FACULTAD
DE CIENCIAS
ECONÓMICAS

Universidad
Nacional
de Córdoba

Métodos estadísticos para la clasificación de observaciones multivariadas:

Un caso de aplicación:

Regionalización económica de la provincia de Córdoba

Tesis de Doctorado

Goldenhersch, Hebe Susana

1975

METODOS ESTADISITICOS PARA LA CLASIFICACION DE OBSERVACIONES MULTIVARIADAS:

Un caso de aplicación: regionalización económica

De la provincia de Córdoba

Hebe Susana Goldenhersch

Trabajo de Tesis

Córdoba, Noviembre de 1975

METODOS ESTADISTICOS PARA LA CLASIFICACION DE OBSERVACIONES MULTIVARIADAS

Un caso de aplicación: regionalización económica de la Provincia de Córdoba.

Hebe Susana Goldenhersch

Trabajo de Tesis

Córdoba, Noviembre de 1975

Agradecimientos:

A mis compañeros del Instituto de Econometría y Estadística y del Departamento de Estadística y Matemática; a todo el personal (sin excepción) del Centro de Computación y Procedamiento de Datos de la Facultad por haber colaborado toda vez que fué necesario en la solución de problemas referidos a los programas de cómputo; el Lic. José Cocilovo con quien intercambiamos bibliografía y comentarios acerca de los métodos multivariados.

I N D I C E

INTRODUCCION

El problema que encaramos es el de clasificar, formando grupos homogéneos, un conjunto de observaciones caracterizadas por los valores que en cada una de ellas asumen varias variables.

Existen métodos estadísticos muy desarrollados y experimentados en los más diversos campos para evaluar este tipo de clasificaciones cuando la caracterización de una observación viene dada por una sola variable (poblaciones univariadas). No ocurre otro tanto cuando las observaciones provienen de poblaciones multivariadas.

En biología, antropología y ciencias de la conducta, la cuestión ha sido encarada mediante técnicas estadísticas (análisis discriminante) o numéricas (taxonomía numérica o "cluster" análisis).

En el terreno de las ciencias sociales, sin embargo, se recurre por lo general a métodos intuitivos que suministran clasificaciones no óptimas.

En este trabajo, se trata de aplicar algunas de las técnicas de discriminación y taxonomía, adaptándolas a problemas económicos.

1. Definiciones

Para definir más concretamente la cuestión que nos ocupa, transcribimos un párrafo de M.G. Kendall (1965) que diferencia tres tipos de problemas de naturaleza parecida:

Discriminación: tenemos una muestra de miembros provenientes de cada una de dos poblaciones. Se determina para cada miembro los valores de cierto número p de variables. Se requiere: construir un método para asignar un nuevo miembro a la población correcta sobre la base de los valores que presenta de las p variables. Como una extensión, podemos tener los datos para k poblaciones y requerir discriminar entre ellas. Un problema importante, que no se ha considerado mucho, deviene cuando alguna o todas las variables son

iv) se desea jerarquizar las poblaciones urbanas en grupos, teniendo en cuenta las variables económico-sociales que caracterizan a cada una de ellas. Es el caso que Kendall llama "disección".

1.1. Particularidades de los problemas planteados en economía u otras ciencias sociales

i. Una primera cuestión radica cuando se trata de "zonificar" una región o país; será necesario, o por lo menos conveniente, que las observaciones (provincias, departamentos u otro tipo de unidad de observación) que forman una zona estén geográficamente unidas. Este hecho no es contemplado por las técnicas de clasificación habitualmente aplicadas en las ciencias biológicas.

ii. En biología, los individuos se clasifican sólo de una o muy pocas maneras posibles, perfectamente definidas. No ocurre lo mismo en el campo económico social, donde coexisten distintos criterios para una clasificación, y la selección de variables es más o menos arbitraria. Este es, posiblemente el aspecto más importante, porque con las mismas observaciones pueden lograrse distintas agrupamientos utilizando diversas variables y es posible, como ocurre en muchos casos, realizar una selección de variables que no incluya las más relevantes para una clasificación adecuada a los requerimientos del problema.

2. Las técnicas posibles

2.1. Métodos numéricos se han desarrollado especialmente desde que la computación ha suministrado la posibilidad del manejo de gran cantidad de información.

Estos métodos presentan el inconveniente de asignar igual importancia a todas las variables intervinientes, sin brindar la posibilidad de señalar las que más contribuyen a la separación entre los grupos. No se comienza con una clasificación a priori, sino con el conjunto de observaciones. Los grupos se forman únicamente de acuerdo a los valores numéricos de las variables, sin tener en cuenta la importancia que puedan tener algunas de ellas para la clasificación en cuestión, ni tampoco la necesidad de formar gru-

pos geográficamente compactos.

El efecto suele ser la formación de grupos en los que es difícil o imposible determinar cuáles son las variables que han contribuído en mayor medida a su formación, dándose el caso en que alguna de poca importancia para el tipo de clasificación deseada pesa igual que otra muy importante.

Es decir que se carece de un modelo teórico previo acerca de los grupos, lo cual representa un serio peligro metodológico similar al que se corre cuando se encara una investigación haciendo regresiones con los datos, para "ver qué ocurre" con ellos.

Por otra parte, la ventaja de los métodos numéricos que permiten manejar numerosas variables, resulta a veces una desventaja cuando se trata de cuestiones económicas, en las que es necesario seleccionar variables de real importancia y no tan numerosas (se introduce "ruido").

Sokal (1963) presenta un cuadro bastante completo de las técnicas de taxonomía numérica; en general, las diferencias radican solo en la manera en que se define una "distancia" entre cada par de elementos, y luego en determinar en qué parte se hacen los "cortes" entre grupos. El estudio de Naciones Unidas (1972) clasifica los países de América Latina según variables económicas y sociales utilizando técnicas de taxonomía, complementadas con análisis factorial.

Varsavsky (1969), Aracx (1968), Wallace (1968) y Boulton (1968) presentan técnicas numéricas que relacionan con la teoría de la información. Plantean estos autores diversos métodos para preparar y mejorar clasificaciones.

Estos, sin embargo, mantienen el problema metodológico indicado más arriba con respecto a los métodos numéricos en general.(*)

(*) Kendall (1965) y (1967) propone un método de "libre distribución" que tiene mucha semejanza con las puramente numéricos, y que hemos experimentado con nuestros datos, manteniéndose las dificultades planteadas.

2.2. Métodos derivados de la estadística multivariada

2.2.1. Consideraciones históricas

El análisis multivariado, como lo señala Kendall (1972)* puede considerarse iniciado con la publicación de Wishart en 1928 sobre la distribución de su mismo nombre (la distribución de las varianzas y covarianzas de p variables normales). La matriz de varianzas y covarianzas resultaba una extensión natural del caso univariado de la varianza. A partir de allí, se desarrollan tres líneas en el análisis multivariado: la primera, el estudio de razones de matrices de dispersión, arribando a generalizaciones multivariadas de la t de Student, el análisis de varianza, tests de regresión, etc., llegando al test F de razón de varianzas (Wilks avanzó mucho en ese sentido).

Una segunda línea fué el estudio de la diferencia de matrices del tipo (A-λB) llegando al tratamiento de componentes principales.

Una tercera línea, fué la medición de distancias entre poblaciones p-variadas y asociado a ella, el estudio de funciones discriminantes.

Con respecto a esta última, su origen puede buscarse aún antes de los trabajos de Wishart, cuando Karl Pearson, alrededor de 1920 definió una "medida de la distancia" entre dos poblaciones; Mahalanobis propuso una alternativa en 1925, la D^2, que resultó más eficiente para cumplir con el objetivo que la de Pearson. Este fué el punto de partida de la escuela hindú, que realizó y continúa realizando importantes aportes en este campo.

En 1931, Hotelling generalizó la t de Student para el caso multivariado, resultando algo muy similar a la D; la relación entre ambos estadísticos es:

$$T^2 = \frac{D^2 \, n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}$$

(*) pág. 105.

y su objeto, igual que el de la D^2 fué el de permitir que se efectúen tests acerca de la diferencia de medias en poblaciones normales multivariadas.

En 1936, Fisher publicó su primer trabajo sobre funciones discriminantes: la diferencia más importante entre su enfoque y el de Mahalanobis fué que en lugar de medir distancias, Fisher pretendió dividir el espacio muestral en regiones y adjudicar cada observación muestral a una u otra población según la región en que estuviere. Se aproximaba así, a la futura teoría de la decisión.

Sin embargo, se siguió trabajando sobre la D^2 que resultaba muy útil ya que además de cumplir el objetivo con que se comenzó a investigar, permite completar el análisis discriminante. Es así que el mismo Fisher, en 1938 generalizó la D^2 para más de dos poblaciones. Se produce entonces una interrupción en el desarrollo de las investigaciones en este sentido, dedicándose la escuela inglesa al estudio de las raíces características y correlaciones canónicas vinculadas a ellas. Por este camino se llega también a nuevas aplicaciones de funciones discriminantes. Son importantes en este sentido, las contribuciones del hindú Rao, y de la escuela anglosajona (Bartlett, Williams).

Debemos remarcar que la mayoría de los trabajos de aplicación de funciones discriminantes se refieren a biología, antropología o ciencias de la conducta. En economía, podemos mencionar los trabajos de Tintner (1946) para dos grupos, K. Gales (1957) para discriminar entre clases sociales y los de Irma Adelman y Cinthia T Morris(*) que aplican el análisis factorial, la función discriminante y otros métodos multivariantes con el fin de plantear un modelo econométrico de cambio político y social en los países subdesarrollados. Alrededor de estos trabajos se desarrolla una polémica metodológica importante (Rayner, Berry, Eckstein

(*) Adelman y Morris (1968-a, 1968-b, 1970-a 1970-b)
Rayner (1970) Berry y Eckstein en Adelman y Morris (1970-a)

critican varios aspectos. Nos remitimos a los artículos citados, no obstante más adelante -Capítulo 5- haremos referencias a algunos puntos en discusión).

En nuestro país, existen los trabajos de Serrato (1969) que aplica el método de Bartlett para obtener regiones homogéneas en el centro-litoral argentino; y Kaminsky (1971) que trabaja para dos grupos a efectos de clasificar explotaciones lecheras en la misma región.

Se han realizado interesantes aplicaciones en geografía: Berry (1965), King (1965), Casetti (1964), Brown y Trott (1968) quienes plantean la utilización de todos los métodos mencionados para obtener regiones homogéneas. Agregan también el análisis factorial.

2.2.2. Análisis factorial, de componentes principales o función discriminante?

Varios autores, (Rogers (1971), Berry (1965), Kendall (1939), Brown y Trott (1968), especialmente en materia de geografía, proponen el uso de componentes principales para formar las zonas homogéneas. Se trata de encontrar "componentes", que son combinaciones lineales entre las variables, las que explican la mayor parte de la varianza de las variables originales.

En base a estas componentes, se forman matrices de similaridades, y se representan en una o dos dimensiones las observaciones, formándose de esa manera los grupos homogéneos. Berry (1965), luego de constituir de esta manera los grupos, plantea el análisis discriminantes para observaciones dudosas, y lo realiza entre cada par de grupos a fin de asignarlas definitivamente.

El análisis factorial, usado por estos mismos autores como técnica alternativa, se propone cierta hipótesis lineal y luego estudia si ese modelo explica satisfactoriamente las correlaciones originales (Rogers, (1971))(*)

(*) pág. 481.

Se mantiene en las dos técnicas señaladas la falta del modelo inicial, agregándose la dificultad para dar una interpretación satisfactoria de esas "combinaciones lineales" de variables originales.

La función discriminante elimina ese problema, el plantearse previamente la formación de grupos, buscando luego la función o combinación lineal de variables adecuada para incorporar nuevas observaciones. Hemos extendido su aplicación utilizando las funciones obtenidas para mejorar la clasificación original, reasignando las observaciones por un procedimiento iterativo que mejora paso a paso la clasificación sin romper el esquema o modelo original. La obtención de funciones canónicas estandarizadas permite, por otrapparte, ir determinando cuáles son las variables que más han contribuído a la separación entre grupos, pudiéndose en esta etapa eliminar algunas que aunque se manifiestan con poder discriminante resultan irrelevantes para el objetivo de la clasificación.

En el próximo capítulo hacemos un análisis detallado de este método, del que luego presentamos aplicaciones. Digamos por ahora, que, la función discriminante introducida por Fisher proporciona una clasificación óptima en un espacio de igual dimensión el número de variables consideradas, dando lugar a un método sencillo para clasificar nuevas observaciones. En efecto, se obtiene una función para cada grupo y solo es necesario asignar la observación al grupo correspondiente a la función para la cual asumió el mayor valor. Incluso, aceptando los supuestos previos acerca de distribución de las variables, pueden calcularse las probabilidades de pertenencia a cada grupo.

La obtención posterior de las funciones canónicas, reduce la dimensionalidad, permite la representación gráfica del conjunto de observaciones, y como ya mencionamos, el análisis de las variables que más contribuyen a la discriminación.

Este último aspecto presenta semejanza con el método de componentes principales o de factores, pero en el análisis discri-

minante no es necesario "descubrir" cuál es el significado real de las combinaciones lineales entre variables, lo cuál es bastante subjetivo, sino que directamente interesa observar cuál o cuales son las variables que separan mejor los grupos.

Por último, señalemos que entre las técnicas utilizadas para el análisis de datos multivariados, se habla de técnicas "Q" y "R" o de matrices de correlaciones "Q" y "R". Las primeras son aquellas que consideran las correlaciones o "distancias" entre las observaciones (individuos); en esta categoría entra la función discriminante; las segundas, consideran las correlaciones entre variables, en ellas ubicamos las componentes principales y el análisis factorial.

3. La función discriminante

3.1. La función discriminante de Fisher para dos poblaciones

3.1.1. Caso general

Se tienen dos poblaciones P_1 y P_2 en el espacio p-variado.(*) A cada población corresponde un vector de medias $\mu_i (i=1,2)$ de p componentes y matriz de varianzas y covarianzas $\Sigma_i (i=1,2)$. Cada observación está caracterizada por un vector X de p componentes (el valor de cada una de las variables en esa observación). Llamamos W al espacio en el cual están definidas ambas poblaciones, y en el que a cada vector le corresponde un punto. Se trata de dividirlo en dos regiones (R y W-R) de modo que en R estén la mayor parte de los puntos de P_1 y en W-R la mayor parte de los puntos de P_2. Suponemos también que $f_1(x)$ y $f_2(x)$ representan las densidades multivariadas de P_1 y P_2 (suponemos variables continuas, sin que ésto implique pérdida de generalidad).

(*) Si se tratara de una sola variable, el problema se reduce al resuelto por la teoría de la decisión: fijar un valor tal que toda observación que lo supere, se asigna a P_2; si no lo supera, se asigna a P_1.

Si las dos poblaciones no presentarán superposición alguna, sería sencillo adjudicar cada observación a la que corresponde; sin embargo, existe superposición (las poblaciones no son conjuntos disjuntos), de allí que ciertas observaciones puedan pertenecer a una u otra; se requiere entonces una regla para decidir a qué población se adjudican. La función discriminante proporciona esa regla, tratando de minimizar la probabilidad de clasificar mal. Esta regla es proporcionada por una función de las variables, que se trata sea lo más simple posible (de allí que en general se trabaje con funciones lineales).

Se trata entonces de buscar un límite que determine en el espacio W una región R tal que:

$$\int_R f_2(x)dx = \int_{W-R} f_1(x)dx = 1 - \int_R f_1(x)dx \qquad (1)^*$$

Esto, suponiendo que es igual en cuanto a costo, la importancia del error de asignar a P_1 una observación perteneciente a P_2 y viceversa.

La condición (1) implica que:

$$\int_R [f_1(x)+f_2(x)]dx = 1 \qquad (2)$$

Por otra parte, se desea minimizar ambos tipos de error, lo cual equivale a minimizar uno de ellos; es decir:

$$\int_R f_2(x)dx = \text{mínimo} \qquad (3)$$

Se minimiza (3) sujeto a (2):

$$\int_R \left[f_2(x) - \lambda(f_1(x)+f_2(x)) \right] dx \quad \text{ó}$$

(*) Las integrales son múltiples, las funciones son de vectores así como las diferenciales

$$\int_R (\beta f_2(x) - f_1(x))dx \qquad (4)$$

El mínimo de esta función se logrará tomando en R todos los puntos para los que $(\beta f_2(x) - f_1(x)) < 0$; de este modo, la cota de R será la recta determinada por:

$$\frac{f_1(x)}{f_2(x)} = \beta \qquad (5)$$

Lo que representa una razón de verosimilitud. Se adjudicará entonces una observación a P_1 si

$$\frac{f_1(x)}{f_2(x)} > \beta \qquad (6)$$

y viceversa.

La probabilidad de clasificar mal una observación será:

$$\int_{\frac{f_1(x)}{f_2(x)} > \beta} f_2(x)dx = \int_{\frac{f_2(x)}{f_1(x)} > \beta} f_1(x)dx \qquad (7)$$

y el valor de β dependerá de las probabilidades que cada observación tiene de pertenecer a una u otra población. Si estas probabilidades son desconocidas, y se suponen iguales, entonces β será igual a 1 y se adjudicará a P_1 si $f_1(x) > f_2(x)$.

