada grupo una puntuación (o score).

Valores de las Funciones Canónicas en los Centroides Estandarizados		
Grupo	estatura	peso
Centro 0	-0.0738	-0.0263
Centro 1	0.0922	0.0328

figura 35

Funciones Canónicas No Estandarizadas		
Variable	FCE 1	FCE 2
estatura	0.0028	0.0006
peso	0.0018	-0.0001
pie	0.0155	0.0163
brazo	0.0059	-0.0074
espalda	0.0051	-0.0123
cráneo	0.0033	-0.0005
pierna	0.0070	0.0312
Constante	-2.3502	-0.9324

figura 36

Valores de las Funciones Canónicas No Estandarizadas		
Grupo	FCNE 1	FCNE 2
0	-0.2746	-0.1199
0	-0.2100	0.0494
0	-0.1742	-0.0617
0	-0.1966	-0.1310
0	-0.1553	-0.0594
0	-0.1962	-0.1043
0	-0.1239	-0.0831
0	-0.0580	-0.1267
0	-0.0727	0.0524
0	-0.1362	-0.1103
0	-0.0127	-0.0041
0	0.0247	-0.0242
0	-0.0674	-0.0130
0	-0.0314	-0.0052
0	-0.0700	0.0664
1	-0.0298	0.0536
1	0.1652	0.0997
1	0.0611	0.0265
1	0.0685	0.0335
1	0.1400	0.0498
1	0.2142	-0.1238
1	0.0961	0.0741
1	0.1754	0.1370
1	0.1367	0.0518
1	0.1757	0.1046
1	0.1990	-0.0650
1	0.3524	0.2328

figura 37

Valores de las Funciones Canónicas en los Centroides No Estandarizados		
Grupo	FCNE 1	FCNE 2
Centro 0	-0.1170	-0.0450
Centro 1	0.1462	0.0562

figura 38

La distancia de Mahalanobis es un tipo de distancia semejante a la distancia euclídeana, pero la característica principal es que esta distancia se ve influenciada por la matriz de varianzas y covarianzas, debido a esto si un elemento pertenece al grupo 0 tendrá menor distancia a su grupo que al grupo 1.

Cuadrado de Distancias Mahalanobis a su Centroide	
Grupo	Valor
0	0.1525
0	0.4254
0	0.0592
0	0.1141
0	0.0996
0	0.1577
0	0.3251
0	0.3841
0	0.2636
0	0.2577
0	0.3020
0	0.2081
0	0.2104
0	0.3798
0	0.4060
1	0.2437
1	0.3141
1	0.1542
1	0.1301
1	0.1086
1	0.5109
1	0.3537
1	0.2849
1	0.0559
1	0.2367
1	0.4593
1	0.4026

figura 39

La gráfica de las funciones canónicas estandarizadas muestra los centroides de cada grupo, mostrar claramente la separación de los grupos, además muestra los pesos que los individuos tienen.

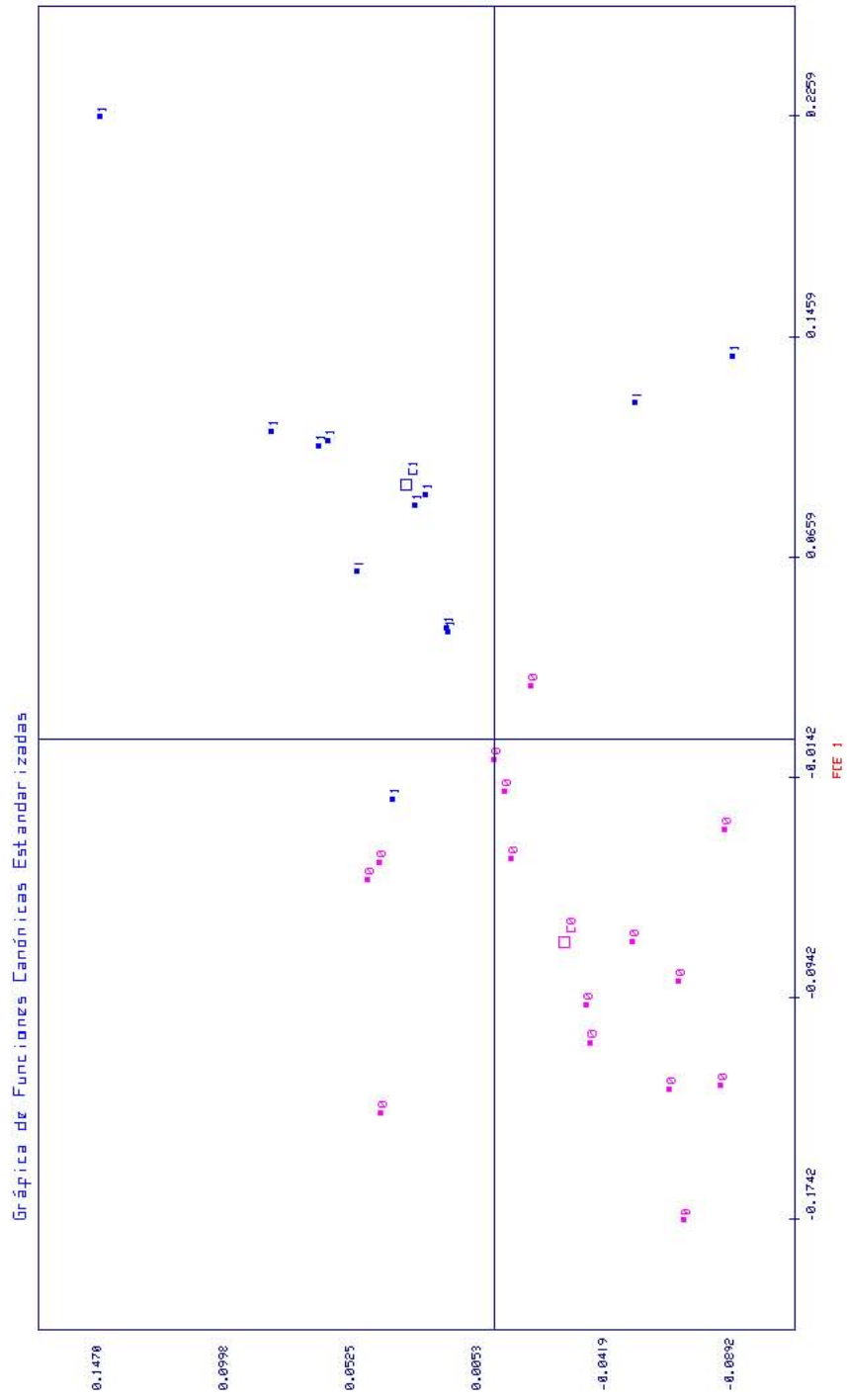


figura 40

CAPÍTULO III: DISCRIMINANTE PARA TUMORES CANCEROSOS (CASANDRA Y SPSS) Y ASIGNACIÓN DE CRÉDITO

III.1 TUMOR: MALIGNO VS BENIGNO.

A continuación se presentará el caso de dos grupos que representan tipos de tumor:

Maligno y Benigno.

Los datos para realizar este ejemplo fueron extraídos de una página de internet (<http://www.cs.wisc.edu/~olvi/uwmp/mpml.html>) que se dedica a analizar datos multivariantes, en ella se da una ligera explicación de los datos así como también de las variables utilizadas.

Entre las 30 variables se encuentran las siguientes:

- a) Radius (mean of distances from center to points on the perimeter).
- b) Texture (standard deviation of gray-scale values).
- c) Perimeter.
- d) Area.
- e) Smoothness (local variation in radius lengths).
- f) compactness ($\text{perimeter}^2 / \text{area} - 1.0$).
- g) concavity (severity of concave portions of the contour).
- h) concave points (number of concave portions of the contour).
- i) symmetry .
- j) fractal dimension ("coastline approximation" - 1).

Casandra hace un análisis estadístico para las estadísticas básicas. Si la media de una variable es significativamente diferente en varios grupos puede decirse que esta variable discrimina entre grupos.³

³Para permitir una mejor visualización se han ocultado algunos renglones de las tablas.

Manova de Grupos					
Estadísticas Básicas					
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestral n-1	coeficiente variación
x1	569	14.127	12.419	3.524	0.2494
x2	569	19.290	18.499	4.301	0.2230
x3	569	91.969	590.440	24.299	0.2642
x4	569	654.889	123843.554	351.914	0.5374
x5	569	0.096	0.000	0.014	0.1460
x6	569	0.104	0.003	0.053	0.5062
x7	569	0.089	0.006	0.080	0.8978
x8	569	0.049	0.002	0.039	0.7932
x9	569	0.181	0.001	0.027	0.1513
x10	569	0.063	0.000	0.007	0.1124
x11	569	0.405	0.077	0.277	0.6844
x12	569	1.217	0.304	0.552	0.4533
x13	569	2.866	4.088	2.022	0.7054
x14	569	40.337	2069.432	45.491	1.1278
x15	569	0.007	0.000	0.003	0.4265
x16	569	0.025	0.000	0.018	0.7029
x17	569	0.032	0.001	0.030	0.9464
x18	569	0.012	0.000	0.006	0.5231
x19	569	0.021	0.000	0.008	0.4024
x20	569	0.004	0.000	0.003	0.6972
x21	569	16.269	23.360	4.833	0.2971
x22	569	25.677	37.776	6.146	0.2394
x23	569	107.261	1129.131	33.603	0.3133
x24	569	880.583	324167.385	569.357	0.6466
x25	569	0.132	0.001	0.023	0.1725
x26	569	0.254	0.025	0.157	0.6188
x27	569	0.272	0.044	0.209	0.7665
x28	569	0.115	0.004	0.066	0.5735
x29	569	0.290	0.004	0.062	0.2133
x30	569	0.084	0.000	0.018	0.2152

figura 41

En el Análisis Discriminante se realizan diferentes desgloses de varianzas para someterlos a pruebas estadísticas y determinar el grado de asociación entre esas varianzas y por lo tanto entre las variables. Se busca determinar cuál de las variables contribuyen a la mejor discriminación entre grupos.

En las dos tablas siguientes se contemplan estadísticas básicas tomando los grupos por separado y para cada una de las variables.

Si se desea se puede calcular el coeficiente de variación para observar que tanto contribuye cierta variable. Este Coeficiente de variación sirve para tener una medida de dispersión no afectada por unidades.

Estadísticas Básicas					
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestral n-1	coeficiente variación
Variable h		B			
Casos		357			
x1	357	12.147	3.170	1.781	0.147
x2	357	17.915	15.961	3.995	0.223
x3	357	78.075	139.416	11.807	0.151
x4	357	462.790	19033.030	134.287	0.290
x5	357	0.092	0.000	0.013	0.145
x6	357	0.080	0.001	0.034	0.421
x7	357	0.046	0.002	0.043	0.943
x8	357	0.026	0.000	0.016	0.619
x9	357	0.174	0.001	0.025	0.142
x10	357	0.063	0.000	0.007	0.107
x11	357	0.284	0.013	0.113	0.396
x12	357	1.220	0.347	0.589	0.483
x13	357	2.000	0.595	0.771	0.386
x14	357	21.135	78.207	8.843	0.418
x15	357	0.007	0.000	0.003	0.425
x16	357	0.021	0.000	0.016	0.763
x17	357	0.026	0.001	0.033	1.266
x18	357	0.010	0.000	0.006	0.579
x19	357	0.021	0.000	0.007	0.340
x20	357	0.004	0.000	0.003	0.808
x21	357	13.380	3.926	1.981	0.148
x22	357	23.515	30.184	5.494	0.234
x23	357	87.006	182.982	13.527	0.155
x24	357	558.899	26765.426	163.601	0.293
x25	357	0.125	0.000	0.020	0.160
x26	357	0.183	0.008	0.092	0.505
x27	357	0.166	0.020	0.140	0.844
x28	357	0.074	0.001	0.036	0.481
x29	357	0.270	0.002	0.042	0.154
x30	357	0.079	0.000	0.014	0.174

figura 42

Estadísticas Básicas					
Variable	M				
Casos	212				
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muest. n-1	coeficiente variacion
x1	212	17.463	10.265	3.204	0.183
x2	212	21.605	14.284	3.779	0.175
x3	212	115.365	477.626	21.855	0.189
x4	212	978.376	135378.355	367.938	0.376
x5	212	0.103	0.000	0.013	0.123
x6	212	0.145	0.003	0.054	0.372
x7	212	0.161	0.006	0.075	0.467
x8	212	0.088	0.001	0.034	0.391
x9	212	0.193	0.001	0.028	0.143
x10	212	0.063	0.000	0.008	0.121
x11	212	0.609	0.119	0.345	0.566
x12	212	1.211	0.233	0.483	0.399
x13	212	4.324	6.597	2.569	0.594
x14	212	72.672	3764.469	61.365	0.844
x15	212	0.007	0.000	0.003	0.426
x16	212	0.032	0.000	0.018	0.570
x17	212	0.042	0.000	0.022	0.517
x18	212	0.015	0.000	0.006	0.366
x19	212	0.020	0.000	0.010	0.492
x20	212	0.004	0.000	0.002	0.502
x21	212	21.135	18.349	4.284	0.203
x22	212	29.318	29.537	5.435	0.185
x23	212	141.370	867.718	29.467	0.208
x24	212	1422.286	357565.422	597.968	0.420
x25	212	0.145	0.000	0.022	0.151
x26	212	0.375	0.029	0.170	0.455
x27	212	0.451	0.033	0.182	0.403
x28	212	0.182	0.002	0.046	0.254
x29	212	0.323	0.006	0.075	0.231
x30	212	0.092	0.000	0.022	0.235

figura 43

Se usa el estadístico de Lambda de Wilks para determinar la significancia de las variables que se introducen. Como se puede ver en esta tabla, la Lambda de Wilks es más cercana a cero y esto indica un solapamiento bajo, lo suficiente como para poder separar bien a los grupos con las 30 variables.

Lamda de Wilks	0.225675347
Grupos	2
Variables	30
Casos	569
Grados de Libertad L1	30
Grados de Libertad L2	538
F	61.531852
Probabilidad Asociada	0

figura 44

Las siguientes dos tablas son procedimientos preliminares a la prueba F y a la Lambda de Wilks. MANOVA se usará para comparar medias.

Análisis de Varianza					
las Variables que intervienen en Manova					
Variable	Suma de Cuadrados	Dentro de Grupos	Total	Cuadrados Medios	Dentro de Grupos
	Entre Grupos			Entre Grupos	
X1	3.759.34	3294.605	7053.947	3759.342	5.811
X2	1.811.25	8696.130	10507.380	1811.250	15.337
X3	184.959.19	150411.006	335370.192	184959.186	265.275
X4	35.358.547.15	34984591.698	70343138.852	35358547.155	61701.220
X5	0.014444	0.098	0.112	0.014	0.000
X6	0.563762	1.020	1.584	0.564	0.002
X7	1.750444	1.859	3.610	1.750	0.003
X8	0.515805	0.339	0.855	0.516	0.001
X9	0.046627	0.380	0.427	0.047	0.001
X10	0.000005	0.028	0.028	0.000	0.000
X11	14.049441	29.631	43.681	14.049	0.052
X12	0.011917	172.840	172.851	0.012	0.305
X13	718.153896	1603.771	2321.925	718.154	2.829
X14	353.292.50	822144.642	1175437.139	353292.497	1449.991
X15	0.000023	0.005	0.005	0.000	0.000
X16	0.015638	0.167	0.182	0.016	0.000
X17	0.03332	0.484	0.518	0.033	0.001
X18	0.003601	0.018	0.022	0.004	0.000
X19	0.000002	0.039	0.039	0.000	0.000
X20	0.000024	0.004	0.004	0.000	0.000
X21	7.999.38	5269.223	13268.607	7999.384	9.293
X22	4.479.38	16977.666	21457.042	4479.376	29.943
X23	393.116.14	248230.178	641346.321	393116.143	437.796
X24	99.152.279.11	84974795.630	184127074.738	99152279.108	149867.364
X25	0.052599	0.244	0.296	0.053	0.000
X26	4.911109	9.150	14.061	4.911	0.016
X27	10.756049	13.966	24.722	10.756	0.025
X28	1.546513	0.909	2.454	1.546	0.002
X29	0.376768	1.797	2.174	0.377	0.003
X30	0.019435	0.166	0.185	0.019	0.000

figura 45

Variable	Grados de Libertad		Estadística F	Probabilidad Asociada	Lamda de Wilks
	Entre Grupos	Dentro de Grupos			
X1	1	567	646.981	0.000	0.467
x2	1	567	118.096	0.000	0.828
x3	1	567	697.235	0.000	0.448
x4	1	567	573.061	0.000	0.497
x5	1	567	83.651	0.000	0.871
x6	1	567	313.233	0.000	0.644
x7	1	567	533.793	0.000	0.515
x8	1	567	861.676	0.000	0.397
x9	1	567	69.527	0.000	0.891
x10	1	567	0.093	0.758	1.000
x11	1	567	268.840	0.000	0.678
x12	1	567	0.039	0.838	1.000
x13	1	567	253.897	0.000	0.691
x14	1	567	243.652	0.000	0.699
x15	1	567	2.558	0.106	0.996
x16	1	567	53.247	0.000	0.914
x17	1	567	39.014	0.000	0.936
x18	1	567	113.263	0.000	0.834
x19	1	567	0.024	0.871	1.000
x20	1	567	3.468	0.060	0.994
x21	1	567	860.782	0.000	0.397
x22	1	567	149.597	0.000	0.791
x23	1	567	897.944	0.000	0.387
x24	1	567	661.600	0.000	0.462
x25	1	567	122.473	0.000	0.822
x26	1	567	304.341	0.000	0.651
x27	1	567	436.692	0.000	0.565
x28	1	567	964.385	0.000	0.370
x29	1	567	118.860	0.000	0.827
x30	1	567	66.444	0.000	0.895

figura 46

Las funciones discriminantes están formadas por los puntajes discriminantes de cada una de las 30 variables con su respectiva constante. CASANDRA nos proporciona estos valores, pero más adelante se mostrará la función discriminante estandarizada.

Funciones Discriminantes		
Variable	F B	F M
X1	132.685	128.878
x2	3.008	3.078
x3	-1.147	-0.778
x4	-0.917	-0.911
x5	-786.143	-776.462
x6	-2733.014	-2802.852
x7	478.288	502.380
x8	-1036.595	-992.185
x9	507.416	510.336
x10	19110.246	19028.810
x11	119.200	123.663
x12	1.296	1.021
x13	-16.886	-16.826
x14	0.625	0.610
x15	-4929.690	-4578.475
x16	2543.499	2500.537
x17	-299.919	-375.282
x18	212.679	384.099
x19	644.562	680.810
x20	-15142.580	-14820.736
x21	-37.099	-33.158
x22	-0.389	-0.237
x23	3.249	3.164
x24	0.083	0.064
x25	1548.181	1553.288
x26	110.912	115.418
x27	-51.616	-42.338
x28	-10.084	0.161
x29	-114.471	-103.873
x30	-1481.869	-1436.323
Constante	-1005.756	-1045.988

figura 47

En la figura 66 se pueden observar las funciones discriminantes evaluadas para cada individuo de la población y junto a esto la probabilidad de que dicho elemento de el grupo tenga un tumor maligno o benigno. Es claro ver que para una probabilidad de asignación alta en el grupo de los que tienen tumores malignos se le asigna ese grupo, aunque pueden existir algunos errores de asignación debido al solapamiento de los grupos en donde se encuentran sujetos no fácilmente diferenciables.