Si las probabilidades de pertenencia de cada observación a P_1 y P_2 son conocidas (sean q_1 y q_2 respectivamente), entonces la probabilidad de que una observación pertenezca a P_1 y esté en la región R, será:

$$\int_R q_1 f_1(x)dx \qquad (8)$$

Definimos también la probabilidad condicional de que una observación pertenezca a P_1 dado su vector de valores X. Lo cual, de acuerdo a la definición de probabilidad condicional y a la fórmula de Bayes, resulta:

$$Pr(P_1/X) = \frac{q_1 f_1(x)}{q_1 f_1(x)+q_2 f_2(x)} \qquad (9)$$

La probabilidad de clasificar mal una observación, teniendo en cuenta los dos tipos de errores posibles, es:

$$q_1 \int_{W-R} f_1(x)dx + q_2 \int_R f_2(x)dx \qquad (10)$$

Si queremos minimizar esta probabilidad, será necesario asignar la observación a la población que tiene mayor probabilidad condicional; si

$$\frac{q_1 f_1(x)}{q_1 f_1(x)+q_2 f_2(x)} > \frac{q_2 f_2(x)}{q_1 f_1(x)+q_2 f_2(x)} \qquad (11)$$

se clasificará en P_1, de lo contrario en P_2. (En caso de igualdad, puede asignarse arbitrariamente). Como los denominadores son iguales, la regla resulta:

Adjudicar a P_1 si: $q_1 f_1(x) > q_2 f_2(x)$

Adjudicar a P_2 si: $q_1 f_1(x) < q_2 f_2(x)$

En caso de igualdad, adjudicar a cualquier población. Se llega a la misma conclusión que en (6), haciendo $s = q_2/q_1$.

Anderson (1958)(*) explica cómo estos procedimientos de

(*) Cap. 6.

decisión constituyen procedimientos "admisibles", esto es, que no pueden ser superados por otro procedimiento posible.

3.1.2. Para dos poblaciones normales

3.1.2.1. Parámetros poblacionales conocidos

Supongamos ahora que $f_1(x)$ y $f_2(x)$ son normales multivariadas, con iguales matrices de varianzas y covarianzas. Es decir:

$$f_1(x) \sim N(\mu_1, \Sigma)$$

$$f_2(x) \sim N(\mu_2, \Sigma)$$

resultando entonces la función de densidad conjunta de las p variables para la población P_i $(i = 1, 2)$:

$$f_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu_i)' \Sigma^{-1} (X-\mu_i)\right] \quad (12)$$

y la razón de verosimilitud:

$$\frac{f_1(x)}{f_2(x)} = \frac{\exp\left[-\frac{1}{2}(X-\mu_1)' \Sigma^{-1} (X-\mu_1)\right]}{\exp\left[-\frac{1}{2}(X-\mu_2)' \Sigma^{-1} (X-\mu_2)\right]} =$$

$$= \exp^{-1/2}\left[(X-\mu_1)' \Sigma^{-1} (X-\mu_1) - (X-\mu_2)' \Sigma^{-1} (X-\mu_2)\right] \quad (13)$$

La región R que contendrá las observaciones asignadas a P_1, estará formada por el conjunto de X para los que (13) sea mayor que una constante k.

Esto puede expresarse con mayor claridad tomando el logaritmo de (13) por tratarse de una función monótona creciente; en cuyo caso, se adjudicará a P_1 si:

$$-\frac{1}{2}\left[(X-\mu_1)'\,\Sigma^{-1}(X-\mu_1)-(X-\mu_2)'\,\Sigma^{-1}(X-\mu_2)\right]>\log\beta \qquad (14)$$

Desarrollando el primer miembro de (14) y agrupando, resulta:

$$U=X'\,\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)>\log\beta$$

Como se observa, U resulta una función lineal de las componentes del vector de observaciones, menos una constante. A esa función lineal se la denomina "función discriminante".

Si las probabilidades q_1 y q_2 son conocidas, entonces

$$\beta=\frac{q_2}{q_1}$$

y si son iguales, $\beta = 1$ y $\log\beta = 0$

La regla de decisión será

Adjudicar a P_1 si:

$$X'\,\Sigma^{-1}(\mu_1-\mu_2)>\frac{1}{2}(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \text{ ó } U>0$$

Adjudicar a P_2 si:

$$X'\,\Sigma^{-1}(\mu_1-\mu_2)<\frac{1}{2}(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2) \text{ ó } U<0$$

Arbitrariamente, si U es igual a 0.

De manera que para construir la función discriminante, se calcula $\Sigma^{-1}(\mu_1-\mu_2)$ lo cual proporciona los coeficientes de la combinación lineal con el vector X y luego la constante:

$$\frac{1}{2}(\mu_1+\mu_2)'\,\Sigma^{-1}(\mu_1-\mu_2)$$

Para cada observación, se calcula el "valor discriminante" y según el resultado, se adjudica a P_1 o P_2.

Si las probabilidades q_1 y q_2 son desconocidas y no se considera conveniente suponerlas iguales, se trata de elegir β igualando los costos por mala clasificación.

Anderson (1958)* obtiene la distribución de U, encontrando que es normal con

$$E(U) = \frac{D^2}{2} \quad y \quad V(U) = D^2$$

Siendo $D^2 = (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$ si $X \sim N(\mu_1, \Sigma)$

Si $X \sim N(\mu_2, \Sigma)$, entonces $E(U) = -\frac{D^2}{2}$ y $V(U) = D^2$

Sobre el importante significado de esta D^2 volveremos más adelante.

El conocimiento de la distribución de U permite calcular las probabilidades de clasificar mal una observación.

3.1.2.2. Parámetros poblacionales desconocidos

En caso de desconocer los vectores medios y las matrices de varianzas y covarianzas, sus mejores estimadores resultan:

$$\mu_i^* = \bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ji}}{n_i} \qquad i = 1, 2$$

$$\Sigma^* = S = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'\right]$$

Se obtiene así la función:

$$V = X'S^{-1}(\bar{X}_1-\bar{X}_2) - \frac{1}{2}(\bar{X}_1+\bar{X}_2)'S^{-1}(\bar{X}_1-\bar{X}_2) \qquad (15)$$

El primer término de (15) es la función discriminante obtenida por Fisher en 1936. Esta función, tiene la propiedad de maximizar, entre todas las combinaciones lineales de las componentes de cada vector, la varianza entre las muestras de ambos grupos con respecto a la varianza dentro de esas muestras.

La distribución de V, tiende a la de U. Para muestras chicas, su distribución es bastante más complicada.(*)

3.1.2.3. Evaluación de una función discriminante.

En el curso del análisis discriminante, pueden plantearse dudas acerca de si existe diferencia significativa entre ambos grupos. En realidad, el caso más frecuente no es la duda sobre si se trata de una sola población sino que a veces las variables seleccionadas no son las más adecuadas para discriminar entre los grupos.

El camino más obvio para evaluar dichas diferencias es la realización de un test sobre igualdad de medias.

Como generalización natural de la t deSStudent, hemos mencionado ya el test T^2 de Hotelling.

Sin embargo, por la posibilidad de extensión para más de dos grupos y por su obtención en el curso de los cálculos para obtener la función discriminante, hemos preferido el estadístido D^2 de Mahalanobis, que mide la "distancia" entre dos poblaciones multivariadas:

$$D_p^2 = (\bar{X}_1 - \bar{X}_2)' \, S^{-1} (\bar{X}_1 - \bar{X}_2) \qquad (16)$$

donde p es el número de variables, y el resto de los elementos tienen el significado ya explicado en el punto anterior.

Rao (1952)(**) deduce la distribución exacta de D^2, que

(*) ver Anderson pp. cit.

(**) pág. 70.

como ya citáramos en 2.2.1., se relaciona con T^2 de la siguiente forma:

$$T^2 = \frac{D^2 \, n_1 n_2}{(n_1+n_2)(n_1+n_2-2)} \quad (n_1 \text{ y } n_2 \text{ son los tamaños muestrales})$$

Plantea también que el estadístico:

$$\frac{n_1 n_2 (n_1+n_2-p-1)}{p(n_1+n_2)(n_1+n_2-2)} \, D^2 \qquad (17)$$

puede ser usado como razón de varianzas con p y (n_1+n_2-p-1) grados de libertad, bajo la hipótesis de igualdad de medias.

Si llamamos d al vector de coeficientes de la función discriminante obtenida anteriormente (primer término de (15)), se tiene:

$$d = S^{-1}(\bar{X}_1 - \bar{X}_2)$$

luego:

$$D^2 = (\bar{X}_1 - \bar{X}_2)' \, d$$

es decir, que D^2 puede obtenerse en el transcurso de la construcción de la función discriminante.

Al tratar el caso general, de varios grupos o poblaciones, veremos otros test sobre igualdad de medias.

3.1.3. Función discriminante y teoría de la información

Varios autores, entre ellos Kullbak (1959), plantean el problema de la clasificación utilizando los elementos de Teoría de la Información. Veamos cómo se vincula esta teoría con la función discriminante, y de qué manera se arriba a iguales resultados.

De acuerdo a la fórmula de Bayes, empleada en (9), una vez realizada la observación X, las probabilidades de pertenencia

de la misma a las poblaciones P_1 y P_2 serán respectivamente:

$$P(P_1/X) = \frac{P(P_1)f_1(x)}{P(P_1)f_1(x)+P(P_2)f_2(x)}; \quad P(P_2/X) = \frac{P(P_2)f_2(x)}{P(P_1)f_1(x)+P(P_2)f_2(x)}$$

obsérvese que $f_i(x)$ representa $P(X/P_i)$

La "evidencia" de X a favor de P_1, vendrá dada por el cociente:

$$\frac{P(P_1/X)}{P(P_2/X)} = \frac{P(P_1)f_1(x)}{P(P_2)f_2(x)}$$

Es decir:

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(P_1/X)}{P(P_2/X)} - \log \frac{P(P_1)}{P(P_2)}$$

Lo que implica: el logaritmo de la razón de verosimilitud, resulta igual a la diferencia entre la "evidencia" a favor de P_1 a posteriori, y la misma evidencia a priori. Esa diferencia, es precisamente la "información" proporcionada por la observación a favor de P_1. Si esa información es mayor que un número β (o mayor que 0, tomando $\beta=0$), la decisión será en favor de P_1.

En el caso de poblaciones normales multivariadas, esta información, como se observa, coincide con U(función discriminante). Si la información resulta positiva, ello implica que hay evidencia suficiente para adjudicar la observación a P_1.

Por otra parte, la "información media", es la esperanza de la información del conjunto de observaciones. Se la denomina información media de P_1 con respecto a P_2. Es claro que esa información media, coincidirá con la esperanza de U, o sea $\frac{D^2}{2}$.

La suma de la información media de P_1 con respecto a P_2 y la de P_2 con respecto a P_1 dará la "divergencia" entre ambas poblaciones, la cual resulta, para poblaciones normales, igual a D^2 (varianza de U).

Así como comprobamos la similaridad de resultados entre la función discriminante y los métodos de teoría de la información, Porebski (1966) hace una comparación bastante exhaustiva para el caso de dos poblaciones, de todos los métodos y estadísticos del análisis multivariado, llegando a la conclusión de que son semejantes y permiten inclusive los mismos test de significación.

Para más de dos poblaciones, los resultados difieren en cierta medida, ya que algunos métodos trabajan con toda la información original (todas las dimensiones o variables) y otros con un número menor, efectuando combinaciones lineales entre variables.

3.2. La función discriminante para más de dos poblaciones

3.2.1. Caso general

Consideramos ahora el caso de las poblaciones P_1, P_2, ... P_k, con sus correspondientes densidades $f_1(x)$, $f_2(x)$,..., $f_k(x)$ y probabilidades "a priori" q_1, q_2, ..., q_k. Deseamos dividir el espacio en k regiones mutuamente excluyentes R_1, R_2, ..., R_k, de modo que si una observación cae en R_i diremos que pertenece a P_i. Nuevamente suponemos de igual importancia (en cuanto a costo) cualquier error en la clasificación.

De esta manera, la probabilidad condicional a posteriori, será para la población P_i:

$$Pr(P_i/X) = \frac{q_i f_i(x)}{\sum_{j=1}^{k} q_j f_j(x)} \qquad (18)$$

Se asignará cada observación a aquella población en que resulta con mayor probabilidad de pertenecer (a posteriori).

Como los denominadores son todos iguales, se asignará a P_i si se cumple:

$$q_i f_i(x) > q_j f_j(x) \qquad \forall j \neq i$$

ó

$$\frac{f_i(x)}{f_j(x)} > \frac{q_j}{q_i} \qquad \forall j \neq i \qquad (19)$$

Igual que en el caso de dos poblaciones, si las q_i son desconocidas, se deberá determinar un número β y se adjudicará a P_i siempre que:

$$\frac{f_i(x)}{f_j(x)} > \beta \qquad \forall j \neq i$$

y si las q_i pueden suponerse todas iguales, será $\beta = 1$.

3.2.2. Poblaciones normales

Consideremos ahora el caso de k poblaciones normales, con iguales matrices de varianzas y covarianzas y distintos vectores medios. Es decir, la población P_i tiene densidad multivariada $N(\mu_i, \Sigma)$.

Dicha densidad es igual que (12). Los cocientes que permitirán asignar cada observación según (18) serán de la forma:

$$\frac{f_i(x)}{f_j(x)} = \exp. - \frac{1}{2}\left[(X-\mu_i)'\Sigma^{-1}(X-\mu_i)-(X-\mu_j)'\Sigma^{-1}(X-\mu_j)\right] \qquad (20)$$

los cuales, tomando logaritmo y reagrupando, se transforman en:

$$\log \frac{f_i(x)}{f_j(x)} = X'\Sigma^{-1}(\mu_1-\mu_j) - \frac{1}{2}(\mu_1+\mu_j)'\Sigma^{-1}(\mu_1-\mu_j) \qquad (21)$$

deberá entonces adjudicarse una observación a P_i si se cumple:

$$\log \frac{f_i(x)}{f_j(x)} > \log \frac{q_j}{q_i} \qquad \forall j \neq 1, \quad j = 1, 2, \ldots, k$$

Si $q_j = q_i$, entonces:

$$\log \frac{f_i(x)}{f_j(x)} > 0 \qquad \forall j \neq i \ , \ j = 1, 2, \ldots, k$$

A fin de llevar a la práctica esta verificación para asignar cada observación, resulta útil el siguiente camino:

Se construyen k funciones de la forma:

$$\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i' \Sigma^{-1}\mu_i \qquad i = 1, 2, \ldots, k$$

Se calcula para cada X:

$$U_i = X'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i \qquad (22)$$

y se adjudica a P_i si se cumple que

$$U_i > U_j \quad j \neq i, \ j = 1, 2, \ldots, k$$

Si se trata de muestras de las k poblaciones, se construirán las funciones:

$$S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_i' S^{-1}\bar{X}_i \qquad i = 1, 2, \ldots, k \qquad (23)$$

calculando para cada X:

$$U_i = X'S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_i'S^{-1}\bar{X}_i \qquad (24)$$

adjudicando igual que en el caso anterior.

3.3. Evaluación de las funciones discriminantes para más de dos poblaciones.

Igual que en el caso de dos poblaciones, se trata de decidir si las k poblaciones en consideración son realmente distintas, o si las diferencias existentes entre ellas no son significativas.

En 3.1.2.3. se planteó la D^2cde Mahalanobis y la T^2 de Hotelling como estadísticos adecuados para ese objetivo.

En este caso se presentan otros problemas: es necesario efectuar un test de diferencia de medias para las k poblaciones, pero además se necesita docimar la misma hipótesis para cada par de ellas. Es evidente que puede ocurrir un rechazo a la igualdad de medias de todos los grupos, y sin embargo algún par de ellos no presentan diferencias significativas. Por otra parte, discutiremos también aquí el papel de las variables incluídas en la observación, qué ocurre cuando se adicionan nuevas variables, cuál es la contribución de cada una de ellas a la discriminación.

Consideraremos también los tests adecuados para docimar la igualdad de matrices de varianzas y covarianzas de todos los grupos, supuesto que se ha hecho para obtener las funciones discriminantes, y que también es necesario para efectuar los tests de igualdad de medias.

3.3.1. D^2 Generalizada para varios grupos

Sean k poblaciones normales p-variadas y n_i el tamaño de la muestra correspondiente a la i-ésima población. Rao (1952)* presenta el siguiente estadístico para docimar la hipótesis de que las medias de las k poblaciones son iguales. Se trata de una generalización para muestras grandes, cumpliéndose el supuesto de igualdad de matrices de varianzas y covarianzas:

$$D^2_{pk} = \sum_{i,j=1}^{p} S^{-1}_{ij} \sum_{r=1}^{k} n_r (\bar{X}_{ir} - \bar{X}_i)(\bar{X}_{jr} - \bar{X}_j)$$

El cual puede expresarse como:

$$D^2_{pk} = tr(d' S^{-1} d) \qquad (25)$$

siendo d la matriz de orden pxk, cuyo elemento genérico es:

(*) Pág. 257.

$$d_{ij} = n_j (\bar{X}_{ij} - \bar{X}_i)$$

donde $\bar{X}_i$ es la media de la variable i para todos los grupos, y $\bar{X}_{ij}$ es la media de la variable i en la muestra correspondiente al grupo j.