Funciones Discriminantes Evaluadas y Probabilidades de Asignación						
Grupo original	Grupo Asignado	F Máxima	F - B	F - M	Probabilidad (F - B)	Probabilidad (F - M)
M	M	1,093.72225	1,083.07635	1,093.72225	0.00002	0.99998
M	M	1,067.45118	1,061.32642	1,067.45118	0.00218	0.99782
M	M	1,027.42913	1,015.29278	1,027.42913	0.00001	0.99999
M	M	1,157.43922	1,145.38191	1,157.43922	0.00001	0.99999
M	M	991.82529	985.19914	991.82529	0.00132	0.99868
M	M	1,039.30366	1,035.59218	1,039.30366	0.02386	0.97614
M	M	1,046.02222	1,039.44895	1,046.02222	0.00140	0.99860
M	M	1,078.38821	1,076.40249	1,078.38821	0.12071	0.87929
M	M	1,033.15793	1,029.17057	1,033.15793	0.01821	0.98179
M	M	942.79078	930.32015	942.79078	0.00000	1.00000
M	M	1,042.22459	1,041.42046	1,042.22459	0.30914	0.69086
M	M	1,013.60307	1,006.63185	1,013.60307	0.00094	0.99906
M	M	1,105.64559	1,102.47365	1,105.64559	0.04024	0.95976
M	B	974.60822	974.60822	973.82910	0.68549	0.31451
M	M	1,016.34582	1,013.65518	1,016.34582	0.06353	0.93647
M	M	1,127.41982	1,118.29875	1,127.41982	0.00011	0.99989
M	M	1,047.67068	1,042.15227	1,047.67068	0.00400	0.99600
M	M	1,053.32882	1,043.80516	1,053.32882	0.00007	0.99993
M	M	1,045.78581	1,034.49079	1,045.78581	0.00001	0.99999
B	B	979.13154	979.13154	976.31347	0.94364	0.05636
B	B	1,052.63574	1,052.63574	1,042.43465	0.99996	0.00004
B	B	1,006.03480	1,006.03480	993.80262	1.00000	0.00000
B	B	1,027.31684	1,027.31684	1,021.65288	0.99654	0.00346
B	B	990.49218	990.49218	979.12083	0.99999	0.00001
B	B	1,033.25461	1,033.25461	1,023.34740	0.99995	0.00005
B	B	1,056.59164	1,056.59164	1,052.40264	0.98507	0.01493
B	B	954.51012	954.51012	942.78608	0.99999	0.00001
B	B	1,029.16220	1,029.16220	1,024.57472	0.98992	0.01008
B	B	908.27631	908.27631	901.43922	0.99893	0.00107
B	B	966.26570	966.26570	952.97530	1.00000	0.00000
B	B	906.92622	906.92622	897.73737	0.99990	0.00010
B	B	990.78050	990.78050	982.21698	0.99981	0.00019
B	B	976.13519	976.13519	970.11358	0.99758	0.00242
B	B	974.44376	974.44376	971.40844	0.95414	0.04586
B	B	943.99540	943.99540	932.70523	0.99999	0.00001
M	M	1,015.27965	1,001.12088	1,015.27965	0.00000	1.00000
M	M	1,072.23137	1,060.12108	1,072.23137	0.00001	0.99999
M	M	1,057.00491	1,042.31469	1,057.00491	0.00000	1.00000
M	M	1,081.98365	1,073.61563	1,081.98365	0.00023	0.99977
M	M	1,044.69066	1,043.46690	1,044.69066	0.22728	0.77272
M	M	1,061.66432	1,041.01124	1,061.66432	0.00000	1.00000
B	B	750.41061	750.41061	738.52040	0.99999	0.00001

figura 48

La consistencia de clasificación verifica el grupo en que fueron asignados los elementos en base a las funciones discriminantes. De los 569 elementos analizados el 96% fueron asignados de manera exitosa. Se realizará un ejercicio con estos mismos datos, pero tomando en cuenta las variables con un mayor peso discriminante para ver si con ellas se puede llegar a un resultado suficientemente satisfactorio y más fácil de obtener al requerir menos variables.

Consistencia de Clasificación			
Grupo Original	Grupo Asignado F - B	F - M	Suma
B	355	2	357
M	18	194	212
Suma	373	196	569
Porcentaje de Consistencia	96%		

figura 49

La primer Función Discriminante explica el 99.99% de la varianza entre grupos, esto se corrobora con la correlación canónica alta (.999517).

Varianza Explicada Eigen			
Valores	% de Varianza Explicada	Varianza Acumulada	Correlación Canónica
1.03468	99.99%	99.99%	0.999517
0.074325	0.01%	100.00%	0.263027

figura 50

Las Funciones Canónicas Estandarizadas proveen un panorama más gráfico de la posición de las variables en un plano bidimensional dado por dos funciones discriminantes.

Coeficientes de Funciones Canónicas Estandarizadas		
Variable	FCE 1	FCE 2
X1	0.000154	0.000000
x2	0.000000	0.010724
x3	0.000095	0.000000
x4	0.000085	0.000000
x5	0.000000	0.000000
x6	-0.008494	0.000000
x7	0.006923	0.000000
x8	0.016094	0.000000
x9	0.000000	0.000000
x10	0.000000	0.000000
x11	0.000000	0.000000
x12	0.000000	0.000000
x13	0.000000	0.000000
x14	0.000039	0.000000
x15	0.029795	0.000000
x16	0.000000	0.000000
x17	-0.001518	0.000000
x18	0.068831	0.000000
x19	-0.028475	0.000000
x20	-0.085164	0.000000
x21	0.000000	0.000000
x22	0.000000	0.000000
x23	0.000095	0.000000
x24	0.000094	0.000000
x25	0.000000	0.000000
x26	0.000000	0.000000
x27	0.000000	0.000000
x28	0.000000	0.000000
x29	0.000000	0.000000
x30	0.000000	0.000000

figura 51

III.1.1 ANALISIS PARA EL CASO REDUCIDO A 7 VARIABLES.. Como se dijo anteriormente el analisis de tumor se hará ahora pero con las variables $X_1, X_3, X_7, X_8, X_{21}, X_{23}$ y X_{28} .

Manova de Grupos						
Estadísticas Básicas						
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1		coeficiente variacion
x1	569	14.13	12.42	3.52		0.2494
x3	569	91.97	590.44	24.30		0.2642
x7	569	0.09	0.01	0.08		0.8976
x8	569	0.05	0.00	0.04		0.7932
x21	569	16.27	23.36	4.83		0.2971
x23	569	107.26	1129.13	33.60		0.3133
x28	569	0.11	0.00	0.07		0.5735

figura 52

Nuevamente Casandra hace un análisis para las estadísticas básicas para cada grupo.

Estadísticas Básicas						
Variable Indicadora (DIAGNT) -> B						
Casos 357						
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1		coeficiente variacion
x1	357	12.15	3.17	1.78		0.1466
x3	357	78.08	139.42	11.81		0.1512
x7	357	0.05	0.00	0.04		0.9432
x8	357	0.03	0.00	0.02		0.6186
x21	357	13.38	3.93	1.98		0.1481
x23	357	87.01	182.98	13.53		0.1555
x28	357	0.07	0.00	0.04		0.4809

figura 53

Se realizará un breve analisis de algunos coeficientes de variación.

Estadísticas Básicas						
Variable Indicadora (DIAGNT) -> M						
Casos 212						
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1		coeficiente variacion
x1	212	17.46	10.27	3.20		0.1835
x3	212	115.37	477.63	21.85		0.1894
x7	212	0.16	0.01	0.08		0.4666
x8	212	0.09	0.00	0.03		0.3907
x21	212	21.13	18.35	4.28		0.2027
x23	212	141.37	867.72	29.45		0.2084
x28	212	0.18	0.00	0.05		0.2541

figura 54

La Lamda de Wilks es muy similar al analisis realizado con 30 variables pero al contemplar siete variables , se puede concluir qe los grupos también se pueden separar de manera exitosa con una lamda de Wilks igual a 0.300984.

Lamda de Wilks	0.30098418
Grupos	2
Variables	7
Casos	569
Grados de Libertad L1	7
Grados de Libertad L2	561
F	186.126477
Probabilidad Asociada	0

figura 55

Como antes se analizan la varianza de las variables que intervienen en el MANOVA

Análisis de Varianza las Variables que intervienen en Manova						
Variable	Suma de Cuadrados			Cuadrados Medios		
	Entre Grupos	Dentro de Grupos	Total	Entre Grupos	Dentro de Grupos	
X1	3,759.34	3,294.60	7,053.95	3,759.34	5.81	
x3	184,959.19	150,411.01	335,370.19	184,959.19	265.28	
x7	1.75	1.86	3.61	1.75	0.00	
x8	0.52	0.34	0.86	0.52	0.00	
x21	7,999.38	5,269.22	13,268.61	7,999.38	9.29	
x23	393,116.14	248,230.18	641,346.32	393,116.14	437.80	
x28	1.55	0.91	2.45	1.55	0.00	

Análisis de las Variables Inter						
Variable	Grados de Libertad		Estadística F	Probabilidad Asociada	Lamda de Wilks	
	Entre Grupos	Dentro de Grupos				
X1	1	567	646.981	0.000	0.467	
x3	1	567	697.235	0.000	0.448	
x7	1	567	533.793	0.000	0.515	
x8	1	567	861.676	0.000	0.397	
x21	1	567	860.782	0.000	0.397	
x23	1	567	897.944	0.000	0.387	
x28	1	567	964.385	0.000	0.370	

figura 56

He aquí las probabilidades de asignación para algunos de los individuos de ambos grupos, se puede observar que algunos están claramente asignados ya que las probabilidades de asignación son relativamente altas, algunos otros son colocados en el grupo que esté más “cercano” es decir al grupo al que tenga la probabilidad de asignación más alta.

Funciones Discriminantes Evaluadas y Probabilidades de Asignación						
Grupo original	Grupo Asignado	F Máxima	F - B	F - M	Probabilidad (F - B)	Probabilidad (F - M)
M	M	34.07	22.39	34.07	0.00	1.00
M	M	53.57	46.92	53.57	0.00	1.00
M	M	45.20	35.63	45.20	0.00	1.00
M	M	22.96	16.99	22.96	0.00	1.00
M	M	39.83	36.31	39.83	0.03	0.97
M	M	19.29	17.58	19.29	0.15	0.85
M	M	42.06	36.54	42.06	0.00	1.00
M	M	21.94	20.92	21.94	0.26	0.74
M	M	21.07	18.38	21.07	0.06	0.94
M	M	27.06	23.39	27.06	0.02	0.98
M	B	33.12	33.12	32.12	0.73	0.27
M	M	31.82	27.93	31.82	0.02	0.98
M	M	22.04	20.51	22.04	0.18	0.82
M	B	31.08	31.08	28.73	0.91	0.09
M	M	28.32	26.86	28.32	0.19	0.81
M	M	27.69	26.36	27.69	0.21	0.79
M	M	31.53	28.81	31.53	0.06	0.94
M	M	26.72	20.37	26.72	0.00	1.00
M	M	50.15	39.22	50.15	0.00	1.00
B	B	26.36	26.36	24.49	0.87	0.13
B	B	19.11	19.11	13.65	1.00	0.00
B	B	14.39	14.39	7.08	1.00	0.00
M	M	36.08	30.84	36.08	0.01	0.99
M	M	52.75	41.74	52.75	0.00	1.00
M	M	32.71	21.97	32.71	0.00	1.00
M	M	28.17	18.72	28.17	0.00	1.00
M	M	33.79	27.93	33.79	0.00	1.00
M	M	40.37	37.77	40.37	0.07	0.93
M	M	28.57	24.13	28.57	0.01	0.99
M	M	31.89	30.39	31.89	0.18	0.82
M	M	32.49	26.24	32.49	0.00	1.00
M	M	20.10	19.05	20.10	0.26	0.74
M	M	33.46	27.29	33.46	0.00	1.00
M	M	42.97	37.60	42.97	0.00	1.00
M	M	30.83	26.71	30.83	0.02	0.98
M	M	40.24	37.25	40.24	0.05	0.95
M	B	30.71	30.71	29.19	0.82	0.18
B	B	24.69	24.69	18.15	1.00	0.00
M	B	28.94	28.94	21.89	1.00	0.00
M	M	35.37	33.15	35.37	0.10	0.90
M	B	29.32	29.32	26.83	0.92	0.08
M	B	15.99	15.99	14.03	0.88	0.12
M	M	45.47	37.79	45.47	0.00	1.00
M	M	19.48	18.12	19.48	0.20	0.80
M	M	25.84	24.97	25.84	0.30	0.70
M	M	46.21	38.11	46.21	0.00	1.00
B	B	9.95	9.95	0.20	1.00	0.00
M	M	29.42	26.44	29.42	0.05	0.95
B	B	20.09	20.09	14.61	1.00	0.00
B	B	28.82	28.82	26.59	0.90	0.10

figura 57

Haciendo este análisis se puede observar que el porcentaje de consistencia se redujo, de un 96% con 30 variables a un 94%. Este análisis podría tomarse como primer estudio. En los casos en los que el individuo haya obtenido una probabilidad de asignación “baja” para su grupo (el investigador decidirá qué es lo que se puede considerar bajo), cabe la posibilidad de realizar un análisis más amplio ya que con una probabilidad de asignación baja es más fácil caer en errores de clasificación. Es importante mencionar que el costo de error de una mala selección puede ser literalmente de vital importancia, ya que de no detectarse pronto un tumor maligno este puede evolucionar.

Consistencia de Clasificación			
Grupo Original	Grupo Asignado		Suma
	F - B	F - M	
B	355	2	357
M	31	181	212
Suma	386	183	569
Porcentaje de Consistencia		94%	

figura 58

Esta tabla explica que la primer función explica casi el 100% de la varianza entre grupos (el 100% exacta se debe a un redondeo de decimales), esto se corrobora con la correlación canónica alta (.836566).

Varianza Explicada			
Eigen	% de Varianza	Varianza	Correlación
Valores	Explicada	Acumulada	Canónica
2.33158	100.00%	100.00%	0.836566
0.000002	0.00%	100.00%	0.001573

figura 59

Aquí tenemos el cuadrado de las distancias de Mahalanobis a su centroide. A cada individuo se le ha asignado una distancia, la cual nos da una idea de que tan “cerca” un individuo está de la media de su grupo.

Cuadrado de Distancias Mahalanobis a su Centroide

Grupo	Valor
M	0.063742
M	0.019119
M	0.01875
M	0.036536
M	0.011179
M	0.010809
M	0.005805
M	0.011937
M	0.012711
M	0.033606
M	0.010773
M	0.007063
M	0.069784
M	0.01267
M	0.027697
M	0.013255
M	0.008566
M	0.019933
M	0.033948
B	0.006208
B	0.006642
B	0.003916
M	0.010923
M	0.032244
M	0.062018
M	0.023768
M	0.03268
M	0.005539

figura 60

III.2 COMPARATIVO CON SPSS.

Análogamente a lo que sucede en en el sistema estadístico CASANDRA, SPSS puede utilizar bases de datos con extensión xls (Excel). Cuando ya se encuentra la base de datos de SPSS que se importarse elige de el menu Analyze→ Classify →Discriminant... tal y como se muestra en la figura:

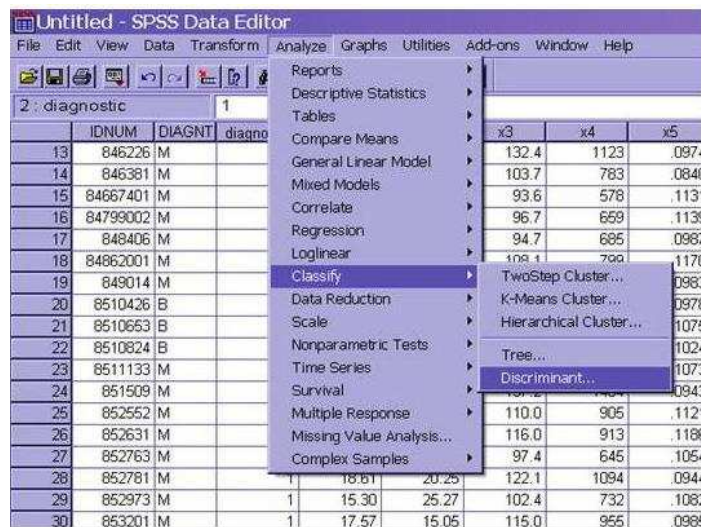


figura 61

Con lo que aparecerá una ventana en la que se elige quien será la variable de agrupamiento y quien o quienes serán las variables independientes. Es importante definir cual será el rango de la variable de agrupamiento , ya que sin esto no se podrá tener el cálculo de salida.

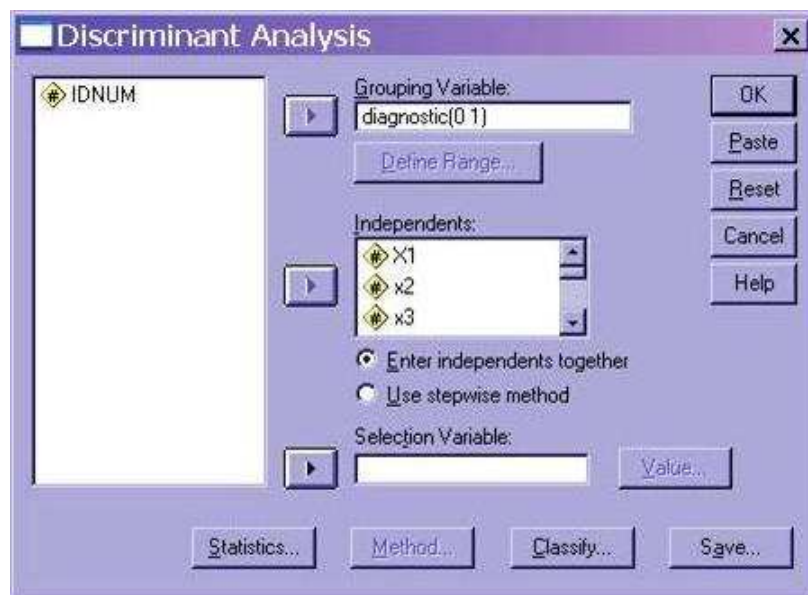


figura 62

Si se pretende tener un mayor detalle de las estadísticas de salida, SPSS preguntará que es lo que necesitas exactamente tener en el “output” , como el tipo de probabilidades a priori, el uso de la matriz de covarianza o incluso si se requiere hacer una depuración de la base de datos en el sentido de intercambiar un dato nulo por la media con la finalidad de no alterar demasiado la base misma.

Para este ejemplo se ha seleccionado todo lo que se muestra en la figura de abajo para poder comparar todas las estadísticas de salida de SPSS con CASANDRA.

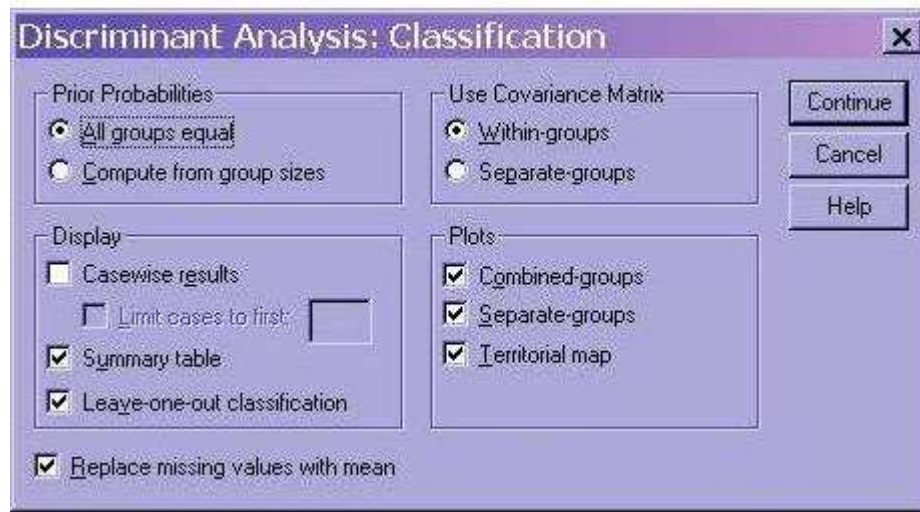


figura 63

También se pueden elegir estadísticas descriptivas ,matrices o inclusive los coeficientes de Fisher's y los no estandarizados.