Este estadístico, que tiene distribución aproximada ^{2} con p(k-1) grados de libertad, puede cumplir con otro objetivo además del ya señalado como dócima de diferencia de medias: analizar la contribución de la adición de nuevas variables al conjunto original. En este sentido, debe señalarse que es frecuente la tentación de añadir más variables para lograr mejores resultados, lográndose solo complicar el análisis sin aumentar el poder discriminatorio. Más adelante consideramos el tema de cuáles deben ser las variables a incluir, pero en este punto queremos ilustrar la aplicación de D^2 para ese objetivo: supóngase que se agregan q variables a las p originales, tendrá entonces:

$$D^2_{(p+q)k} = tr(d'S^{-1}d)$$

(siendo éstas las matrices del caso planteado arriba, sólo que cambia el orden de las mismas). Este estadístico tiene distribución ^{2} con (p+q)(k-1) grados de libertad. Los q(k-1) grados de libertad adicionales, contribuyen a la discriminación en

$$D^2_{(p+q)k} - D^2_{pk}$$

y su significación puede docimarse sabiendo que la diferencia se distribuye como ^{2} con q(k-1) grados de libertad.

Por otra parte, cualquiera sea el resultado obtenido por el test D^2, es conveniente efectuar el mismo para par de grupos, o aún para otros subconjuntos, a fin de determinar si son o no significativas las diferencias entre sus medias.

3.3.2. Matrices de dispersión entre grupos, intra grupos y total.

Definimos a continuación algunas matrices que serán de aplicación necesaria en éste y los próximos capítulos. Ellas son: W, T, A.

El elemento genérico de W es:

$$W_{ij} = \sum_{g=1}^{k} \sum_{n=1}^{n_g} (x_{ign} - \bar{x}_{ig})(x_{jgn} - \bar{x}_{jg}) \quad (*) \qquad (26)$$

O, para expresarlo en forma matricial, definimos previamente una matriz DIG_g formada por los vectores desvíos de las observaciones del grupo g-simo con respecto a las medias de cada variable en dicho grupo. Entonces, la matriz de suma de productos cruzados de desvíos intra-grupo (para el grupo g) será:

$$\underset{pxp}{W_g} = \underset{p\,x\,n_g}{DIG'_g} \quad \underset{n_g\,x\,p}{DIG_g}$$

La matriz W es la suma de las matrices W_g:

$$W = \sum_{g=1}^{k} W_g$$

Obsérvese que al ser p una matriz de rango completo, éste será igual a p. (Es una matriz formada por los numeradores de varianzas y covarianzas).

En cuanto a la matriz A, matriz de suma productos cruzados de desvíos entre los grupos, se obtiene restando W de la matriz de suma de productos cruzados de desvíos general, o total (T):

$$t_{ij} = \sum_{n=1}^{N} (x_{in} - \bar{x}_i)(x_{jn} - \bar{x}_j) \qquad (27)$$

(*) k es el número de grupos; n_g es el número de observaciones en el grupo g-simo; N es el total de observaciones; x_{ign} es el valor de la variable i-ésima. en la n-sima observación del grupo g-simo.

O, en forma matricial, siendo DT la matriz formada por los vectores desvíos de las observaciones con respecto a la media de cada variable para todos los grupos (media general):

$$T = DT'DT$$
$$p.p \quad p.NN.p$$

Su rango también es p. (rango completo)

Por último:

$$A = T-W \tag{29}$$

A es la matriz de suma de productos cruzados de desviaciones entre los grupos:

$$A = DEG'\ DEG$$
$$pxp \quad pxk \quad kxp$$

DEG es la matriz de desviaciones entre las medias de cada grupo y la media general de la variable correspondiente. Su elemento genérico es de la forma:

$$DEG_{ij} = (\bar{x}_{ij} - \bar{x}_j) \qquad i = 1,k;\ j = 1,p$$

Como de las k medias de los grupos sólo hay k-1 independientes, el rango de DEG es a lo sumo k-1 por lo que el rango de A será el menor entre k-1 y p.

La forma práctica de calcular A, resulta aplicando para a_{ij} la fórmula siguiente:

$$a_{ij} = \sum_{g=1}^{k} n_g(\bar{x}_{ij} - \bar{x}_i)(\bar{x}_{jg} - \bar{x}_j) \tag{29}$$

3.4. Otros tests para evaluar diferencia de medias

Existen en el análisis multivariado otros tests para decimar hipótesis acerca de diferencias de medias. Todos ellos se

basan en el supuesto de distribuciones normales, y de igualdad de matrices de varianzas y covarianzas. Los estimadores con que operan son, en general, los máximo verosímiles, y suelen ser preferidos a D^2 por cuanto su distribución no depende tanto de los grados de libertad. (D^2 tiene distribución asintótica).

Kendall (1967) trata extensivamente en el tomo III la obtención y propiedades de estos estadísticos. Aquí los presentamos brevemente.

En general, se trata de docimar las siguientes hipótesis:

$$H_1: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$$ (igualdad de matrices de varianzas y covarianzas).

$$H_2: \mu_1 = \mu_2 = \ldots = \mu_k$$ (igualdad de vectores medios)

Los estimadores a utilizar son: 1) matrices de varianzas y covarianzas para todos los grupos (S) y para cada grupo (S_g):

$$S = \frac{1}{N-1} W; \quad S_g = \frac{1}{N_g - 1} W_g \quad g = 1, 2, .., k$$

2) Vectores medios muestrales para cada grupo:

$$\bar{X}_g = (\bar{X}_{g1}, \bar{X}_{g2}, \ldots, \bar{X}_{gp}) \quad g = 1, 2, \ldots, k$$

siendo

$$\bar{X}_{gi} = \frac{1}{n_g} \sum_{n=1}^{n_g} X_{i_n} \quad i = 1, 2, \ldots, p$$

En muchas investigaciones se omite el test H_1, lo cual es incorrecto, a pesar de ser explicable su ausencia porque el test H_2 es relativamente insensible a la desigualdad de matrices de dispersión (varianzas y covarianzas). Sin embargo, es este un supuesto necesario también para la construcción de las funciones

discriminantes de Fisher, que, de no cumplirse, resultarían cuadráticas o aún más complicadas, en lugar de lineales. Es necesario efectuarlo para conocer las limitaciones en caso de no cumplirse.

El test H_2 es una generalización del análisis de la varianza en el caso univariado. El mismo requiere el cálculo de matrices de sumas de productos de desviaciones intra grupos, entre grupos y totales.

El criterio descripto por Rao (1952)(*) es el de Lamda de Wilks, cuya distribución aproximada para varias variables fué propuesta por Bartlett.

Se define:

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

con W y T matrices de sumas de productos cruzados de desviaciones entre grupos y total respectivamente definidas en 3.3.2 (26) y (27).

La distribución aproximada, proporcionada por Rao es la siguiente:

siendo

$$S = [p^2(k-1)^2-4]/[p^2+(k-1)^2-5]$$

$$m = (N-1)-(p+k)/2; \quad \lambda = -[p(k-1)-2]/4$$

$$r = p(k-1)/2 \qquad y = \Lambda^{1/S}$$

resulta

$$F^{2r}_{ms+2} = \left(\frac{1-y}{y}\right)\left(\frac{ms+2}{2r}\right)$$

Como este test supone el cumplimiento de H_1, digamos que este último fué planteado por Bartlett, una de cuyas formulaciones presenta Box (1949) de la siguiente forma:

(*) Cap. 7.

$$M = (N-1)\ \ln|S| = \sum_{g=1}^{k} (n_g-1)(\ln|S_g|)$$

se requieren los parámetros:

$$A_1 = \sum_{g=1}^{k} \frac{1}{n_g-1} - \frac{1}{N-1} \quad \frac{2p^2+3p-1}{6(k-1)(p+1)}$$

$$A_2 = \sum_{g=1}^{k} \frac{1}{(n_g-1)^2} - \frac{1}{(N-1)^2} \quad \frac{(p-1)(p+2)}{6(k-1)}$$

Si

$$A_2 - A_1^2 > 0 \ , \ \text{entonces:}$$

$$f_1 = \frac{1}{2}(k-1)p(p+1) \ ; \ f_2 = (f_1+2)(A_2-A_1^2)$$

$$b = f_1(1-A_1 - f_1/f_2)$$

Y

$$F_{f_2}^{f_1} = \frac{M}{b}$$

Si

$$A_2 - A_1^2 < 0$$

$$f_1 = 0{,}5(k-1)p(p+1) \ ; \ f_2 = (f_1+2)/(A_1^2-A_2)$$

$$b = f_2/(1-A_1+2/f_2)$$

Y

$$F_{f_2}^{f_1} = f_2 M \ / \ f_1(b-M)$$

Una forma alternativa para calcular Λ es planteada por Rao (1952)(*) a partir de los valores característicos de W^{-1} en lugar de hacerlo por un cociente de determinantes. Resulta por este camino:

$$\Lambda = \prod_{i=1}^{p} \frac{1}{(1+\lambda_i)} \qquad (30)$$

Williams (1952 y 1955) propone algunos tests con distribuciones exactas para evaluar la significación de las funciones discriminantes. Sin embargo, sólo son planteados para un número pequeño de variables y grupos. Los que presentamos en este capítulo resultan buenas aproximaciones para cualquier número de variables y grupos, aunque asintóticos con respecto al número de observaciones.

3.5. Funciones discriminantes canónicas (Método de Bartlett)

Obsérvese que, geométricamente, las k poblaciones están centradas en k puntos del espacio p-dimensional (esos puntos son sus vectores medios). Las observaciones se distribuyen alrededor de esos puntos y los ejes pueden o no ser ortogonales según las variables estén o no incorrelacionadas. Bajo el supuesto de normalidad multivariada, uniendo los puntos correspondientes a igual densidad alrededor de cada media, se obtendrán elipsoides con centro en la media y cuyo tamaño y orientación dependerán de las dispersiones de cada grupo (varianzas y covarianzas).

Si las medias poblacionales están sobre una recta, y las probabilidades "a priori" de pertenencia a cada grupo son iguales, todas las funciones resultarán proporcionales y con una de ellas se podrá discriminar, particionando el espacio en planos paralelos, es decir, midiendo distancias a lo largo de la línea recta que une las medias. Se tendrá entonces una sola función discriminante que será suficiente para realizar las asignaciones.

(*) Cap. 8.

Este hecho ocurre muy raras veces en forma exacta, pero siempre resultará posible ajustar por medio de una recta las medias, y discriminar en base a las proyecciones de las medias sobre esa línea de ajuste. Si no fuera bueno el ajuste por medio de una recta, podrá seleccionarse otra función, ortogonal a la primera, de modo que las proyecciones de las medias se realicen sobre un plano y discriminar midiendo distancias en ese plano, y así sucesivamente.

Este es el camino que lleva al método de Bartlett para encontrar las funciones discriminantes, y que desarrollamos a continuación.

3.5.1. Planteo del método

Siguiendo la línea de investigación señalada en 2.2.1. trabajaron Bartlett, Williams, Rao, Kullback y otros autores. Partían para su análisis del estudio de las raíces características o valores propios de matrices de varianzas y covarianzas.

Rao (1952), Kendall (1967), desarrollan el método que, como señaláramos más arriba, trata de obtener un número reducido de funciones discriminantes, o dicho de otro modo, de disminuir el número de dimensiones en que se opera con las variables originales, Otros autores: Cooley (1962), Hope (1968), plantean en forma muy completa la aplicación del método.

Para encontrar la ecuación de la recta que ajuste las medias de los k grupos, se busca una combinación lineal de las variables originales. A fin de que dicho ajuste cumpla con el objetivo de servir como función discriminante, la combinación lineal deberá maximizar la varianza entre grupos con respecto a la intra grupos. En otras palabras, la combinación lineal debe explicar la mayor parte de la variabilidad (diferencia) entre los grupos. Para ello, se determinará cuál es el vector V de la combinación lineal:

$$f = V'X$$

donde X es el vector correspondiente a una observación; f resulta el valor correspondiente a la observación en el espacio unidimensional (es un escalar), y resulta de proyectar la observación X sobre la línea de ajuste de las medias. La media de cada grupo se obtendrá:

$$\hat{f}_g = V'\hat{X}_g \quad \text{ó} \quad \bar{f}_g = V'\mu_g$$

(dependiendo de que el vector de medias sea estimado o poblacional).

Este vector V debe maximizar la razón de la varianza entre grupos con respecto a la intra grupos; o sea, maximizar:

$$\lambda = \frac{V'AV}{V'WV} \quad (*) \qquad (30)$$

Para maximizar esta expresión procedemos a maximizar λ suponiendo constante el denominador. Esto es, debe maximizarse (30) sujeto a $V'WV = c$ ó $c-V'WV = 0$.

Por el método de los multiplicadores de Lagrange:

$$Q = V'AV+\lambda(c-V'WV)$$

Derivando con respecto a V e igualando a cero:

$$\frac{\partial Q}{\partial V} = 2AV - 2\lambda WV$$

$$AV - \lambda WV = \emptyset$$

$$(A-\lambda W)V = \emptyset$$

Premultiplicando por W^{-1}

$$(W^{-1}A-\lambda I)V = \emptyset \qquad (31)$$

Los vectores V que satisfacen esta ecuación, son los vec

(*) A, W y T son las matrices definidas en 3.3.2.

tores característicos de $W^{-1}A$, asociados a sus valores característicos λ que satisfagan la ecuación característica:

$$|W^{-1}A - \lambda I| = 0$$

Como se desea maximizar la expresión (30), como ésta se satisface para todos los vectores característicos de $W^{-1}A$, y es igual a λ, el máximo se alcanzará tomando el V asociado al mayor λ.

Los componentes de V son los coeficientes de la combinación lineal que constituye la primera "variable canónica", o sea la recta que mejor ajusta las medias de todos los grupos, (en el sentido ya indicado, de maximizar la varianza entre los grupos), sobre la cual se proyectan las observaciones efectuando el producto interno del vector de coeficientes V por el vector X correspondiente a una observación, es como se obtiene la proyección de dicha observación sobre la recta de ajuste, cuyo valor resulta un valor particular de esa variable canónica, también llamado "score" o "puntaje" discriminante o canónico.

Si el ajuste no se considera lo suficientemente bueno (más adelante se plantean los criterios para decidir en eses aspecto), se toma una segunda "variable canónica": la combinación lineal que surge al utilizar como coeficientes los componentes del vector característico asociado al segundo valor en orden decreciente, y así sucesivamente.

Si bien los vectores característicos de $W^{-1}A$ no son en general ortogonales, las variables canónicas sí lo son; de modo que en conjunto, forman un sistema de ejes cartesianos ortogonales de dos, tres o más dimensiones(*)

(*) Siendo V_1 y V_2 dos vectores característicos de $W^{-1}A$, asociados a los valores $\lambda_1 \neq \lambda_2$ (no nulos), probamos que $V_1'AV_2=0$ y $V_2'W^{-1}V_1=0$:

3.5.2. Test acerca del número de variables canónicas

El conjunto de valores característicos de $W^{-1}A$ representa todo el poder discriminante de las funciones. Como el número máximo de valores no nulos, que depende del rango de dicha matriz, es igual al mínimo entre p(número de variables y rango de W) y

---

$$(W^{-1}A-\lambda_1 1)V_1 = \emptyset$$

$$(W^{-1}A-\lambda_2 1)V_2 = \emptyset$$

ecuaciones que pueden expresarse cómo:

$$AV_1-\lambda_1 W^{-1}V_1 = \emptyset \quad (1)$$

$$AV_2-\lambda_2 W^{-1}V_2 = \emptyset \quad (2)$$

Premultiplicando la primera ecuación por V_2' y la segunda por V_1' y restando:

$$V_2'AV_1 - \lambda_1 V_2'W^{-1}V_1 = 0$$
$$V_1'AV_2 - \lambda_2 V_1'W^{-1}V_2 = 0$$
$$(\lambda_1-\lambda_2)V_2'W^{-1}V_1 = 0$$

Luego $V_2'W^{-1}V_1 = 0$

De igual modo, multiplicando (1) y (2) previamente por $\frac{1}{\lambda_1}$ y $\frac{1}{\lambda_2}$:

$$\frac{1}{\lambda_1} V_2'AV_1 - V_2'W^{-1}V_1 = 0$$
$$\frac{1}{\lambda_2} V_1'AV_2 - V_1'W^{-1}V_2 = 0$$
$$(\frac{1}{\lambda_1} - \frac{1}{\lambda_2})V_1'AV_2 = 0$$

Luego $V_1'AV_2 = 0$.

Esto implica que en las variables canónicas, las observacioes están incorrelacionadas tanto entre grupos (las medias lo están) como intra grupos.
Como consecuencia de ello también están incorrelacionadas en general (independientemente de la formación de los grupos). Es decir: $V_1'TV_2=0$ (siendo T la matriz de suma de productos cruzados de desviaciones total).

k-1 (número de grupos menos uno y, si es menor que p, rango de A), ése es el mayor número de variables canónicas que pueden obtenerse.(*)

La suma de todos los valores característicos, al ser igual a la traza de la matriz que relaciona la variabilidad entre grupos con respecto a la intra grupos, condensa todo el poder discriminante: maximiza en todas las dimensiones (p ó k-1) las diferencias entre las medias de los grupos, con respecto a las observaciones dentro de cada grupo.