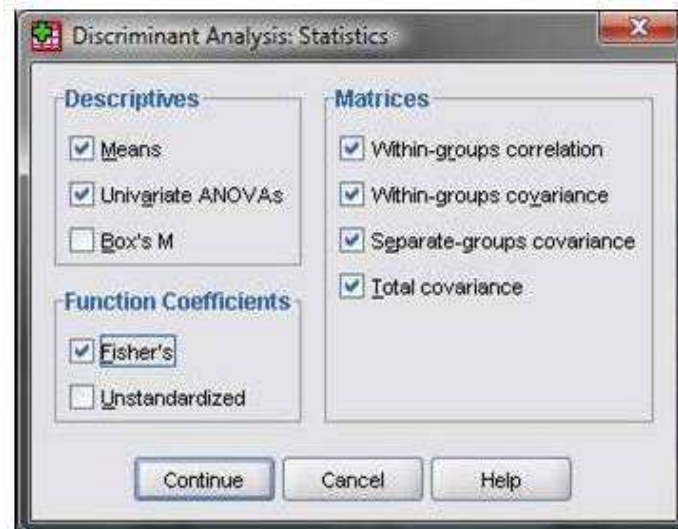


figura 64

SPSS analiza primeramente cuales han sido los datos tomados, cuales han sido excluidos y su total, de manera que el usuario podra percatarse si ha perdido en este primer proceso de analisis informacion importante y de ser así se dará a la tarea de verificar su base y de ser posible depurarla.

Analysis Case Processing Summary		
Unweighted Cases	N	Percent
Valid	569	100.0
Excluded		
Missing or out-of-range group codes	0	.0
At least one missing discriminating variable	0	.0
Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
Total	0	.0
Total	569	100.0

figura 65

Tal y como ocurre para el sistema CASANDRA , este proporciona estadísticas de grupo, es decir , estadísticas para cada uno de los grupos de análisis. Para el las pruebas de igualdad de medias SPSS tiene la siguiente vista:

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
X1	.467	646.981	1	567	.000
x2	.828	118.096	1	567	.000
x3	.448	697.235	1	567	.000
x4	.497	573.061	1	567	.000
x5	.871	83.651	1	567	.000
x6	.644	313.233	1	567	.000
x7	.515	533.793	1	567	.000
x8	.397	861.676	1	567	.000
x9	.891	69.527	1	567	.000
x10	1.000	.093	1	567	.760
x11	.678	268.840	1	567	.000
x12	1.000	.039	1	567	.843
x13	.691	253.897	1	567	.000
x14	.699	243.652	1	567	.000
x15	.996	2.558	1	567	.110
x16	.914	53.247	1	567	.000
x17	.936	39.014	1	567	.000
x18	.834	113.263	1	567	.000
x19	1.000	.024	1	567	.877
x20	.994	3.468	1	567	.063
x21	.397	860.782	1	567	.000
x22	.791	149.597	1	567	.000
x23	.387	897.944	1	567	.000
x24	.462	661.600	1	567	.000
x25	.822	122.473	1	567	.000
x26	.651	304.341	1	567	.000
x27	.565	436.692	1	567	.000
x28	.370	964.385	1	567	.000
x29	.827	118.860	1	567	.000
x30	.895	66.444	1	567	.000

figura 66

Se hace un análisis de las matrices de suma de cuadrados entre grupos, y análisis de las matrices de covarianzas y correlaciones total⁴

Después de el encabezado “Analysis 1” se presenta la siguiente tabla de variables que no pasan la prueba de tolerancia:

⁴vea anexo 2 para el detalle figura A_1y A_2

Analysis 1

	Within-Groups Variance	Tolerance	Minimum Tolerance
x13	2.829	.041	.001
x18	3.18E-005	.202	.001
x19	6.85E-005	.502	.001
x21	9.293	.011	.001
x22	29.943	.103	.001
x24	149867.364	.024	.001
x25	.000	.197	.001
x26	.016	.138	.001
x27	.025	.133	.001
x28	.002	.190	.001
x29	.003	.370	.001
x30	.000	.159	.001

All variables passing the tolerance criteria are entered simultaneously.

a. Minimum tolerance level is .001.

figura 67

SPSS hace una prueba de tolerancia que permite no considerar a las variables que resulten redundantes. Como se puede ver a partir de este momento, ya no contemplará a las variables $X_{13}, X_{18}, X_{19}, X_{21}, X_{22}, X_{23}, X_{24}, X_{25}, X_{26}, X_{27}, X_{28}, X_{29}$ y X_{30} y tomará como base el complemento de esas variables. Dará valores para las funciones discriminantes, coeficientes canónicos de la función discriminante y otros.

Summary of Canonical Discriminant Functions

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.705 ^a	100.0	100.0	.854

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.270	730.777	18	.000

figura 68

Desde un inicio se sabía que el máximo de funciones discriminantes sería 1 y por lo tanto la función explica el 100% de la varianza entre grupos, lo cual hace sentido con una correlación canónica alta de .854 . Por su parte CASANDRA llega a resultados semejantes cuando contempla otras variables ($X_1, X_3, X_7, X_8, X_{21}, X_{23}$ y X_{28}).⁵⁶

La Lambda de Wilks tiene con SPSS .270 mientras que para CASANDRA es .2256 .⁷ ⁸CASANDRA es muy claro en decir cuantas variables está usando para el cálculo de lambda de Wilks.

SPSS proporciona los coeficientes de la función canónica estandarizada (standardized Canonical Discriminant function coefficients) sin contemplar a las variables de la figura 67.

⁵vea sección III.1 figura 59

⁶Cabe decir que el cuando se hizo el análisis con CASANDRA en un grupo reducido de variables, se consideraron a las variables que no presentaban correlaciones altas con las demás, por otro lado es mucho más fácil trabajar con seis variables que con 30 o inclusive con las que termina usando SPSS.

⁷Existe en SPSS una duda en cuanto a si se consideran las 30 variables o no

⁸En la siguiente sección se presentará un resultado contemplando las variables que SPSS usa para hacer el discriminante.

**Standardized
Canonical
Discriminant
Function
Coefficients**

	Function
	1
X1	5.427
x2	.334
x3	-4.496
x4	-1.623
x5	.094
x6	-.430
x7	.952
x8	.280
x9	.112
x10	.190
x11	.327
x12	-.110
x14	-.330
x15	.166
x16	.169
x17	-.479
x20	-.086
x23	1.091

figura69

La matriz de estructura que ordena todos los puntajes discriminantes de mayor a menor de todas las variables⁹

⁹Esta estructura da una idea clara del peso o importancia que tiene cada variable en la función discriminante.

Structure Matrix

	Function	
	1	
x23	.765	
x8	.750	
x21 ^a	.738	
x24 ^a	.676	
x3	.674	
X1	.650	
x28 ^a	.646	
x4	.611	
x7	.590	
x6	.452	
x11	.419	
x13 ^a	.412	
x27 ^a	.410	
x14	.399	
x26 ^a	.319	
x2	.277	
x18 ^a	.251	
x5	.234	
x22 ^a	.229	Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.
x9	.213	
x16	.186	
x17	.159	
x25 ^a	.149	
x29 ^a	.107	
x30 ^a	.060	
x20	.048	
x15	-.041	
x19 ^a	-.036	
x10	-.008	a. This variable not used in the analysis.
x12	-.005	

figura 70

Los coeficientes de la función canónica discriminante sin estandarizar.

**Canonical Discriminant
Function Coefficients**

	Function
	1
X1	2.251
x2	.085
x3	-.276
x4	-.007
x5	7.117
x6	-10.128
x7	16.626
x8	11.453
x9	4.317
x10	26.876
x11	1.430
x12	-.199
x14	-.009
x15	55.443
x16	9.845
x17	-16.404
x20	-32.574
x23	.052
(Constant)	-13.493

Unstandardized
coefficients

figura 71

Los centroides de cada grupo, un proceso que depura la base de datos y contempla solo los individuos que tengan toda la información.¹⁰

Functions at Group Centroids

	Function
diagnostic	1
0	-1.265
1	2.130

Unstandardized canonical discriminant
functions evaluated at group means

figura 72

¹⁰Los centroides dan una visión gráfica que permite identificar como estan colocados los grupos y por lo tanto los individuos

Las probabilidades a priori de los grupos la cual se puede manipular dependiendo de si el problema que se trata cuenta con informacion de alguna probabilidad más alta de pertenencia a algún grupo.

Classification Statistics

Classification Processing Summary

Processed		569
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		569

figura 73

Prior Probabilities for Groups

diagnostic	Prior	Cases Used in Analysis	
		Unweighted	Weighted
0	.500	357	357.000
1	.500	212	212.000
Total	1.000	569	569.000

figura 74

La función lineal discriminante de Fisher es importante para la obtención de los puntajes discriminantes para luego poder hacer la clasificacion en el grupo que más semejanza tenga. CASANDRA a diferencia de SPSS muestra las variables que se hayan seleccionado para el análisis multivariado.

Classification Function Coefficients

	diagnostic	
	0	1
X1	80.535	88.179
x2	2.463	2.753
x3	3.475	2.538
x4	-.837	-.859
x5	452.540	476.705
x6	-2487.882	-2522.274
x7	367.637	424.091
x8	-837.777	-798.887
x9	460.388	475.046
x10	15751.569	15842.828
x11	-56.709	-51.854
x12	-1.752	-2.428
x14	.840	.811
x15	3463.719	3651.982
x16	2080.353	2113.783
x17	-234.073	-289.774
x20	-15798.3	-15908.9
x23	.594	.771
(Constant)	-936.653	-983.938

Fisher's linear discriminant functions

figura 75

En la tabla 75 se muestra el porcentaje clasificación correctamente, tanto para las 30 variables como para las variables que SPSS selecciona como no dependientes y de mayor importancia en el sentido de aporte de información para la creación de las funciones discriminantes. En el primer caso ¹¹ hay un resultado muy semejante¹². Para poder hacer una comparación real en cuanto al porcentaje de consistencia del “Cross validation” es necesario comparar lo que ocurriría si se usan las mismas variables.

¹¹Variables originales

¹²Vea figura 49

Classification Results^{b,c}

			Predicted Group Membership		Total
			0	1	
Original	Count	0	357	0	357
		1	17	195	212
	%	0	100.0	.0	100.0
		1	8.0	92.0	100.0
Cross-validated ^a	Count	0	355	2	357
		1	24	188	212
	%	0	99.4	.6	100.0
		1	11.3	88.7	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 97.0% of original grouped cases correctly classified.

c. 95.4% of cross-validated grouped cases correctly classified.

figura 76

III.3 ANALISIS DISCRIMINANTE CON CASANDRA USASNDO LAS VARIABLES QUE SPSS CONSIDERA.

Al hacer uso de las mismas variables de SPSS, CASANDRA llega a una lambda igual a .27992¹³ mientras que SPSS a .270 , ambos valores son bastante semejantes y por lo tanto llegan al mismo resultado: los grupos se encuentran suficientemente separados ¹⁴

Lamda de Wilks	0.26992
Grupos	2
Variables	18
Casos	569
Grados de Libertad L1	18
Grados de Libertad L2	550
F	82.64750
Probabilidad Asociada	0

figura 77

Si se hace el comparativo con las funciones discriminantes, se llega a que son (como se esperaba) muy parecidos los resultados , como por ejemplo: para la variable X1 CASANDRA obtiene 75.95 y el otro paquete un 80.53, para X& un -2,475.94 contra 2,487 , o por último $x^2_3 = 0.32$ contra un .594 de SPSS. En resumen se puede afirmar que se obtiene básicamente el mismo resultado.

¹³CASANDRA lo hace con seis variables

¹⁴Vea la figura 58

III.3 ANALISIS DISCRIMINANTE CON CASANDRA USASND0 LAS VARIABLES QUE SPSS CONSIDER03

Funciones Discriminantes		
Variable	F 0	F 1
x1	75.953169	83.585227
x2	2.494365	2.783458
x3	3.488269	2.53012
x4	-0.771209	-0.792651
x5	791.533949	818.135562
x6	-2.475.94	-2,503.91
x7	347.769514	404.817223
x8	-932.218651	-893.748932
x9	416.809946	431.918384
x10	15,641.93	15,676.25
x11	-31.092076	-26.602023
x12	0.331866	-0.348111
x14	0.703202	0.675078
x15	263.889841	450.880252
x16	2,313.29	2,333.95
x17	-224.841579	-283.176111
x20	-15,823.79	-15,788.06
x23	0.323341	0.502301
Constante	-914.6151464	-958.4943285

figura 78

En el caso de la varianza explicada CASANDRA toma las variables que se le den. Son igules los eigenvalores en este caso porque SPSS al proporcionarle las 30 variables, son esas precisamente las que toma para su calculo. Algunas ocacionesSPSS usa las 30 variables y en otras usa solo las que para el programa considere “importantes” .

Ahora bien, comparando este resultado con el que se tiene con CASNDRA de las 30 variables, se concluye que son prácticamnete iguales.¹⁵

Varianza Explicada			
Eigen Valores	% de Varianza Explicada	Varianza Acumulada	Correlación Canónica
514.793705	99.95%	99.95%	0.99903
0.250975	0.05%	100.00%	0.447911

figura 79

¹⁵Vea figura 50 y 76

En cuanto al porcentaje de consistencia, CASANDRA tiene un 95% mientras que SPSS llega a un 95.4% lo cual indica que ambos paquetes estadísticos se podrían usar indistintamente, pero la diferencia es el control que se puede tener en CASANDRA. Es importante decir que no se deben eliminar las variables solo con un algoritmo sin pensar en las implicaciones que se puede tener.

Consistencia de Clasificación			
Grupo Original	Grupo Asignado		Suma
	F - 0	F - 1	
0	357	0	357
1	26	186	212
Suma	383	186	569
Porcentaje de Consistencia	95%		

figura 80

III.4 ANÁLISIS PARA ASIGNACIÓN DE CRÉDITO

Caso de crédito.

Para un banco que se dedique al otorgamiento de créditos es importante saber a quién se le asignará cierta línea. Existe información de la cual se puede valer, como por ejemplo el “Buró de Crédito” que le da a cada persona un score o calificación, la cual indica que tan buen pagador es. Pero esta información en algunas ocasiones no viene completa, como por ejemplo el hecho de no tener un score asignado, esto se debe a dos razones principales, la primera que el sujeto nunca ha solicitado un crédito y la segunda que el sujeto solo ha tenido crédito en determinados bancos los cuales por política no comparten su información crediticia. Pero aún con un score asignado este podría no ser del todo exacto, ya sea por un mal reporte de los bancos o simplemente por la actualización de la información en el Buró.¹⁶

¹⁶Buró de Crédito es una empresa que ofrece un servicio para agilizar el proceso de evaluación de riesgo y asignación de créditos, fomenta empresas más rentables y procura un manejo transparente de la información entre consumidores e instituciones otorgantes de crédito.

Por ello, la información que maneja Buró de Crédito permite ampliar las oportunidades de acceso al crédito para un mayor número de personas y empresas mexicanas.

Buró de Crédito integra información histórica de los datos generales y el comportamiento crediticio de personas físicas y empresas.

Toda la información es proporcionada por los usuarios* y se conservará en la base de datos durante un plazo de 72 meses o seis años para personas físicas y morales.

La base de datos está resguardada garantizando con ello su confidencialidad e integridad.

La información completa acerca de los clientes, contenida en su historial crediticio, únicamente se presenta en el Reporte de Crédito Especial, documento al que sólo tiene acceso el cliente. Para el caso de los usuarios e instituciones que hacen uso de nuestros servicios, la información es presentada a través del Reporte de Crédito Ordinario y éste no muestra el nombre de los otorgantes de crédito.

La información utilizada en este ejemplo de Análisis de crédito es de una base de datos la cual, considerando el hecho de que muchos de los individuos nunca antes habían solicitado una tarjeta de crédito (por su nivel socioeconómico), como primer método de análisis se usó el de discriminante ya que se tienen clasificados dos grupos:

- (1) Sujetos a los que se les asignará línea de crédito.
- (2) Sujetos a los que NO se les asignará línea de crédito.

Es necesario decir que para que una persona solicite crédito en alguna institución, este deberá aportar datos cualitativos así como cuantitativos, dentro de los primeros se pueden agrupar las variables:

- Estado civil
- Profesión.
- Lugar en donde radica
- Dirección (la cual da un indicativo del tipo de estatus social que tiene).

Para las variables que pertenecen al conjunto de variables cuantitativas se consideran para la solicitud de crédito las siguientes (entre otras):

- Capacidad de pago (lo que la persona declara poder pagar mensualmente).
- Edad.
- Número de carros.
- Ingresos mensuales (serán los ingresos mensuales promedio).

Para esta tesis se consideraron las variables cuantitativas (nota: si se pretende realizar un análisis en el que se incluyan las variables cualitativas se puede realizar un análisis de discriminante logístico). Contemplando entonces estas variables se pretende hacer un primer acercamiento a lo que podría ser una buena clasificación de individuos a los que se les dará un crédito.

Primero se analizarán las variables estadísticas básicas, tal y como se hizo anteriormente.

Estadísticas Básicas					
Variable	No. Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muest. n-1	
capag	2.558	1,051.71	3,964,911.13	1,991.21	
Edad	2.558	34.83	125.87	11.22	
numcar	2.558	0.32	0.27	0.52	
ingremens	2.558	7,539.20	48,507,447.64	6,964.73	

figura 81

En la base de datos se ha etiquetado al grupo de los malos pagadores como M y al de los buenos pagadores como B.

* Usuarios: Bancos, arrendadoras, empresas de financiamiento automotriz, hipotecario, y de bienes en general, tiendas departamentales, empresas comerciales y compañías de servicios (ej. Televisión por cable, telefonía), entre otras.

Estadísticas Básicas				
Variable Indicadora (clas) -> B				
Casos 1957				
Variable	No Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1
capag	1.957	994.57	3.160.721.18	1.777.84
Edad	1.957	35.02	129.04	11.36
numcar	1.957	0.32	0.27	0.52
ingremens	1.957	7.438.03	46.298.895.64	6.804.33

figura 82

Estadísticas Básicas				
Variable Indicadora (clas) -> M				
Casos 601				
Variable	No Observaciones	Media Aritmética	Varianza Muestral n-1	Desv. Est. Muestr. n-1
capag	601	1.237.77	6.547.854.35	2.558.88
Edad	601	34.22	115.29	10.74
numcar	601	0.31	0.27	0.52
ingremens	601	7.868.63	55.646.081.90	7.459.63

figura 83

Si se comparan las medias aritméticas de cada una de los grupos , se nota lo siguiente:

- La capacidad de pago de los individuos que pertenecen al grupo de los buenos pagadores es “extrañamente” menor al del grupo de los malos pagadores
- La edad es muy semejante en los grupos.
- El número de carros es también muy parecido, aunque ligeramente menor el de los malos.
- Nuevamente se presenta un aspecto raro en el comportamiento de los ingresos mensuales ya que los malos pagadores reportan tener (en promedio) ingresos más altos que el de los buenos pagadores.

A diferencia de los análisis anteriores en esta tesis, se puede observar que se tiene una lamda de Wilks muy alta, lo cual puede implicar que los grupos estan muy solapados y que por lo tanto este análisis puede presentar dificultades para separar bien los grupos.

Lamda de Wilks	0.995780234
Grupos	2
Variables	4
Casos	2.558
Grados de Libertad L1	4
Grados de Libertad L2	2.553
F	2.704679
Probabilidad Asociada	0.02843819

figura 84

Como análisis complementario se efectua el analisis de varianza multiple que nos permite probar la hipótesis de igualdad de grupos.

Haciendo pruebas de hipótesis para cada una de las variables se observa que para edad, número de carros e ingresos, se rechaza la hipótesis de diferencia entre las medias de las variables de cada grupo ya que la probabilidad es mayor a 0.05.