La relación:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \; 100 \qquad (32)$$

indica el porcentaje de la traza representado por cada valor característico, y da una idea acerca de cuál es el número de variables canónicas que resultan suficientes para explicar la mayor parte de la variabilidad. Generalmente dos o tres valores absorben casi el 100% de la traza.

Bartlett, citado en Rao(1952)(**) plantea la distribución aproximada de las raíces características, de la cual resulta una alternativa para docimar la diferencia de medias, así como para decidir con cuántas variables canónicas se debe trabajar en el espacio reducido.

El estadístico es:

$$\ln \prod_{i=1}^{p} (1+\lambda_i)^{(N-1)-\frac{p+k}{2}}$$

(*) En adelante nos referimos a p como el número máximo de valores no nulos, debiendo entenderse que si k-1 es menor que p, será k-1 ese número máximo.

(**) pág. 373.

Su distribución es χ^2 con p(k-1) grados de libertad.

Estadístico que puede expresarse como:

$$\left[N-1-\frac{p+k}{2}\right]\sum_{i=1}^{p}\ln(1+\lambda_i)\sim\chi^2_{p(k-1)} \qquad (33)$$

Un valor significativo indica que las medias están separadas. Por el carácter aditivo de la χ^2, puede efectuarse un test para cada una de las raíces características, lo que implica la dócima de diferencia de medias en cada una de las variables canónicas. Si sólo la primera resulta significativa, una recta proporciona un buen ajuste entre las medias, y pueden medirse las distancias entre las observaciones de cada grupo y sus medias. No obstante, para alguna de las variables, o para algún par de grupos, la diferencia puede ser significativa, o el tamaño de las muestras puede no ser suficiente para mostrar la real separación entre los grupos; de allí que en muchos casos es interesante efectuar el análisis en alguna de las dimensiones que no aparecen como significativas.

Con respecto al test para dimensión, digamos que el estadístico a utilizar para cada λ_i es:

$$\left(N-1-\frac{p+k}{2}\right)\ln(1+\lambda_i) \qquad (34)$$

Los p(k-1) grados de libertad se distribuyen entre los valores característicos como sigue:

$$p(k-1) = (p+k-2)+(p+k-4)+\ldots\ldots$$

es decir, la distribución correspondiente al primer valor tiene (p+k-2) grados de libertad; les restantes [p(k-1)-(p+k-2)]; la segunda (p+k-4) y las restantes (luego de eliminar las dos primeras: [p(k-1)-(p+k-2)-(p+k-4)]. Así sucesivamente, perdiendo dos grados de libertad por cada valor.

La representación gráfica de las medias en el espacio reducido y en cada una de las dimensiones significativas, (y aún en algunas no significativas) aporta interesantes conclusiones y ayuda al análisis del agrupamiento realizado.

Agreguemos que, el valor de D^2 obtenido al efectuar una reducción en el número de dimensiones, resulta siempre inferior al que se obtendría de trabajar con las funciones discriminantes de Fisher en el espacio p-dimensional. Esto ocurre también para los valores de D^2 entre cada partde grupos.

Teniéndo en cuenta las distribuciones de estos estadísticos, es posible docimar la significación de la reducción en el valor de D^2 operada al pasar de un método al otro. Se tratará de un cociente de D^2 cuya distribución aproximada corresponde a una F. El test resulta equivalente al descripto más arriba.

3.6. Utilización de las funciones canónicas para clasificar y reclasificar observaciones.

3.6.1. Obtención de medias y dispersiones en el espacio reducido.

Ya se ha mencionado que si efectuado el test correspondiente, resulta que una sola función discriminante es suficiente, ocurrirá que las medias de los grupos están situadas aproximadamente sobre una recta; si se trata de dos funciones discriminantes, un plano resultará el marco adecuado para representar las medias de los grupos. Alrededor de dichas medias, se encuentran dispersas las observaciones.

En 3.4. se plantea la manera de encontrar el valor de cada observación, así como de las medias de cada grupo en dicho espacio reducido:

$$f = V'X \qquad \text{para cada observación } X$$

$$\bar{f}_g = V'\bar{X}_g \quad \text{ó} \quad \bar{f}_g = V'\mu_g \qquad \text{para la media del grupo } g.$$

$$g = 1, 2, \ldots, k$$

Si se han utilizado n funciones discriminantes (variables canónicas), f es un vector de n componentes ya que V' resultará de orden nxp y X es de orden px1.

Con los k vectores de medias se forma la matriz de medias en el espacio reducido.

De igual modo, se obtiene la matriz de dispersión para cada grupo en el espacio reducido:

$$D_g = V'S_gV \qquad g = 1, 2, \ldots, k \qquad (35)$$

La distribución de las variables canónicas (en el espacio reducido) es también normal (de cumplirse la hipótesis para los valores originales) por ser combinaciones lineales de variables normales. De allí que, conociendo las medias y dispersiones sea posible construir las elipses o elipsoides que corresponden a la forma cuadrática de la distribución normal bivariada o multivariada según el caso.

En esas elipses o elipsoides es posible ubicar también cada observación, prestando el análisis gráfico una importante ayuda para la clasificación de las observaciones.

En el Apéndice II se incluye un método práctico para graficar las elipses en el plano.

Por otra parte, y en este punto hay cierta semejanza con la interpretación de las componentes principales, las componentes de cada variable canónica (vector característico) permiten analizar la contribución de cada variable a la discriminación. Esas componentes son las que proporcionan las ponderaciones para cada variable, a fin de obtener el valor canónico de cada observación (valor de la observación en el espacio reducido). Sin embargo, no es necesario caer en interpretaciones subjetivas acerca del "significado" de cada función discriminante; es suficiente un estudio detenido (auxiliado por la interpretación gráfica) de las variables que más contribuyen a separar los grupos. Puede ocurrir que alguna variable de escasa importancia económica

(o en general, de escasa importancia en relación al problema que se investiga) resulta con gran poder discriminante. Es necesario preguntarse si tendrá sentido la separación de grupos así obtenida, o si debe eliminarse esa variable del conjunto original.

Como los coeficientes de las funciones discriminantes están afectados por la varianza de cada variable, y éstas en general son distintas, es conveniente homogeneizarlas multiplicando cada elemento de la función discriminante por la desviación estandar correspondiente. Esto ocurre porque la desviación estandar es una medida de dispersión de cada variable. Mediante esta multiplicación vuelven a tratarse todas las variables con el mismo peso, porque al usar como coeficientes de las variables canónicas los vectores característicos de $W^{-1}A$ se ha normalizado la variabilidad intra-grupo (se ha hecho igual a 1 al multiplicar por W^{-1}), resultando las variables con una ponderación inversa a su dispersión. En la práctica, se puede multiplicar cada componente del vector característico por el numerador de la desviación estandar (raíces cuadradas de elementos diagonales de W) dado que los denominadores (grados de libertad) son todos iguales.

3.6.2. Clasificación de las observaciones utilizando las variables canónicas, o las funciones discriminantes de Fisher.

Según la definición estricta de discriminación, el objetivo de este análisis termina cuando se obtiene el conjunto de funciones que permitirán asignar nuevas observaciones al grupo al cual tiene mayor probabilidad de pertenecer. Por el criterio de Bayes presentaremos la fórmula para calcular las probabilidades de cada grupo, que permitan asignar las observaciones.

De acuerdo a (17) en el punto 3.2.1., la probabilidad condicional a posteriori para cada grupo es:

$$Pr(\pi_i/X) = \frac{q_i f_i(x)}{\sum_{g=1}^{k} q_g f_g(x)}$$

En el caso de poblaciones normales multivariadas se tendrá:

$$Pr(P_i/X) = \frac{\frac{q_i}{|S_i|^{1/2}} \exp - \frac{1}{2}\left[(X-\bar{X}_i)' S_i^{-1} (X-\bar{X}_i)\right]}{\sum_{g=1}^{k} \frac{q_g}{|S_g|^{1/2}} \exp - \frac{1}{2}\left[(X-\bar{X}_g)' S_g^{-1} (X-\bar{X}_g)\right]} \qquad (36)$$

Y trabajando en el espacio reducido; con los valores canónicos de las variables:

$$P(P_g/f) = \frac{\frac{q_g}{|D_g|^{1/2}} \exp - \frac{1}{2}\left[(f-\bar{f}_g)' D^{-1} (f-\bar{f}_g)\right]}{\sum_{i=1}^{k} \frac{q_i}{|D_i|^{1/2}} \exp - \frac{1}{2}\left[(f-\bar{f}_i)' D_i^{-1} (f-\bar{f}_i)\right]} \qquad (37)$$

Las q_i, que como se recordará son las probabilidades "a priori", pueden ser reemplazadas por los tamaños de las muestra de cada grupo.

Llamando χ^2 a la forma cuadrática del exponente, ya que precisamente ésa es su distribución, la fórmula puede expresarse:

$$Pr(P_i/X) = \frac{\frac{q_i}{|D_i|^{1/2}} \exp - \frac{\chi_i^2}{2}}{\sum_{g=1}^{k} \frac{q_g}{|D_g|^{1/2}} \exp - \frac{\chi_i^2}{2}} \qquad (38)$$

El valor de la probabilidad indicará en qué medida una observación X se aleja del centro del elipsoide correspondiente al grupo i-ésimo; de manera que cada observación presentará para algún grupo, mayor probabilidad que para los demás, (salvo casos de igualdad).

Si las probabilidades (tamaños de los grupos) son igua-

les, así como las matrices de dispersión, entonces la determinación del grupo con mayor probabilidad puede hacerse comparando las χ^2_g: el menor valor de χ^2_g indicará el grupo en el cual la observación se encuentra más cercana a la media.

Si se ha trabajado con las funciones discriminantes de Fisher, entonces (ver 3.2.2.) al adjudicar la observación al grupo para el que se obtiene mayor valor discriminante, ya se sabe que a él corresponde la más alta probabilidad de pertenencia de esa observación. La fórmula (36) puede ser aplicada sólo a ese grupo a fin de calcular cómo es de elevada dicha probabilidad.

Una simplificación se obtiene al suponer iguales las matrices de dispersión y los tamaños muestrales en este caso. Si llamamos U_M el mayor valor discriminante (correspondiente al grupo M) de una observación, la fórmula (36) calculada para el grupo M se reduce a:

$$P(P_M/X) = \frac{1}{\sum_{g=1}^{k} e^{(U_g - U_M)}} \qquad (*)$$

(*) Porque según (24) $U_i = X'S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_i'S^{-1}\bar{X}_i$ y en (36) si $q_i = q_j\ (\forall i,j)$ y $S_i = S\ (\forall_i)$, se simplifican los cocientes $\frac{q_i}{|S_i|^{1/2}}$; la expresión:

$$\exp -\frac{1}{2}\left[(X-\bar{X}_i)'S_i^{-1}(X-\bar{X}_i)\right] = \exp(X'S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_i'S^{-1}\bar{X}_i - \frac{1}{2}X'S^{-1}X) =$$

$$= \exp(X'S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_iS^{-1}\bar{X}_i)\exp(\frac{1}{2}X'S^{-1}X)$$

Luego (36) queda: $P(P_i/X) = \dfrac{\exp(X'S^{-1}\bar{X}_i - \frac{1}{2}\bar{X}_i'S^{-1}\bar{X}_i)\exp -\frac{1}{2}(X'S^{-1}X)}{\sum_{g=1}^{k}\exp(X'S^{-1}\bar{X}_g - \frac{1}{2}\bar{X}_g'S^{-1}\bar{X}_g)\exp -\frac{1}{2}(X'S^{-1}X)}$

Simplificando $\exp(X'S^{-1}X)$: $P(P_i/X) = \dfrac{\exp U_i}{\sum_{g=1}^{k}\exp U_g} = \dfrac{1}{\sum_{g=1}^{k}\exp(U_g - U_i)}$

3.6.3. Reclasificación de observaciones

Si el problema que se encara es el de encontrar la o las funciones que mejor discriminen entre varios grupos a partir de una muestra, a fin de asignar nuevos elementos, es decir el clásico planteamiento del análisis discriminante, el proceso estaría concluído en el punto anterior.

Sin embargo, nos proponemos un objetivo distinto: no se trata de asignar "nuevas observaciones" a los grupos preestablecidos, sino de clasificar de una manera aceptable el conjunto original de observaciones que en numerosos casos resulta toda la población y no una muestra (problema que consideraremos en especial más adelante).

Cuál es en ese caso la utilidad de la función o funciones discriminantes? Se parte de una clasificación previa, realizada por algún criterio razonable del investigador; esa clasificación es, por supuesto, susceptible de errores tanto relativos al número de grupos que se proponen, como a la ubicación de las observaciones en dichos grupos, como a la cantidad y calidad de las variables que se han usado para caracterizar cada observación.

Para cada uno de estos casos, contemplaremos las soluciones posibles.

3.6.3.1. Existencia de grupos distintos de los planteados originalmente.

Puede ocurrir que en la primera clasificación, existan varias observaciones con baja probabilidad de pertenencia a todos los grupos. Esta circunstancia suele deberse a la existencia en más grupos que los propuestos, y en ese caso se estudiará la posible conformación de los mismos de acuerdo al criterio del investigador (las observaciones con baja probabilidad, en ese caso, se supone que tendrán algunas características comunes).

Las bajas probabilidades pueden originarse también en la existencia de menos grupos, hecho éste más fácilmente verifica

ble a través de las D^2 entre grupos (ver 3.3.1.): cuando dos grupos presentan un valor de D^2 no significativo, se los unifica.

Si persistieran las bajas probabilidades luego de estos pasos, puede recurrirse a la separación de esas observaciones para obtener las funciones discriminantes a fin de que no influyan sobre su construcción, adjudicándolas a posteriori.

3.6.3.2. Pertenencia de las observaciones a grupos distintos del original

En este caso, un procedimiento iterativo permite mejorar paso a paso las clasificaciones, aumentando el valor del estadístico D^2.

Se procede a clasificar con el primer conjunto de funciones discriminantes obtenidas; aquellas observaciones que por su probabilidad, son asignadas a un grupo diferente del original, se cambian de grupo, volviendo a obtener funciones discriminantes y repitiendo el procedimiento.

Se comprueba que en cada paso, el valor de D^2 aumenta, mostrando una mayor diferenciación entre grupos lo cual siempre resulta beneficioso para la clasificación.

El valor de D^2, según surge de su definición [(25) en 3.3.1. y (16) en 3.1.2.3] depende de las diferencias entre las medias de los grupos y la media general para cada variable $[(\bar{x}_{ig}-\bar{x}_g),\ i = 1,2, \ldots, k]$ o sea de la magnitud de las varianzas covarianzas entre grupos (cuyos numeradores forman la matriz A). Recordemos que la variabilidad total (T) puede descomponerse en

$$T = W+A \qquad \text{(ver 3.3.2)}$$

y su elemento genérico:

$$t_{ij} = w_{ij}+a_{ij}$$

$$\sum_{n=1}^{N}(x_{in}-\bar{x}_i)(x_{jn}-\bar{x}_j)=\sum_{g=1}^{k}\sum_{n=1}^{n_g}(x_{ign}-\bar{x}_{ig})(x_{jgn}-\bar{x}_{jg})+$$

$$+\sum_{g=1}^{k}n_g(\bar{x}_{ig}-\bar{x}_i)(\bar{x}_{jg}-\bar{x}_j) \qquad (38)$$

Toda vez que se reclasifica una observación en un grupo en el que resulta más cerca de la media para el conjunto de las variables, se logra una disminución en el primer término del segundo miembro de (38), con lo que aumenta el segundo término (dado que t_{ij} no altera). De esta manera, al trasladar una observación a otro grupo que el original, siempre que se encuentre más cerca de su media, aumentará el valor de D^2(la variabilidad entre grupos) y su significación. Este procedimiento necesariamente tiene un límite ya que por ser finito el número de observaciones y determinado el número de grupos, siempre llegará un momento en que todas las observaciones estarán en aquél grupo en que la media resulta más cercana y aquí la aplicación de las funciones discriminantes no llevará a ninguna reclasificación.

Debemos hacer una distinción entre lo que resulta si se utilizan las funciones discriminantes de Fisher o un número reducido de funciones canónicas.