Análisis de Varianza las Variables que Intervien en Manova						
Variable	Suma de Cuadrados			Cuadrados Medios		
	Entre Grupos	Dentro de Grupos	Total	Entre Grupos	Dentro de Grupos	
capag	27.194.519.19	10.111.083.231.04	10.138.277.750.23	27.194.519.19	3.955.822.85	
Edad	290.17	321.569.53	321.859.70	290.17	125.81	
numcar	0.02	696.40	696.42	0.02	0.27	
ingremens	85.254.610.46	123.948.289.005.45	124.033.543.615.91	85.254.610.46	48.493.070.82	

Entre Grupos	Grados de Libertad		Estadística F	Probabilidad Asociada	Lamda de Wilks
	Entre Grupos	Dentro de Grupos			
1	1	2.556	6.874554	0.008678	0.997318
1	1	2.556	2.30644	0.124831	0.999098
1	1	2.556	0.001151	0.772876	0.999968
1	1	2.556	1.758078	0.181613	0.999313

figura 85

Para poder colocar a un nuevo individuo son necesarias las Funciones Discriminantes

Funciones Discriminantes		
Variable	F B	F M
capag	0.00022	0.00027
Edad	0.27250	0.26594
numcar	-0.17631	-0.23464
ingremens	0.00009	0.00010
Constante	-5.44988	-6.51783

figura 86

Si pretendemos entonces colocar a un sujeto en alguno de los grupos es necesario evaluar en las funciones discriminantes y asignar a esa persona al grupo que halla presentado mayor probabilidad de pertenencia.

Funciones Discriminantes Evaluadas y Probabilidades de Asignación						
Grupo original	Grupo Asignado	F Máxima	F - B	F - M	Probabilidad (F - B)	Probabilidad (F - M)
B	B	4.949846	4.949846	3.713799	0.774875197	0.225124803
M	B	6.681781	6.681781	5.432451	0.777183903	0.222816097
B	B	0.648263	0.648263	-0.501683	0.759501198	0.240498802
B	B	0.657573	0.657573	-0.405614	0.74329915	0.25670085
M	B	1.525846	1.525846	0.406961	0.753781784	0.246218217
B	B	1.695039	1.695039	0.636769	0.742359864	0.257640136

figura 87

La consistencia de clasificación sólo es de un 77%, y eso puede representar un costo muy alto para el banco, tanto si se pretende otorgar crédito como si se pretende denegar el crédito. De hecho con estas variables se observa que casi todos los individuos que son del grupo de los buenos pagadores se ha clasificado correctamente casi al 100% , pero clasificar mal a esta clase de individuos no representa un “costo” tan importante, como el error que existe al clasificar mal a los individuos que pertenecen al grupo de los malos pagadores. En este caso practicamente todos los individuos del grupo de los malos pagadores fueron clasificados como buenos. lo cual rectifica lo que se encontró en la Lambda de Wilks. El grupo M está trasladado en el grupo de los del grupo B. Es en casos como este que el porcentaje de clasificación puede ser en terminos “absolutos” no tan malo como lo es analizando los grupos uno a uno.

Consistencia de Clasificación			
Grupo Original	Grupo Asignado		Suma
	F - B	F - M	
B	1,956	1	1957
M	600	1	601
Suma	2,556	2	2558
Porcentaje de Consistencia		77%	

figura 88

Continuando con el output de CASANDRA, se puede verificar que la primer función discriminante explica casi el 100% , pero en este caso aunque la función explique dicho porcentaje el lector se habrá percatado que esta , no basta para poder tener una buena regla de clasificación.

Varianza Explicada			
Eigen Valores	% de Varianza Explicada	Varianza Acumulada	Correlación Canónica
0.004273	100.00%	100.00%	0.065230
0	0.00%	100.00%	0.000016

figura 89

Coeficientes de Funciones Canónicas Estandarizadas		
Variable	FCE 1	FCE 2
capag	-0.000005	0.000009
Edad	0.000003	0.000015
numcar	0.000001	-0.000011
ingremens	-0.000002	-0.000001

figura 90

Matriz de Valores de las Variables Estandarizadas (Z)				
Grupo	capag	Edad	numcar	ingremens
B	-0.402626	0.104219	-0.612749	-0.220999
B	-0.025970	-1.411019	-0.612749	0.066162
M	-0.126412	-1.054493	-0.612749	-0.220999
B	0.476237	-1.054493	-0.612749	-0.364580

figura 91

Centroides de las Variables Estandarizadas por Grupo				
Grupo	capag	Edad	numcar	ingremens
B	-0.028696	0.016636	0.003122	-0.014526
M	0.093440	-0.054171	-0.010166	0.047300

figura 92

Valores de las Funciones Canónicas Estandarizadas		
Grupo	FCE 1	FCE 2
B	0.000002	0.000005
B	-0.000007	-0.000002
M	-0.000001	0.000004
B	-0.000001	-0.000014
B	-0.000005	-0.000015
M	-0.000002	-0.000001
B	-0.000005	-0.000005

figura 93

Valores de las Funciones Canónicas en los Centroides Estandarizados		
Grupo	capag	Edad
Centro B	0.000000	0.000000
Centro M	-0.000001	0.000000

figura 94

En los ejemplos anteriores el porcentaje de clasificación resultó ser adecuado, pero en este caso desde los estadísticos básicos, los grupos parecían estar solapados, es decir, no existía una diferencia a “simple vista” suficientemente grande, lo cual quedó demostrado con las pruebas de hipótesis.

En la vida real no siempre se puede aplicar este método para la clasificación de individuos, aunque se tengan los elementos para su uso. En este caso se puede tener errores desde la originación de llenado de la solicitud de crédito hasta problemas con la creación de ciertos campos en la base de datos o inclusive, la captura de la misma.

CONCLUSIONES

El análisis discriminante es una técnica del análisis multivariado que se encarga de clasificar a individuos que provienen de dos o más grupos. Esta técnica necesita dar información precisa de los grupos, es decir, se requiere saber de primera instancia a que grupo pertenecen los individuos, o bien realizar un análisis previo como un cluster (análisis de conglomerados) para saber el número de grupos que se necesite. En algunos casos se tiene información a priori de las probabilidades de pertenencia a un determinado grupo, pero cualquiera que sea el caso es importante decir que aunque existen paquetes estadísticos como SPSS, SAS, CASANDRA, R etcétera que determinan si un individuo pertenece al grupo X, el trabajo de verificar la pertenencia a un grupo a otro se debe validar de acuerdo al interés del investigador y no a un algoritmo del programa. Es por lo anterior que se ha decidido mostrar a CASANDRA como una buena herramienta de cálculo ya que se tiene control total de lo que se pretende analizar.

En el ejemplo de tumor maligno y benigno, se mostró el uso tanto de CASANDRA como de SPSS y se puede ver que se llega a datos muy semejantes, la variante aquí es que con SPSS algunas veces es ambiguo el uso de variables originales o variables con cross validated, sin embargo cuando se corrió el proceso con CASANDRA con las mismas variables de SPSS se llega a resultados casi iguales. El investigador tiene la facilidad de usar uno u otro paquete pero hay que recordar que SPSS no es un paquete de uso libre. Por otra parte CASANDRA actualmente se encuentra instalado en muchas de las máquinas de la facultad de ciencias y está en constante actualización.

El investigador debe saber qué porcentaje de consistencia de clasificación está dispuesto a soportar. Un porcentaje bajo tiene distintos costos que pueden afectar drásticamente el objetivo del uso del análisis discriminante. Es importante decir que el uso del análisis discriminante no debe ser usado para cualquier problema. Dado un problema se busca un modelo que pueda dar respuesta a lo que se busca.

ANEXO 1 : Análisis de Conglomerados

El objetivo básico en el análisis de conglomerados es descubrir los grupos naturales de los elementos (o variables). Este método puede ser usado para la reducción de datos, y de esta manera se pueden definir agrupaciones inesperadas que sugieran relaciones a ser investigadas. Se debe descubrir primero una escala cuantitativa en el, para medir la asociación (similaridad) entre objetos.

Existen distintos tipos de cluters (conglomerados):

Algoritmos jerárquicos: producen dendogramas y comienzan con el cálculo de distancias de cada objeto con los demás, los grupos son formados por un proceso de aglomeración o división. Los de aglomeración parten de los elementos individuales (comienzan siendo uno solo), los grupos cercanos son gradualmente mezclados hasta que finalmente todos los objetos forman un único grupo. En el algoritmo de la división, todos los objetos comienzan en un grupo (parten del conjunto de elementos total), este grupo se divide en dos grupos y estos a la vez se dividen en dos más y así sucesivamente hasta llegar a los elementos individuales, es decir el número de grupos al final es igual al número de objetos que tenemos.

Métodos clásicos de partición (métodos no jerárquicos): Las técnicas no jerárquicas son diseñadas para agrupar puntos, más que variables, en una colección de K clusters (conglomerados). El número de conglomerados K pueden ser especificados en avances o determinados como partes de un procedimiento de cluster. Debido a que una matriz de distancias (similaridades) no tiene que ser determinada y las bases de datos no tienen que estar almacenadas mientras la computadora está ejecutando, los métodos no jerárquicos pueden ser aplicados a muchos conjuntos de datos grandes que el de los métodos jerárquicos. Los métodos no jerárquicos empiezan de cualquier partición de puntos en grupos o empiezan de un conjunto inicial de puntos semilla (seleccionados previamente) los cuales formaran el núcleo del conglomerado. Una forma de comenzar es seleccionar las semillas de puntos de forma aleatoria de entre los puntos o aleatorizar la partición de puntos en grupos iniciales.

Algoritmos jerárquicos. Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en estas distancias. Si todas las variables son continuas la distancia más utilizada en la distancia euclídea entre las variables estandarizadas univariadamente. La distancia de Mahalanobis, no es muy recomendable ya que la única matriz de covarianzas disponible es la de toda la muestra que puede mostrar unas correlaciones muy distintas a las que existen entre las variables dentro de los grupos. Si no estandarizamos, la distancia euclídea dependerá sobre todo de las variables con valores más grandes. Si estandarizamos, estamos dando a priori un peso semejante

a las variables con independencia de su variabilidad original, lo que puede no ser siempre adecuado.

Cuando en la muestra existen variables continuas y atributos el problema se tendrá que abortar con mucho cuidado ya que la variables binarias tomarán valores igual a cero o a uno dependiendo si el atributo esta o no, y al compararlas con variables estandarizadas continuas pueden tener un peso mayor. Para efectos de esta tesis, no se considerarán este tipo de casos ya que se trabajará solo con variables cuantitativas.

Las técnicas jerárquicas de conglomerados proceden de cada una de las series de una unión sucesiva o de una serie de divisiones sucesivas. Los métodos de *aglomeración Jerárquica* comienzan con objetos individualizados. De esta manera hay tantos conglomerados como objetos (elementos). Los objetos más similares son agrupados primero y estos grupos iniciales son unidos de acuerdo a sus similitudes. Eventualmente, a la vez que las similitudes decrecen, todos los grupos son fusionados en un único conglomerado.

Los métodos de división jerárquica trabajan en forma opuesta. Un solo grupo inicial de objetos es dividido en dos grupos tal que los objetos en un subgrupo estén “lejos de” los objetos en el otro. Estos subgrupos son divididos más adelante en grupos no similares, el proceso continua hasta que son tantos grupos como objetos, es decir hasta que cada objeto forma un grupo.

El resultado de ambos métodos aglomerativos y divisivo pueden mostrarse en la forma de un diagrama de dos dimensiones (dendograma). El dendograma ilustra las uniones o divisiones que han sido hechas en niveles sucesivos.

Los métodos de enlace (linkage methods) son apropiados para conglomerados de elementos, como para variables. Esto no es cierto para todos los procedimientos de aglomeración jerárquica.

La unión simple se da cuando los grupos son fusionados de acuerdo a la distancia entre sus miembros más cercanos. La unión completa (complete linkage) se presenta cuando los grupos son fusionados de acuerdo a la distancia entre sus más lejanos miembros. Para la unión promedio (average linkage) o unión media los grupos son fusionados de acuerdo a la distancia promedio entre las parejas miembros en sus respectivos conjuntos.

Los pasos para aglomeración jerárquica de conglomerados para grupos de N objetos (elementos o variables) son:

- (1) Comenzar con N conglomerados, cada uno contiene una sola entidad y una matriz de distancias (similitudes), dada por:

$$D = \{d_{ik}\}$$

- (2) Buscar la matriz de distancias para las parejas de conglomerados (clusters) mas cercanos (más similares). La distancia entre los conglomerados y “más similares” es d_{uv}

$$d[C; (UV)] = \min(d_{CU}, d_{CV})$$

- (3) Unir los conglomerados U y V .Etiquetar el conglomerado recientemente formado por (UV) . Actualizar las entradas de la matriz de distancias de la unión simple borrando los renglones y columnas correspondientes a los conglomerados U y V para la unión completa agregar una columna dando las distancias entre el conglomerado (UV) y los conglomerados que quedan.
- (4) Repetir los pasos 2 y 3 un total de $N - 1$ veces (todos los objetos deben ser un conglomerado simple después de terminar el algoritmo).Registrar la identidad de los conglomerados que son unidos y los niveles (distancias o similitudes) para los cuales las uniones toman lugar.

El método de unión simple (encadenamiento simple o vecino más próximo) tiene el inconveniente de que no puede discernir o identificar conglomerados mal separados.

En el método de la unión completa (encadenamiento completo o vecino mas alejado). La distancia entre los nuevos grupos es la mayor de las distancias entre grupos antes de la fusión.

Este método sigue siendo de la misma manera que el anterior, con la diferencia de que en cada estado, la distancia entre los conglomerados esta determinada por la distancia entre los elementos, uno de cada conglomerado, que son los mas distantes. Por lo tanto la unión completa asegura que todos los puntos en el conglomerado son de distancia máxima (o mínima similitud) uno a otro.

El algoritmo general aglomerativo comienza de nuevo encontrando la mínima entrada en:

$$D = \{d_{ik}\}$$

Y si unimos los correspondientes objetos como U y V , los cuales serán los conglomerados (UV) . Para el paso 3 del algoritmo general la distancia entre (UV) y cualquier otro cluster es calculada de la siguiente manera:

$$d_{(UV)W} = \max(d_{UW}, d_{VW})$$

Aquí d_{UW} y d_{VW} son distancias entre los miembros del conglomerado U y V respectivamente.

Como ocurre en la unión simple, una “nueva” asignación de distancias que tienen el mismo orden relativo como la distancia inicial no cambiará la configuración de la unión completa.

La unión promedio considera la distancia entre dos conglomerados como el promedio de la distancia entre todos los pares de puntos donde un miembro de una pareja pertenece a cada conglomerado. El método puede ser usado para agrupar objetos (elementos) o variables. El algoritmo de la unión promedio precede en la manera general. Empezamos buscando la matriz de distancias $D = \{d_{ik}\}$ para encontrar los objetos más cercanos por ejemplo U y V . Estos objetos son fusionados para formar el cluster (UV) . Por el paso 3 del algoritmo general aglomerativo las distancias (UV) y los otros clusters W son determinados por:

$$d_{(U,V)W} = \frac{\sum_i \sum_k d_{ik}}{N_{UV}N_W}$$

Donde d_{ik} son las distancias entre el objeto i en el cluster UV y el objeto k del cluster

N_{UV} es el número de puntos (elementos) en el cluster (UV) y

N_w número de puntos en el cluster W

Algoritmos NO jerárquicos. Dentro de estos algoritmos encontramos al Método de las K – medias, y recibe ese nombre porque hace referencia a que cada cluster tiene un centroide más cercano (media).

Este algoritmo es:

- (1) Particionar los puntos en K conglomerados iniciales.
- (2) Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clusters y se reasignan a los que estén más próximos. Se vuelven a re-calcular los centroides de los k clusters después de las reasignaciones de los elementos.
- (3) Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

Contabilidad Multigrupos																																				
	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31				
1			
2		
...	
Total	12.471	4.938	88.487	122.402	508	324	910	12	374	243	308	102	328	917	2.441	1.632	24.005	22.638	1.919	1.342	1.717	1.824	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968	117.968

En el 31 de diciembre de 2010 se han dejado de tener.

figura A_2

Bibliografía

- (1) ANDERSON,T.W (1984). AN INTRODUCTION TO MULTIVARIATE ANALYSIS. NEW YORK WILEY
- (2) EVERIT, B.S Y DUNN, G(2001). APPLIED MULTIVARIANTE DATA ANALYSIS EDWARD ARNOLD, LONDON.
- (3) FAHRMEIR, LUDWING J.(2004). MULTIVARIATE STATISTICAL METHODS. CHAPMAN & HALL/CRD.
- (4) FLORES, J GIL (2001). ANÁLISIS DISCRIMINANTE CUADERNOS DE ESTADÍSTICA. HESPÉRIDES.
- (5) HAIR,J.F. ANDERSON, R.E.,.(1999). ANÁLISIS MULTIVARIANTE.PRENTICE-HALL, MADRID .
- (6) HUBERTY,C.J.(1994). APPLIED DISCRIMINANT ANALYSIS. WILEY. INTERSCIENCE.
- (7) JIMENEZ ,E.URIEL.(1995). ANALISIS DE DATOS:SERIES TEMPORALES Y ANÁLISIS MULTIVARIANTE. AC, MADRID.
- (8) JOHNSON RICHARD A.& WICHERN DEAN (1998). APPLIEDMULTIVARIATE STATISTICAL ANALYSIS. PRENTICE-HALL
- (9) JOHNSON,D.E.(2000). MÉTODOS MULTIVARIADOS APLICADOS AL ANÁLISIS DE DATOS. INTERNATIONAL THOMSON EDI
- (10) MARDIA,K.V. (1995). MULTIVARIATE ANALYSIS. ACADEMIC PRESS.
- (11) PEÑA,D(2002).ANÁLISIS DE DATOS MULTIVARIANTES. MCGRAW-HILL.
- (12) TINSLEY HOWARD E. A.. HANDBOOK OF APPLIED MULTIVARIATE STATISTICS AND MATHEMATICAL MODELING.
- (13) VICENTE Y OLIVA (1999). ANÁLISIS MULTIVARIANTE PARA LAS CIENCIAS SOCIALES. DYKINSON,S.L
- (14) <http://portal.acm.org/citation.cfm?id=1273633>.
- (15) http://www.ucm.es/info/socivmyt/paginas/D_departamento/materiales/
- (16) http://www.ugr.es/~bioestad/_private/cpfund8.pdf .
- (17) http://www.ugr.es/~bioestad/_private/cpfund8.pdf .
- (18) <http://www.seh-lelha.org/clasifica.htm#STEPWISE>
- (19) http://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF
- (20) <http://www.psychstat.missouristate.edu/multibook/mlt03.htm>
- (21) <http://www.dtrek.com/lda.htm>
- (22) <http://www.ece.osu.edu/~aleix/pami06.pdf>
- (23) <http://www2.uca.es/serv/ai/formacion/spss/Imprimir/21conglk.pdf>

- (24) http://es.geocities.com/r_vaquerizo/Manual_R11.htm
- (25) <http://www.emis.de/journals/RCE/V30/v30n2a06PardoDelCampo.pdf>
- (26) <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema5am.pdf>
- (27) <http://marketing.byu.edu/htmlpages/tutorials/discriminant.htm>
- (28) <http://portal.acm.org/citation.cfm?id=1273633>
- (29) http://www.ucm.es/info/socivmyt/paginas/D_departamento/materiales
- (30) http://www.ugr.es/~bioestad/_private/cpfund8.pdf
- (31) http://www.ugr.es/~bioestad/_private/cpfund8.pdf
- (32) <http://www.seh-lelha.org/clasifica.htm#STEPWISE>
- (33) http://www.uam.es/personal_pdi/economicas/rmc/documentos/cluster.PDF
- (34) <http://www.psychstat.missouristate.edu/multibook/mlt03.htm>
- (35) <http://www.dtreg.com/lda.htm>
- (36) <http://www.ece.osu.edu/~aleix/pami06.pdf>
- (37) <http://zonecours.hec.ca/documents/200589.AnalyseDiscriminante.ppt> #257,2,
Qu'est-ce qu'une analyse discriminante?
- (38) <http://geo.polymtl.ca/~marcotte/glq3402/chapitre5.pdf>
- (39) http://www.aiaccess.net/French/Glossaires/GlosMod/f_gm_analyse_discriminante.htm
- (40) <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema6am.pdf>
- (41) <http://www.pilando.com/categorias/administracion/ASIG.5%5B1%5D.doc>