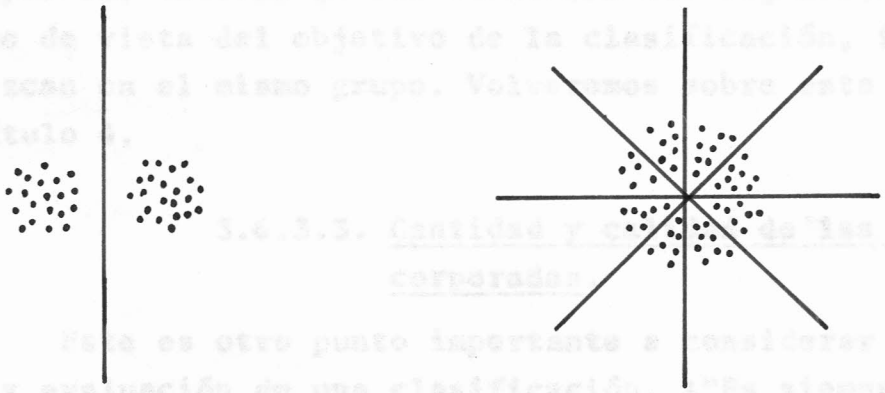
Con el método de Fisher, se trabaja en el número máximo de dimensiones (que será el menor entre la k-1 y p)(*). Si agregamos el hecho de no considerar los tamños de grupos como ponderaciones (es decir, igual probabilidad de pertenencia de cada observación a todos los grupos) tal como lo hemos hecho en el presente trabajo, entonces lo afirmado más arriba se cumple estrictamente.

Si en cambio, se ha utilizado un número reducido de va

(*) Si bien el programa que utilizamos en el trabajo siempre proporciona k funciones, solamente k-1 (o p si es menor) es el número de ellas que son independientes.

riables canónicas, (se trata de ajustes efectuados entre las medias) pueden resultar clasificaciones ligeramente diferentes a las del caso anterior, dado que la cercanía a las medias de los grupos se mide por combinaciones lineales de las variables. Otras diferencias se obtienen cuando para asignar las observaciones a un grupo, se pondera con la probabilidad de pertenencia al mismo (puede ser utilizando el tamaño del grupo como factor de ponderación).

Podemos preguntarnos si se llega a un mismo agrupamiento comenzando por diferentes puntos de partida (se entiende que con un número predeterminado de grupos). Un sencillo ejemplo gráfico nos muestra que no es así:



(a) (b)

En el caso (a), hay una sola posibilidad de clasificación final, aunque se haya comenzado con diferentes particiones. En el caso (b), el resultado final dependerá del punto de partida (hay infinitas posibilidades). Entre estos dos casos extremos, pueden presentarse muchos otros y es claro que, mientras menor sea el número posible de clasificaciones iniciales, más determina

da quedará la final.(*)

Sin embargo, dada una clasificación inicial, entonces podemos afirmar que el agrupamiento final será unívoco siguiendo el proceso señalado de reclasificación.

A pesar de estas propiedades óptimas del proceso mecánico de reclasificación, entendemos que no es conveniente aplicarlo de esta manera. Antes de cambiar de grupo una observación, es necesario analizar cuidadosamente sus características: los valores de las variables económicamente más importantes, la cercanía geográfica con el grupo al que se asigna, etc. Pueden existir dentro de un grupo elementos con baja probabilidad, que están más o menos alejados de la media, pero que por determinadas características, o por los valores de sus variables más significativas desde el punto de vista del objetivo de la clasificación, interesa que pertenezcan en el mismo grupo. Volveremos sobre esta cuestión en el Capítulo 4.

3.6.3.3. Cantidad y calidad de las variables incorporadas.

Este es otro punto importante a considerar en el mejoramiento y evaluación de una clasificación. ¿"Es siempre necesario o beneficioso incluir nuevas variables? ¿Son relevantes todas las que se consideraron en primera instancia? Sokal (1963) hace consideraciones acerca de las variables que no deben ser incluidas en cualquier proceso de clasificación multivariada. Por ejemplo, no deben incluirse variables "lógicamente correlacionadas", esto es, cuando una de ellas es consecuencia lógica de la otra (ej.: ingre

---

(*) Este ejemplo es de Casetti citado por King (1969) pág.214.En el mismo libro (pág.210)King reproduce conclusiones de otro trabajo de Casetti(1964) acerca de los vectores característicos de $W^{-1}A$; afirma que son ortogonales. La demostración efectuada por Casetti en ese trabajo adolece de un error en los primeros pasos. Dicha propiedad es inexistente para este tipo de matrices no simétricas. Numerosos ejemplos así lo prueban. Lamentablemente no hemos podido ubicar el segundo trabajo de Casetti en que trata acerca del proceso iterativo de clasificación. Nos remitimos a las partes citadas por King.

so, número de habitantes e ingreso per cápita). Cuando esa dependencia es parcial, es decir, dadas dos variables A y B la presencia de B depende parcialmente de A, pero también de otros factores, la decisión acerca de incluir o no B, dependerá de la importancia de esos otros factores.

Demás está insistir en la no inclusión de variables irrelevantes desde el punto de vista del objetivo central del trabajo.

En cuanto a la cantidad de variables, ya mencionamos que la posibilidad de computadoras potentes es una fuerte tentación para aumentar el número de las mismas sin mayor análisis previo.

Debe hacerse notar que en el cálculo de D^2 influyen las correlaciones, por lo tanto si se incluyen nuevas variables muy correlacionadas, éstas no producen aumentos significativos en el valor de dicho estadístico.

En 3.3.1. presentamos los test propuestos por Rao (1952) para decidir acerca de la significación de nuevas variables en el proceso de discriminación. Por otra parte, el mismo autor(*) presenta un interesante ejemplo en que D^2 deja de ser significativo al aumentar el número de variables.

Los trabajos citados de Adelman y Morris (1968, 1970) utilizan el criterio de Rao para seleccionar variables, lo que originó críticas que analizaremos en el capítulo 5.

3.7. El cumplimiento de las hipótesis de normalidad e igualdad de matrices de varianzas-covarianzas.

Como es sabido, al trabajar con una distribución multivariada, el cumplimiento de la hipótesis de normalidad para las distribuciones marginales no asegura el mismo para la conjunta; excepto, por supuesto, el caso de independencia. De allí, que, an

(*) pág. 252.

te lo complicado de posibles tests para pruebas de normalidad conjunta, y su inexistencia hasta el momento cuando las variables son numerosas, proponemos la transformación ortogonal de variables, mediante los vectores característicos de la matriz de varianzas-covarianzas original. Sin embargo, de acuerdo con Lockhart (1967), pensamos que esta transformación solo se hará una vez verificada la normalidad para las distribuciones marginales y en caso de su cumplimiento, ya que el incumplimiento de las mismas imposibilita una distribución conjunta normal.

La transformación propuesta es la siguiente:

Sea $X = (x_1, x_2, \ldots, x_p)$ el vector de variables aleatorias con distribución p-variada, y Σ es matriz de varianzas y covarianzas (supuesta no diagonal). Por ser esta última una matriz simétrica, existe la matriz ortogonal V formada por los vectores característicos normalizados de Σ tal que:

$D = V'\Sigma V$ resulta diagonal y sus elementos diagonales son los valores característicos de Σ.

Luego, si X tenía distribución normal p-variada, con sus p variables dependientes, el vector aleatorio:

$Y = V'X$ tendrá también distribución normal p-variada y sus variables serán independientes.

Por la incidencia de cada variable en la variación total, es necesario señalar que serán más importantes las consecuencias de un alejamiento de la normalidad de las variables correspondientes a los mayores valores característicos.

Los trabajos de Shapiro y Wilk (1965), Box (1953) y Pearson (1964) proporcionan métodos adecuados para los tests de normalidad, al mismo tiempo que permiten, sobre todo el último de los mencionados, evaluar las consecuencias de la falta de normalidad. Como es sabido, ésta no afecta demasiado las pruebas de igualdad de medias, aún en el caso de varianzas desiguales, cuando los tamaños muestrales son parejos. Sin embargo, sí afecta los tests de

igualdad de matrices de varianzas-covarianzas. En determinados casos, el rechazo de esta hipótesis puede ser consecuencia de la no normalidad antes que de la diferencia de matrices de dispersión. (ver Ito (1969)).

Por otra parte, la cantidad de variables influyen notablemente sobre la sensibilidad de estos tests.

Holloway y Dunn (1967) muestran que, en caso de poblaciones normales, trabajando con más de dos variables, el incumplimiento de la igualdad de matrices de dispersión produce un crecimiento notable en los niveles de significación, que se agrava en cuanto mayor es la diferencia de dichas matrices, el número de variables y menor el tamaño de las muestras.

Zhezhel (1968) analiza la eficiencia de la función discriminante para dos grupos en caso de distribuciones arbitrarias, llegando a conclusiones semejantes: la probabilidad de clasificación incorrecta aumenta cuando la población no es normal, pero la magnitud del error depende del grado de separación entre las poblaciones, haciéndose insignificativo cuando D^2 es relativamente grande. Si $D^2 \geq 4$ ocurre siempre que la probabilidad de clasificación errónea es menor que 0.5, lo que hace preferible la utilización de la función a una clasificación al azar.

Gilbert (1969) por su parte, analiza el efecto de la desigualdad de matrices de varianzas-covarianzas sobre la función discriminante llegando a la misma conclusión a la que agrega la verificación de que la función discriminante, de no cumplirse dicha hipótesis, resulta igualmente útil para clasificar, no así para evaluar riesgos.

4. Un caso de aplicación: regionalización económica de la Provincia de Córdoba.

4.1. Unidades de observación

El interés del presente trabajo fué inicialmente efectuar una regionalización socioeconómica de la Provincia, aplicando los métodos aquí expuestos. El primer problema a resolver fué el de la unidad de observación: ¿debía ser del Departamento u otra unidad más reducida? Con solo hechar un vistazo al mapa de la Provincia de Córdoba se aprecia la heterogeneidad económica que puede existir en cada Departamento. Estos son muy extensos y abarcan zonas de características bastante diferentes.

La única alternativa era la Pedanía, unidad política menor que el Departamento; surge entonces la dificultad con los datos: hay muy pocos datos de tipo socio-económico a nivel de Pedanía.

Se prefirió en última instancia sacrificar algunas variables importantes, inexistentes a nivel de Pedanías, antes que una regionalización por departamentos, que hubiera resultado ineficaz para hacer una aplicación completa de la función discriminante.

4.2. Variables consideradas

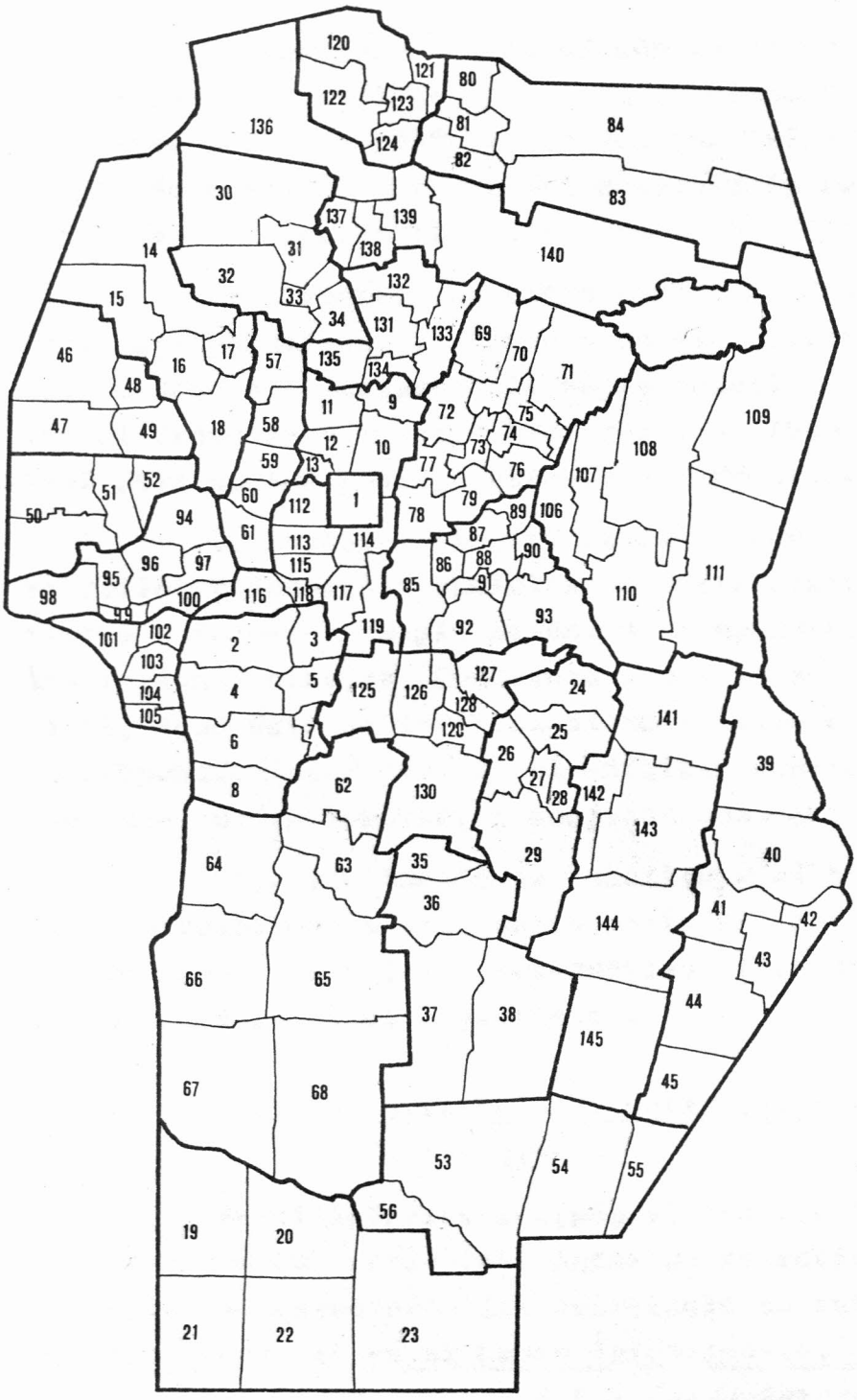
Son las siguientes:

1. Ganado bovino (excluído tambo)
2. Maní
3. Sorgo
4. Fruticultura
5. Trigo
6. Patentamiento de vehículos
7. Hotelería
8. Crecimiento de la población.

Las variables 1, 2, 3, 4, y 5 se miden en cabezas de ganado o hectáreas sembradas por cada 100 hectáreas bajo explota-

MAPA I

DIVISION POLITICA DE LA PROVINCIA DE CORDOBA



ción agropecuaria en la Pedanía. (la variable 4 incluye plantes nuevas y en producción de ciruelos, damascos, duraznos, manzanos y perales)(corresponden al año 1965).

La variable 6 mide el número de vehículos patentados por cada 1000 habitantes (en 1972); la 7, el número de establecimientos hoteleros por cada 1000 habitantes (en 1972) y la 8 mide el crecimiento demográfico intercenssl mediante la relación: Población según censo 1970/Población según censo 1960.(*)

La presencia de cinco variables destinadas a medir fenómenos agropecuarios, y sólo dos de tipo económico y uno demográfico, se debe a la disponibilidad de aquellos datos proporcionado por el Censo Agropecuario Provincial de 1965 (se realizó otro en 1969 pero no fué posible obtener los resultados).

Lamentablemente, no se contó con más variables aptas para reflejar fenómenos sociales, ni estructura de propiedad de la tierra o formas de explotación, ni tampoco con una que detectara las regiones mineras. En cuanto a las zonas con desarrollo industrial, como éste se verifica esencialmente en regiones urbanas no correspondía incorporarlas al análisis general por pedanías; simplemente pueden tomarse en conjunto como una zona más.

Otra limitación la constituye el hecho de que las variables corresponden a diferentes períodos. Nos vemos en la necesidad de suponer que las agropecuarias (las más antiguas) no han sufrido modificaciones importantes.

4.2.1. Correlaciones entre las variables, e importancia de las mismas

En el Apéndice I-C(pág.v) aparece la matriz de correlaciones entre las variables. Antes de seleccionar las de tipo agropecuario, se estudiaron las correlaciones entre casi todas las variables presentes en el Censo Agropecuario.

(*)Fuentes:Variables 1,2,3,4,5,: Censo Agropecuario provincial(1965) Variable 6:Boletín Estadístico-Area Estadística de la Provincia de Córdoba,Enero/Abril 1973. Variable 7:Idem Setiembre/Diciembre 1972. Variable 8: Censo de 1970(Area Estadística de la Provincia-Cifras provisorias).

Dada la capacidad limitada de la computadora, y la necesidad de incluir otras variables no agropecuarias, pretendíamos que sólo quedaran cinco de ellas. Se eliminaron las que presentaban altas correlaciones y las de menor importancia económica para la Provincia. Hubo una primera etapa, en la que se aplicaron las funciones discriminantes para una zonificación exclusivamente agropecuaria(*). En ella se usaron nueve variables: bovino, caprino, maiz, trigo, maní, lino, ovino, sorgo, avicultura.

En esa etapa ya fueron eliminadas: porcino, por estar altamente correlacionada con maiz; fruticultura, por su alta correlación con avicultura, alfalfa y centeno por la correlación con bovino; avena, mijo, poroto, cebada por no tener gran significación económica; girasol, por ser complementaria con otros cultivos (maiz, por ejemplo).

Para la segunda etapa, ante la necesidad de seleccionar cinco, se analizó la contribución de cada una de las nueve anteriores a la discriminación (según los vectores característicos estandarizados); eliminando caprino, maiz, lino, ovino, y reemplazando avicultura por fruticultura (más apta que la anterior para caracterizar la zona de riego permanente). Las correlaciones más elevadas entre las variables que quedaron resultan:

$$r_{15} = 0,46$$
$$r_{16} = 0,44$$
$$r_{13} = 0,42$$

Las restantes son a lo sumo igual a 0,20 (casi todas mucho menores).

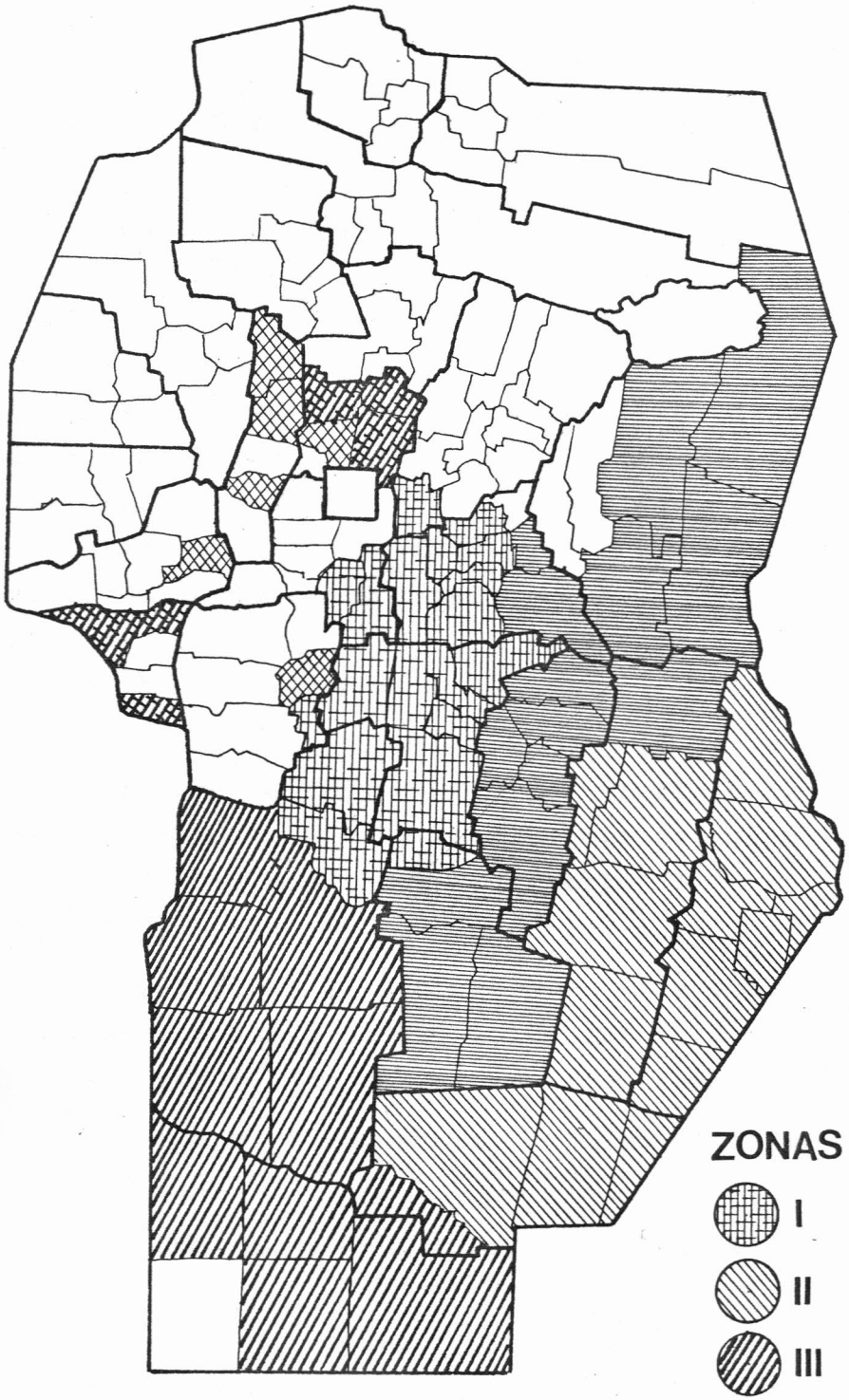
La limitación mencionada en capacidad de computadora, así como la carencia de datos a nivel de Pedanía, redujo el número de variables y los posibles alcances del trabajo. Este mantiene su in-

---

(*) Goldenhersch, Hebe (1973).

MAPA II

CLASIFICACION INICIAL



terés en cuanto a la coherencia de los resultados alcanzados con otros trabajos empíricos tendientes a zonificar la provincia(*) y a la aplicación de los métodos expuestos en capítulos anteriores. En el capítulo 5 se discute más ampliamente este aspecto de las limitaciones. A fin de superarlas, resultará interesante reiterar el mecanismo con los datos proporcionados por el Censo Económico de 1974.

Debe insistirse en la necesidad de eliminar variables que aunque pueden discriminar bien, carezcan de importancia económica. Ejemplifiquemos: en una región se cosecha gran cantidad de "piquillín", fruto que está ausente en el resto de las observaciones. Si se discrimina utilizando esta variable, aparecerá seguramente una zona caracterizada por las toneladas de piquillín que se cosechan, y similar en otros aspectos a otras observaciones. ¿Tendría algún sentido esa zona separada del resto, aunque los test revelen diferencias significativas?.

4.3. La clasificación inicial

En la etapa previa (**), en la que como mencionamos, se aplicó la función discriminante únicamente para variables agropecuarias, se tomó como base un conjunto de mapas "sombreados" elaborados en la Secretaría de Desarrollo de la Provincia de Córdoba en 1970 (***).

Se dividió la provincia en seis zonas considerando nueve variables; resultó no significativa la diferencia entre dos de esas zonas, por lo que se unieron resultando finalmente cinco.

Con esas cinco zonas, se hizo el estudio de variables expuesto en 4.2.1.

Quedaron de esa manera, para comenzar la presente zoni-

(*) Ríos, Raúl A. (1968)
(**) Goldenhersch Hebe (1973).
(***) Secret. Ministerio de Desarrollo (1970).

ficación, cinco variables de tipo agropecuario, a las que se agregaron las otras tres ya mencionadas. Al incorporar fruticultura y hotelería, resultó posible separar otras dos zonas caracterizadas por cultivos con riego permanente y por el turismo respectivamente.

Las siete zonas de partida son, por lo tanto, las que se observan en el Mapa 2; el listado de las Pedanías que las integran, puede consultarse en el Apéndice III9B.

Debe notarse que fué excluido el departamento Capital por resultar muy atípico: poca superficie bajo explotación agropecuaria, muy elevados índices de fruticultura, que hacía aumentar exageradamente el promedio general de esa variable deformando la zona con riego. Al margen del análisis aquí efectuado, debe incluirse en el grupo formado por los centros urbanos industrializados, y su periferia puede también ubicarse en la zona 6.

Insertamos a continuación el cuadro con la matriz de medias de las variables en las siete zonas, y las medias generales.

Cuadro 1

Medias de las variables en la clasificación inicial (*)

Variable / Zona	1 Bovino	2 Maní	3 Sorgo	4 Frutic.	5 Trigo	6 Patent. de Vehiculos	7 Hotelería.	8 Crecim. Pob.
1	40	26	2	9	5	122	1	102
2	70	1	1	4	22	199	0	1011
3	57	1	2	4	5	200	1	109
4	78	4	3	2	6	223	1	105
5	27	1	0	12	1	213	16	122
6	35	0	3	92	0	148	2	107
7	29	0	2	6	0	60	1	96
Medias Generales	42	4	2	10	4	119	1	101

(*) Se trata de cifras redondeadas.

Fuente: Apéndice I-B. pág. i.

No hacemos aquí un análisis de las características de cada zona, dejándolo para la última etapa de la clasificación.

4.3.1. Verfficación de hipótesis

El valor del estadístico D^2 de Mahalanobis para todos los grupos resulta:

$$D^2 = 2.415 \quad \text{(ver 3.3.1.)}$$

altamente significativo.

En cuanto a las D^2 entre cada par de grupos, puede apreciarse en el Apéndice I-B que aparece como no significativa la correspondiente a los grupos 3 y 4. No obstante se los mantiene separados a fin de observar su comportamiento en las siguientes etapas.

Es de interés señalar el valor del determinante de la matriz de varianzas-covarianzas intra grupos (varianza generalizada intra-grupos). El cual resulta:

$$|S| = 28.10^{13}$$

cuyo valor irá disminuyendo a medida que D^2 aumenta en cada una de las iteraciones.

Efectuado el test sobre igualdad de matrices de varianzas-covarianzas (ver3.4.4 pág. 25) resulta:

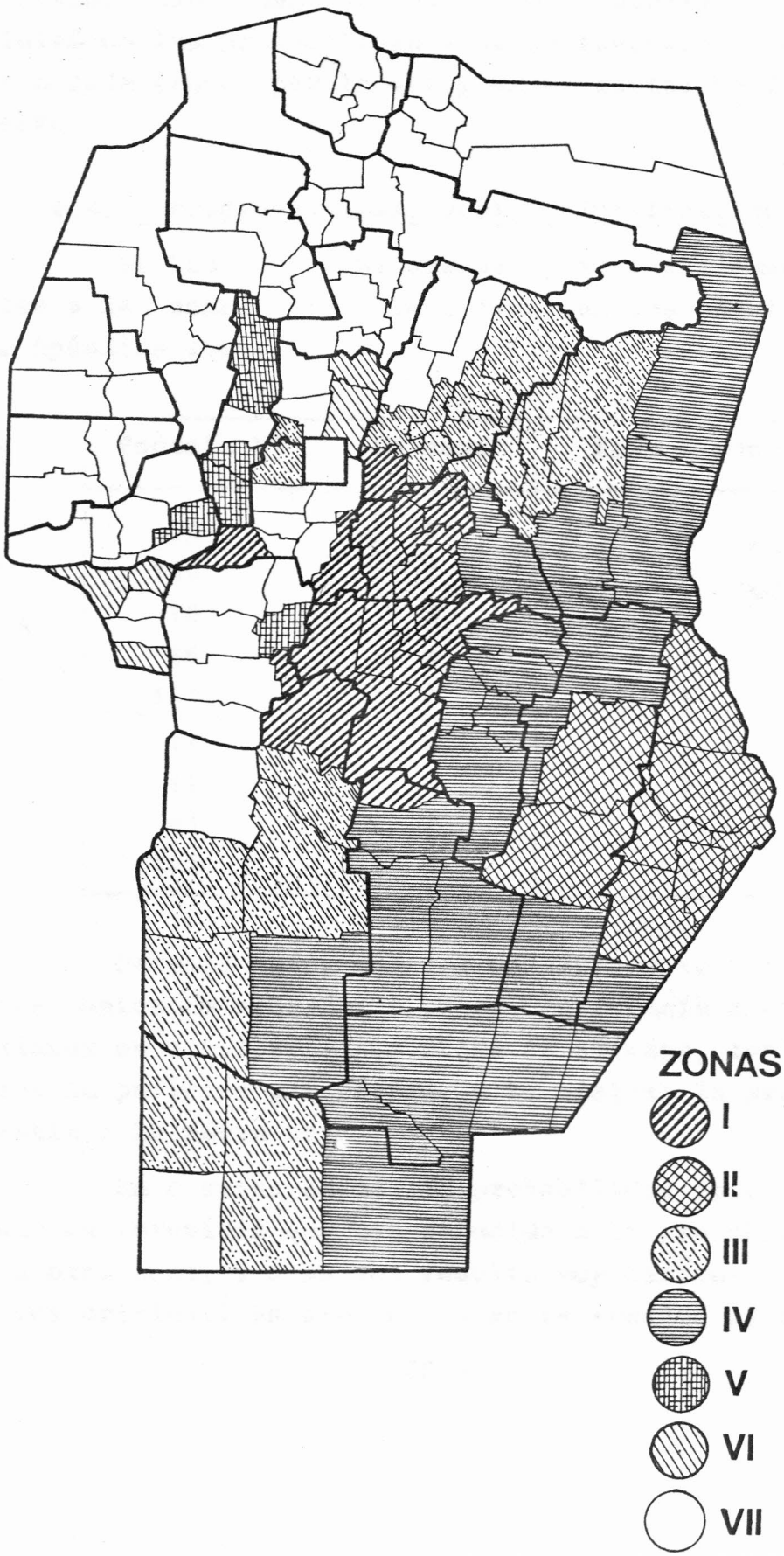
$$F_{216,2342} = 4,69$$

lo que implica el rechazo de la hipótesis.

En cuanto a la hipótesis de normalidad multivariada, podemos afirmar que no se cumple dado el rechazo de la hipótesis univariada para algunas de las distribuciones marginales (Apéndice IV).

No obstante ello, se ha continuado con el trabajo, teniendo en cuenta la coincidencia de los autores al señalar que el

MAPA III
CLASIFICACION FINAL



incumplimiento de las citadas hipótesis afecta la medida de las probabilidades de error, pudiendo realizarse las clasificaciones con buenos resultados (ver 3.7). Debe advertirse que este hecho influirá en las probabilidades de pertenencia de cada observación a cada grupo, por lo que pueden existir variaciones en ese aspecto.

4.4. Iteraciones realizadas y clasificación final

En base al resultado de aplicar las funciones discriminantes a las zonas originales, resultan los siguientes cambios: (ver Apéndice I-B).

Pedanía N°	De zona	Pasa a zona
54	2	4
145	2	4
19	3	7
56	3	4
108	4	3
12	5	7
11	6	7
61	7	5
107	7	3

Debemos notar que los cambios no se hacen en forma automática. Antes de proceder a pasar una Pedanía a otra zona, como lo señalamos en 3.6.3.2, se la ubica en el mapa, teniendo en consideración su posición geográfica, y se analiza la probabilidad de pertenencia a la nueva zona.

En ciertos casos, la probabilidad no es muy elevada, por lo que es verosímil que sea parecida a la probabilidad de pertenencia a otra zona, y a su vez resulta muy alejada geográficamente de su zona original. En esos casos no se realiza el cambio.

En las sucesivas iteraciones (que no transcribimos), nuevos cambios se fueron realizando; los valores de D^2 crecían y simultáneamente decrecía la varianza generalizada intra grupos.

Luego de tres iteraciones, se llegó a la clasificación reflejada en el Mapa 3 y que se describe en el Apéndice I-C.

En ella, $D^2 = 3197$

la varianza generalizada intragrupos:

$$|S| = 76.10^{12}$$

y todas las D^2 entre cada par de grupos resultan significativas.

En este punto es necesario hacer algunas aclaraciones: no se siguió con las iteraciones pese a que vuelven a aparecer pedanías en zonas diferentes a las asignadas:

i) En zona 5 y 6 aparecen respectivamente las observaciones 57 y 101 como transferibles a la zona 7. Estas pedanías en la clasificación original permanecían en su zona. Al retirar otras, el promedio cambió, aumentando la media del grupo en las variables más significativas (Hotelería y Fruticultura respectivamente) por lo cual hay nuevas Pedanías relativamente alejadas de la media. Un análisis del conjunto de variables en dichas observaciones, nos indujo a mantenerlas en esos grupos (por otra parte el retirarlas, nuevamente aumenta la media y se reduce más el grupo).

ii) En la zona 1 aparece la observación 88 pasando a la 3., pero con baja probabilidad; igual ocurre en la zona 3 con la observación 65 que pasaría a la 4. Nuevamente se analizaron las variables más importantes y se optó por dejarlas.

iii) En la zona 7 aparecen las observaciones 75 y 134 pasando a zona 3. La N°134 está geográficamente alejada de la zona 3 por ello no se la pasó y la 75 simplemente por considerar innecesaria una nueva iteración con esa sola modificación. Sin embargo, en la clasificación final (y en el mapa) la hemos incluido en dicha zona.

Cuadro 2

Medias de las variables en la clasificación final

Variable / Zonal	1 Bovino	2 Maní	3 Sorgo	4 Fruticultura	5 Trigo	6 Patent. de Vehiculos	7 Hotelería	8 Crecim. Pob.
1	40	26	2	9	5	122	1	102
2	63	0	1	6	29	231	0	104
3	55	1	5	5	3	160	1	113
4	78	2	2	2	8	222	0	102
5	24	1	0	11	1	191	16	127
6	36	0	4	102	6	135	1	109
7	25	0	1	7	0	48	1	93
Medias Generales	42	4	2	10	4	119	1	101

(*) Se trata de cifras redondeadas.

Fuente: Apéndice I-C pág. i.

Comparando con el Cuadro 1 (pág. 55) se nota la mayor diferenciación entre las zonas en casi todas las variables: i) la variable (bovino) ha disminuído en la zona 2, diferenciandola más claramente de la 4; ii) la variable 2 (maní) ha disminuído en la zona 4 quedando con un valor elevado sólo en la zona 1; iii) la variable 3 (sorgo) ha aumentado en la zona 3 convirtiéndola en la más importante con respecto a ella (aparece también algo más elevada en la zona 6); iv) la variable 4 (fruticultura) se ha elevado en la zona 6 contribuyendo a diferenciarla netamente; v) la variable 5 (trigo) caracteriza a la zona 2 con mayor nitidez que antes, vi) la variable 6 (patentamiento de vehículos) aumenta en las zonas 2 y 4 presentándolas como la de mayor concentración vehicular; vii) la variable hotelería no ha sufrido modificaciones importantes, aunque ya en la clasificación original señalaba claramente una zona; viii) tampoco se ha modificado la variable 8 (crecimien

to de la población) que ya en la primera clasificación mostraba ciertas características destacables.

4.4.1. Características relevantes de cada zona

Zona 1: es la región manicera de la Provincia. Recuérdese que Córdoba produce cerca del 99% del total nacional. Comprende 19 Pedanías, que integran fundamentalmente los Departamentos Tercero Arriba, Río Segundo y algunas Pedanías de Santa María, Calamuchita, Río Primero, General San Martín, J.Celman y Río Cuarto. La existencia de ganado bovino es cercana al promedio provincial y lo mismo ocurre con casi todas las variables (excepto maní) Es una zona de estancamiento en cuanto a población (102 el coeficiente de 1970/1960).

El trabajo del Banco Córdoba (*) sobre tenencia de la tierra en la Provincia muestra a esta zona como una de las menos afectadas por latifundios y minifundios. El coeficiente de Lorenz resulta de los más bajos de la Provincia (Promedio Provincial: 53,6, Departamento Tercero Arriba 39,1; Departamento Río Segundo 42,2).(**)

Zona 2: Es la zona triguera, en la que también se presentan valores elevados para ganado bovino. Pertenece a la pampa húmeda y, coincidiendo con la regionalización de los trabajos del Dr. Ríos y Banco de Córdoba podemos caracterizarla como de explotaciones intensivas mixtas, con predominio de la agricultura.(***) Económicamente es una de las zonas más ricas, hecho que en nuestro trabajo se refleja en el alto índice de patentamiento de vehículos (231 por cada mil habitantes).

(*) Banco de la Provincia de Córdoba (1975)

(**) En varias de las zonas no podemos utilizar este índice para caracterizarlas, porque no ha sido incluído como variable; (no lo disponemos a nivel de Pedanías) por lo tanto algunas zonas son heterogéneas con respecto al mismo.

(***) Ríos R.A. (1968) op. cit.

El índice de crecimiento demográfico es, no obstante, bajo, lo que podría explicarse por el fenómeno general de despoblamiento del agro. Abarca casi todo el departamento Marcos Juarez y parte de Unión (8 Pedanías).

En cuanto a la tenencia de la tierra, siempre según el trabajo citado del Banco de Córdoba presenta un coeficiente elevado (cercano al promedio provincial) para el Departamento Marcos Juarez, con presencia de latifundio y minifundio. El Departamento Unión en cambio, tiene características más parejas en la distribución de la tierra (45,6).

Zona 3 Abarca 19 Pedanías: parte de los Departamentos Río Primero y San Justo por un lado (entre las zonas 1 y 4 por el norte); gran parte del Departamento Río Cuarto y algunas Pedanías de General Roca al sud de la Provincia. Se agregan dos Pedanías en Colón y Santa María vecinas al Departamento Capital. No fué posible establecer esta zona de manera que sean geográficamente vecinas todas las Pedanías.

Presenta ciertas características de las zonas 2 y 4, aunque más pobre que ambas en casi todas las variables. Tiene un índice más elevado que el promedio provincial en ganado bovino. No registra, como la zona 2 importante producción triguera; aparece sin embargo el sorgo, magnitud actual. El índice de patentamiento de vehículos es bastante más bajo que en las otras dos zonas. Curiosamente, registra mayor incremento intercensal en la población. Ello se debe a la presencia de la ciudad de Río Cuarto, que con sus alrededores constituye, según el Censo de 1970 una región con saldo neto de inmigración en el período.

Esta zona es también heterogénea con respecto a la tenencia de la tierra.

Zona 4 Es la zona eminentemente ganadera, presenta la mayor concentración de ganado vacuno, con poca actividad agrícola (sólo la complementaria, cuyas variables no hemos usado). Si bien no hemos considerado la actividad tambera, la misma se concentra

en esta zona en cantidad y calidad.

Presenta elevado índice de patentamiento de automotores y estancamiento demográfico.

Los datos de tenencia de la tierra, muestran distribución pareja en la parte norte de la zona, con un índice de Lorenz que va creciendo hacia el sur. Abarca 23 Pedanías: parte de los Departamentos San Justo, Río Segundo, San Martín, Unión, Juarez Celman, Río Cuarto, General Roca y todo el Roque Saenz Peña.

Zona 5. Es la zona turística de la provincia, que se limita a una pocas Pedanías (seis) situadas en los Departamentos Punilla, San Alberto y Calamuchita. La variable hotelería es la que define esta zona; si bien presenta algo elevado el índice de fruticultura, por comprender algunas regiones de riego. Es elevado el índice de patentamiento de vehículos (posiblemente por patentamiento de vehículos de otras regiones del país que se realizan en la zona) y presenta el mayor índice de crecimiento demográfico intercensal (127). Ello se debe al importante incremento de población producido por saldos netos de inmigración en el Departamento Punilla. Esta zona registra cierta actividad minera, pero no ha sido posible incluir la variable en este trabajo por lo tanto ello no se refleja en el mismo.

Zona 6. Caracterizada por el elevado índice de fruticultura (entre las variables consideradas en este trabajo), es la zona de riego permanente; también se reduce a pocas pedanías (5) en los Departamentos Colón y San Javier. En el Departamento Cruz del Eje hay una zona similar, pero no se ha podido reflejar en nuestra clasificación debido a la gran extensión de la Pedanía que la contiene la cual incluye una vasta zona árida.

Los índices de patentamiento de vehículos y crecimiento demográfico son relativamente bajos. Igual ocurre con los indicadores de otro tipo de actividad agropecuaria.

Zona 7. La constituyen 64 Pedanías, casi toda la región

norte y nor-oeste más zonas serranas áridas y sin desarrollo turístico (gran parte del Departamento Calamuchita y Santa María). Presenta los valores más bajos en todos los índices considerados, debiendo recalcarse un decrecimiento de la población en el período intercensal.

El trabajo mencionado del Banco de Córdoba muestra algunas regiones de esta zona los más altos valores de concentración en la tenencia de la tierra, con la existencia de latifundio y minifundio en elevadas proporciones.

4.5. Análisis de resultados utilizando variables canónicas

4.5.1. Significación de cada uno de los valores característicos de $W^{-1}A$.

Los valores característicos no nulos, resultantes de la clasificación final son los siguientes: (ver Apéndice I-C).

Cuadro 3

Valores característicos de la matriz $W^{-1}A$

Valor Característico		Porcentaje de la traza	Estadístico χ^2 (*)	
Orden	Valor	%	Valor	Grados de Libertad
1	11,2	48,1	336,0	13
2	5,7	24,2	257,4	11
3	2.8	11,9	180,2	9
4	2.0	8,6	149,0	7
5	1.3	5,6	112,4	5
6	0.4	1,6	46,1	3
Suma	23,4	100,0	982,1	48

(*) Se usó la fórmula expuesta en 3.5.2.

Fuente: Apéndice I-C pág. vi.

Existen seis valores no nulos (es decir k-1) y el test de hipótesis efectuado a partir del Cuadro 3 (ver pág. 35) indica que todos ellos son significativos para explicar la varianza entre grupos. La suma de los seis valores (traza de $W^{-1}A$) era de esperar que resultara significativa en base al test ya efectuado con D^2.

Lo curioso de este caso, es que en todas las dimensiones existe diferencia significativa entre las medias, de allí que el análisis en el espacio bidimensional (al que agregaremos una tercera variable canónica) mostrará las características esenciales de las zonas, pero dejará una parte importante sin reflejar.

No obstante, los tres primeros valores característicos absorben el 84,2% de la traza. Es destacable que en la mayoría de las aplicaciones este porcentaje es mucho más elevado. La causa puede residir en que las variables seleccionadas son todas importantes para separar los grupos, y están poco correlacionadas por lo que cada una de ellas incide sobre la discriminación.

4.5.2. Comparación de clasificaciones por ambos métodos

Hemos seleccionado tres variables canónicas para asignar las observaciones en el espacio tridimensional por un problema de capacidad de la computadora. No obstante, los resultados obtenidos difieren en muy pequeña medida con los que surgen al aplicar las funciones discriminantes en el espacio original (siete funciones, de las que son independientes). En el Apéndice I-C, se pueden comparar las clasificación de observaciones mediante las funciones de Fisher -páginas iii y siguientes- y las canónicas -página ix y siguientes-. Debe recordarse que para asignar las observaciones y calcular sus probabilidades, en el primer caso se trabajó sin considerar el tamaño de cada zona (se asignaron iguales probabilidades a priori a todos los grupos) ni las diferencias entre las varianzas-covarianzas de las zonas (que se supusieron iguales) en tanto que en el segundo se tuvieron en cuenta los tamaños de las zonas y sus respectivas matrices de varianzas-covarianzas (ver

3.6.2.). De allí que sea difícil decidir si es mejor uno u otro método sobre todo en este caso en que, en contra del segundo se da la significación de otras dimensiones dejadas de lado y en contra del primero, los diferentes tamaños de los grupos y el rechazo de la hipótesis de igualdad de matrices de dispersión.

Las diferencias observadas son:

i) en la Zona 1, el primer método reasigna la observación 88 (pero con baja probabilidad por lo que la hemos dejado en esa zona) y el segundo mantiene todas las Pedanías originales.

ii) en la Zona 2, no hay diferencias.

iii) en la Zona 3 el primer método reasigna la Pedanía N°65 y el segundo la N°13 (ambas con baja probabilidad; por lo tanto se mantuvieron en la zona 3).

iv) en la zona 4 no hay diferencias.

v) en las Zonas 5 y 6 el segundo método (mantiene todas las observaciones, en tanto el primero reasigna una en cada zona. Por los motivos citados en 4.4. hemos mantenido también estas observaciones en sus zonas.

vi) en la Zona 7 hay una observación (la Pedanía N°75) que es reasignada en ambos métodos y así se ha hecho (pasada a zona 3); existen en el primer método otras dos reasignaciones (115 y 134) que difieren a su vez con dos del segundo (114 y 64).

Por lo expuesto en 4.4 no se han transferido.

4.5.3. Ponderación de cada variable en los vectores característicos estandarizados

Análisis gráfico

Cuadro 4

Vectores característicos asociados a los tres mayores vectores característicos(*)

Componente	V_1		V_2		V_3	
	Normaliz.	Estandar.	Normaliz.	Estandariz.	Normal.	Estand.
1	0.18	25	-0.01	-1	-0.06	-8
2	0.25	11	0.93	39	0.07	3
3	0.06	2	0.06	2	-0.36	-10
4	-0.08	-11	-0.02	-3	0.31	44
5	0.71	25	-0.35	-12	0.31	111
6	0.02	17	-0.00	-5	0.00	-1
7	-0.62	-18	0.12	3	-0.81	-23
8	-0.03	- 6	-0.00	-1	-0.06	-11

(*) Se han redondeado las cifras.

(**)El segundo vector ha sido multiplicado por (-1) a fin de dar ponderación positiva a la variable más importante.

Fuente: Apéndice I-C págs. vi/vii

Cuadro 5

Medias de las zonas en las variables canónicas(*)

Zona / Variable	1	2	3	4	5	6	7
1	15	32	10	20	-6	-3	2
2(**)	21	-12	-1	-2	1	-3	0
3	-3	1	-10	8	-18	22	-6

(*) Se han considerado solo tres por capacidad de computadora

(**) Resultan multiplicadas por (-1) en razón de haber multiplicado por (-1) el segundo vector característico.

Fuente: Apéndice I-C pág. vii.

GRAFICO 1

ELIPSES DE EQUIPROBABILIDAD (90%)
SEGUN LAS DOS PRIMERAS
VARIABLES CANONICAS
MEDIAS Y OBSERVACIONES

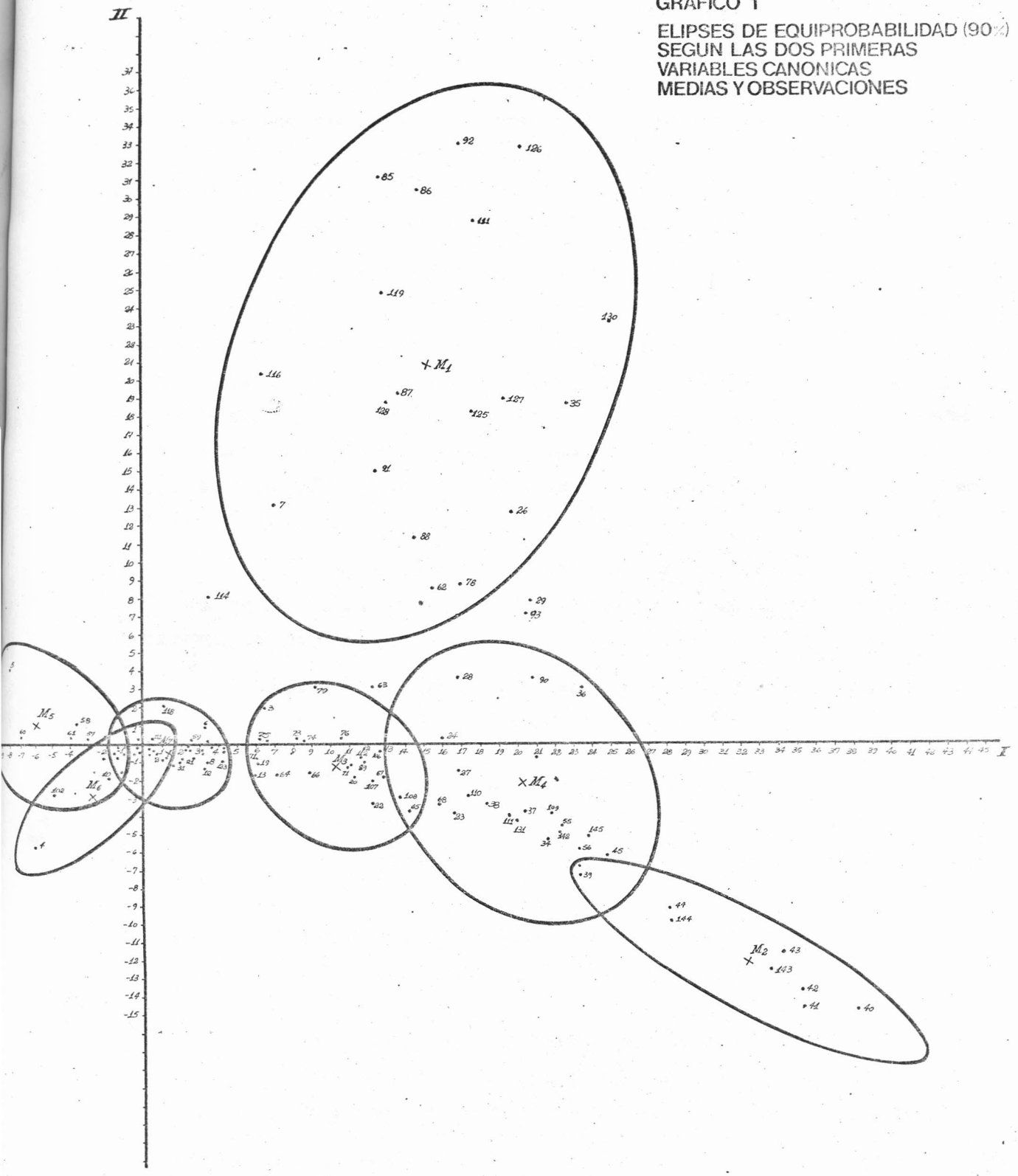
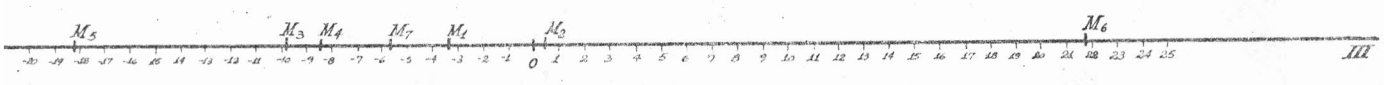


GRAFICO 2

MEDIAS SEGUN LA TERCERA
VARIABLE CANONICA



En el Gráfico 1, se ubican las medias y observaciones de las variables en el espacio bidimensional (de acuerdo a los dos primeros vectores característicos) y en el Gráfico 2 en una tercera dimensión. El cuadro 5 nos presenta los valores de las medias de acuerdo a esas tres variables canónicas.

En el Gráfico 1 se aprecian las diferencias significativas entre las zonas 1, 2, 3 y 4. Las otras tres están más cercanas entre sí, aunque bien diferenciadas del resto. Los dos valores característicos cuyos vectores se reflejan en este gráfico, absorben el 72,3% de la traza. La tercera variable canónica (cuyo valor característico correspondiente absorbe el 11,9% de la traza) permite hacer notar con claridad las diferencias entre las zonas 5,6 y 7.

Ello se explica analizando en el Cuadro 4 la contribución de cada variable a la discriminación en cada uno de los vectores característicos.

Ellos son, en orden decreciente en cada vector:(*)

V_1:

1º) Bovino (con ponderación positiva).
2º) Trigo (con ponderación positiva)
3º) Hotelería (con ponderación negativa).

V_2:

1º) Maní (con ponderación positiva)
2º) Trigo (con ponderación negativa)

V_3:

1º) Fruticultura (con ponderación positiva)
2º) Hotelería (con ponderación negativa)
3º) Crecimiento de Población (con ponderación negativa)

(*) Es muy importante señalar que interesa el valor relativo de cada componente en un vector característico y no su valor absoluto, que puede modificarse sustancialmente con pequeños cambios en los métodos iterativos de cálculo. PUede también multiplicarse por (-1) un vector sin que se altere el significado de sus elementos; cambiará la ubicación de las observaciones en esa variable canónica, pero se mantendrán las distancias y la posición relativa de cada una.

Estas ponderaciones permiten confirmar la importancia de las variables agropecuarias en la clasificación (exceptuando hotelería, que define claramente una zona y crecimiento de población que es coincidente con ella).

Si se observa en los Gráficos la ubicación de las zonas, las conclusiones resultan coherentes con las obtenidas en 4.4.1:

Zona 1: elevado en la segunda variable canónica (maní); y relativamente elevado en la 1 (bovino); cercano a cero en la 3 (fruticultura, hotelería).

Zona 2: elevada en 1 (la más elevada) y altamente negativa en la segunda. Es interesante analizar la ubicación de esta zona; a pesar de que en ganado bovino no tiene un promedio tan alto como la cuatro, aparece mayor su valor en la primera variable canónica; hecho que se explica por la importancia del trigo en la zona 2, (que tiene gran peso en el primer vector característico); el cual está ausente de la zona 4. La posición negativa con respecto al segundo valor característico, también se explica por la variable trigo, que en esta dimensión está ponderada negativamente.

En la tercera variable canónica aparece muy cercana al cero, por la ausencia de fruticultura y hotelería.

Zona 3: Confirma su posición intermedia entre las zonas 2 y 4 por un lado, y las 5, 6 y 7 por otro; sobre todo en cuanto a las variables importantes de la primera y segunda dimensión. En la tercera, aparece con un mayor valor negativo que las otras dos; explicaremos esta circunstancia por el peso relativamente elevado del crecimiento demográfico en la tercera función (con signo negativo), siendo esta zona una de las que presenta mayor aumento intercensal de población. Otro tanto ocurre con sorgo.

Zona 4: Ubicada entre la 3 y la 2 en las tres dimensiones, refleja su elevada existencia de bovino pero sin gran activi-

dad agrícola (como explicáramos al tratar la zona 2). Esa posición en la variable tres puede explicarse porque presenta en sorgo un valor intermedio.

Zona 5: Cercana al cero en la primera y segunda variable canónica, su posición elevada (aunque negativa) en la tercera está dado por la importancia de hotelería y crecimiento demográfico, ambas con ponderación fuerte y negativa en dicha dimensión.

Zona 6: Definida por la fruticultura, la variable con mayor ponderación (y positiva) en la tercera dimensión, se refleja claramente en el Gráfico 2. El resto de las variables, similares a las de la zona 7 explican su posición cercana al cero en el Gráfico 1.

Zona 7: Cercana al cero en las tres dimensiones, ninguna de las variables aquí consideradas presenta en ella valores importantes.

4.5.4. Los coeficientes en las funciones discriminantes de Fisher

Incluímos por último un cuadro con los coeficientes de cada variable en cada una de las siete funciones discriminantes a fin de destacar la coherencia con el análisis efectuado en el punto anterior, en cuanto a las ponderaciones de cada variable en cada una de las zonas.

Cuadro 6

Constantes y Coeficientes de las funciones discriminantes*

Función	Constante	Coeficientes							
		1	2	3	4	5	6	7	8
1	-50	0.3	2.1	0.4	-0.2	0.7	0.0	-0.2	0.3
2	-89	0.5	0.1	0.4	-0.2	3.7	0.0	-1.3	0.3
3	-33	0.3	0.1	1.0	-0.2	0.5	0.0	-0.1	0.4
4	-45	0.6	0.4	0.2	-0.3	1.1	0.0	-0.7	0.3
5	-58	-0.0	-0.1	0.4	-0.1	-0.3	0.0	3.6	0.5
6	-49	0.0	0.2	-0.3	0.7	0.0	0.0	-0.1	0.3
7	-16	0.1	0.0	0.2	-0.1	-0.1	0.0	0.4	0.3

* cifras redondeadas

Fuente: Apéndice 1-C pág. ii/iii

Observando por columnas, los círculos indican en qué zona resulta importante cada una de las variables: bovino en la 2 y 4; maní en la 1, sorgo en la 3, fruticultura en la 6, trigo en la 2 y 4; patentamiento de vehículos en ninguna; hotelería en la 5; crecimiento demográfico en la 3 y 5.

Recordando que mediante estas funciones se adjudica directamente cada observación o la zona para la cual asume el mayor valor la función correspondiente; se entenderá el papel jugado por estas ponderaciones.

5. El problema metodológico. Conclusiones

5.1. Relevancia de la zonificación obtenida, Limitaciones

La zonificación resultante de la presente aplicación de métodos multivariados, pone en primer plano la discusión metodológica acerca del empirismo como método de investigación científica

Resulta claro que esta zonificación refleja parcialmente la realidad de la provincia: al margen de que agreguemos o no

los centros urbanos industriales y la zona minera como dos regiones más, ésta presenta las zonas homogéneas desde el punto de vista de algunos aspectos productivos y demográficos; fundamentalmente los diferentes tipos de productos agropecuarios, la existencia o no de turismo y las tasas de crecimiento demográfico.

No permite, en cambio, una comprensión de las características productivas, económicas y sociales más importantes de cada zona. La ausencia de variables que puedan reflejar esos aspectos, como tenencia de la tierra, grado de mecanización, formas de explotación, número de personas ocupadas en los distintos sectores de la producción, posición de las mismas en el proceso productivo, valor del producto de cada sector, número de empresas y personal ocupado en las mismas; analfabetismo, educación, vivienda, etc.

Como consecuencia, las zonas son homogéneas desde el punto de vista de algunas de sus características productivas y demográficas, pero heterogéneas en cuanto a aquellos aspectos que tal vez son los más importantes para comprender y analizar una región en su situación actual y perspectivas.

Lamentablemente, ésto es lo que ocurre en muchas aplicaciones de los métodos multivariados, y sirve de base para las críticas a ellos formuladas.

En realidad, podemos encarar las críticas metodológicas desde dos aspectos de naturaleza diferente: i) el punto de vista interno de la estadística (limitaciones de los métodos multivariados, falta de conocimiento de algunos puntos, características de las unidades a que se aplica); ii) el punto de vista metodológico general que cuestiona la utilización de la estadística como constructora de teorías (no así como técnica para la prueba de las mismas).

5.2. Críticas desde el punto de vista de la estadística

Galtung (1966) y Kowalsky (1972) tratan de manera muy

completa estos aspectos: el primero, con una profunda discusión acerca de la importante cuestión de si es posible aplicar las técnicas de la inferencia estadística a partir de una población y no de una muestra aleatoria; el segundo, particulariza sobre el uso de los métodos multivariados en la investigación antropológica, pero sus enfoques son aplicables a otros campos.

5.2.1. Muestra o población?

En numerosas oportunidades se aplican el análisis estadístico, los tests de hipotesis y la inferencia, a un universo de datos, como si se tratara de muestras aleatorias. En cierta medida, esestratamiento ha sido el otorgado a nuestras observaciones que constituyen el total de pedanías de la Provincia de Córdoba. Esta cuestión merece una discusión. Galtung (1966)(*) afirma:

> "Por lo tanto, nuestra tesis principal es que los tests estadísticos están fuera de lugar si no tenemos una muestra (también pueden estar fuera de lugar por otras razones). Este es un punto muy debatido, de manera que hay que defender la tesis. Aunque puede haber acuerdo en el sentido de que los test estadísticos son más apropiados para las muestras (esto es para las hipótesis de generalización), muchos investigadores piensan que hay excepciones en que los tests son apropiados también para universos. Vamos a estudiar algunos de estos casos.
> 1. Los datos constituyen un universo, pero son imperfectos.... no hay duda de que las técnicas de la inferencia estadística pueden ser apropiadas. Sin embargo, la condición para utilizarlos es que... los datos deben poder concebirse como una muestra generada de un universo a través de un modelo más o menos especificado... Sin embargo, no existe ninguna razón a priori para suponer que los modelos que pueden funcionar bien para fluctuaciones muestrales también funcionan bien para fluctuaciones en las observaciones. Así, es mucho más fácil obtener independencia entre las unidades al muestrear, que obtener independencia entre las observaciones, debido a un cierto fenómeno de inercia en las mentes humanas y el error tiende más bien a ser sesgado

(*) pág. 434 y sigs. (Tomo II)

que al azar.

Por lo tanto, si existe un modelo par la generación de errores de observación y de medición y se satisface un número razonable de condiciones, se pueden utilizar las técnicas de la inferencia estadística para poner a prueba la hipótesis nula de que los hallazgos se deben a imperfección; sin embargo, no es legítimo pedir prestados inescrupulosamente modelos apropiados para tests de fluctuaciones.muestrales...

2. Los datos constituyen un universo, pero el universo puede concebirse como una muestra de un superuniverso. Esta es la más común línea de defensa cuando se utilizan tests de significación respecto de datos que constituye un universo y parecen existir tres subcasos que deben considerarse separadamente.

a) La muestra es un subuniverso en el espacio. El caso típico consiste en seleccionar una unidad que es una colectividad y tomar una muestra consistente en todos los individuos de esa unidad, y a continuación tratar de generalizar todas las colectividades del mismo tipo. Así, se pueden entrevistar todos los trabajadores de una fábrica y utilizar esto como la base para un estudio sobre los"trabajadores" -queriendo significar todos los trabajadores industriales calificados-. O se pueden muestrear todas las municipalidades de una provincia y tratar de utilizar esto como base para una teoría acerca de las municipalidades en la nación, o todas las provincias de una nación y utilizar esto como una base para un estudio de las provincias de América Latina o del mundo en esesrespecto. El error es obvio. Como un estudio de esa fábrica, esa provincia, esa nación, los datos son perfectos en la medida en que las observaciones o mediciones sean perfectas. Como estudio de las fábricas en general, las provincias en general y las naciones en general, es una muestra de dos etapas en que la primera contiene solamente una unidad (seleccionada al azar o intencionalmente) y la segunda contiene todas las unidades (que entonces puede llamarse un subuniverso).Aunque podemos decir todo lo que queramos acerca del subuniverso, una muestra de una unidad es insuficiente para hacer inferencias cerca del superuniverso, a menos que tengamos alguna información adicional en el sentido de que podemos suponer la homogeneidad entre subuniversos... Para generalizar el superuniverso, la muestra de la primera etapa debe contener más unidades, diseño que habitualmente es llamado "comparativo", "re-

plicación externa"...

b) La muestra es un subuniverso en el tiempo. El caso típico consiste en seleccionar todas las unidades de una cierta clase en un cierto punto o intervalo del tiempo, a menudo llamado "trozo de tiempo". Así, se pueden seleccionar a todos los prisioneros que están en una prisión y sostener que es una muestra del universo total de prisioneros. En este caso se define al universo en intención y no en extensión, no como una lista de elementos en el universo (del cual se puede extraer una muestra), sino como un conjunto de criterios que deben satisfacerse para que algo llegue a ser un elemento. El universo es abierto si pueden generarse nuevos elementos y entrar en él a medida que el tiempo pasa, de modo que en ese sentido el universo seleccionado es un subuniverso y se pueden hacer inferencias al subuniverso o al universo en ese punto del tiempo. Sin embargo para hacer inferencias en cuanto al superuniverso, será necesario remediar los mismos defectos de la muestra en dos etapas con una sola unidad en la primera etapa (un solo trozo de tiempo) o justificando un supuesto de homogeneidad (caso puro) o a través de un muestreo más extenso en la primera etapa...

En todo caso, en la medida en que la condición para aplicar los tests estadísticos no es la existencia de una muestra, sino de una muestra que tenga ciertas características probabilísticas, los universos que son muestras en el tiempo no constituyen una base para las técnicas de la inferencia estadística: como subuniverso, solamente es una muestra de tamaño 1; como muestra tiene m unidades, pero no es una muestra probabilística del universo total, ya que los elementos que están fuera del intervalo en el tiempo tienen una probabilidad 0 de ser incluídas en la muestra...

c) La muestra procede de un universo hipotético. Este es el caso más difícil de analizar. En primer lugar, debe notarse que todo test estadístico supone universos hipotéticos, el universo de la hipótesis nula. Sin embargo, este caso no se refiere al universo del modelo, sino al universo empírico, y dice lo siguiente: si no hay un universo empírico, porque la muestra es idéntica al universo, supongamos que existe y llamémoslo un universo hipotético. Así, si se tienen datos acerca de las fiestas nacionales de muchas naciones del mundo y se está poniendo a prueba la hipótesis de que tienen una tendencia a agruparse en el tiempo de primavera más que en el otoño y en el invierno, ¿se deben aplicar tests estadísticos? Si no tomamos en cuenta la circunstancia irrelevante de que es imposible obtener información respecto de algunas naciones y

suponemos que tenemos datos acerca de todas ellas, diríamos que no. La hipótesis se refiere a las naciones de este mundo tal como existen hoy en día y la tendencia de los datos, haciendo la corrección en cuanto a la diferencia entre los hemisferios norte y sur, es u que más del 70% de los días de fiesta se encuentran en primavera y en verano. En nuestra opinión este es el caso número 1 de la lista dada anteriormente(*)y nos parece que no tiene sentido invocar un universo hipotético, excepto si se trata de legitimar el uso de técnicas que pueden parecer convincentes, a pesar de que son irrelevantes. En efecto, la significación estadística dependería del número de naciones y sería más alta para la misma diferencia porcentual mientras más alto fuera el número de países. Esto significa que la sustentabilidad de nuestra hipótesis acerca de las fiestas nacionales de todas las naciones del mundo dependería de la estructura política del mundo -si el sistema dependería de la estructura política del mundo- si el sistema internacional favorece una división en muchas pequeñas naciones o en unas pocas grandes naciones- Esto nos parece poco razonable, ya que tenemos el universo en nuestros datos... La objeción puede ser rechazada simplemente diciendo que "estas naciones son mi universo y no hay nada más allá de ellas en que yo esté interesado; un descubrimiento basado en ellas es la respuesta final". Estaríamos tentados de aceptar este razonamiento, pero supongamos en todo caso que haya algo válido en la objeción. En primer lugar, admitido esto, de ello no se desprende que las técnicas de la inferencia estadística proporcionen la respuesta al problema de cómo evaluar el grado de confirmación. No hay nada en los datos que corresponda al modelo muestral o al modelo de error. Tal vez existan otras respuestas, en la lógica inductiva, o en otra parte... Por lo tanto, para concluir con una respuesta categórica: nos parece que la idea del universo hipotético con el muestreo hipotético cuando ya tenemos datos acerca del universo que queremos, está mal ubicada, por lo menos en la medida en que nadie ha presentado un buen caso para una definición operacional. Esto concluye nuestra discusión de la aplicación de los tests estadísticos a universos: tal aplicación no tiene sentido sino en algunos casos y cuando se cumplen algunas condiciones".

(*) 1. El caso de los datos perfectos sobre el universo.... En este caso se pueden poner a prueba hipótesis directamente sobre los datos y los tests de hipótesis de estadística no tienen sentido, ya que las hipótesis sustantivas y las hipótesis de generalización coinciden. No hay nada que generalizar pues los resultados ya son generales (pág. 432).

¿Cuál de estos planteos se ajusta a nuestro caso? No creemos que se pueda hablar de ninguno de los formas de subuniverso (ni en el tiempo ni en el espacio); tampoco, por lo expuesto en el punto 1 de Galtung, hablaríamos de datos imperfectos sobre el universo. Nos inclinamos por el último de los enfoques: si nos interesa solamente la provincia de Córdoba, y allí se agota nuestra investigación, de ella obtenemos nuestra respuesta final (la zonificación) y no deseamos inferir nada acerca de un universo mayor. Pero entonces ¿cuál es el papel de las pruebas estadísticas? Hemos efectuado tests sobre diferencia de medias, basados en D^2. Las diferencias expresadas por dicho estadístico son válidas aún en el supuesto de que se trate de la población (ver su definición original). y la realización de los tests sólo serviría para tener una idea de la magnitud de las diferencias por él expresadas. El no cumplimiento de los tests de normalidad y de igualdad de matrices de varianzas covarianzas, podría deducirse, al tratarse de una población, del sólo análisis de los datos y permite alertar acerca de la posibilidad de errores en la clasificación mediante el uso de las funciones discriminantes construídas para poblaciones normales y con dispersiones comunes. Posiblemente sería mayor la dificultad si se hubieran aceptado las hipótesis en esos casos: se trata de la población, los datos permiten constatar que las matrices de varianzas-covarianzas no son iguales, y los tests en ese sentido indican que las diferencias no son significativas . ¿Cómo actuar? concidimos con Galtung al afirmar que las técnicas estadísticas no sirven para evaluar la magnitud de las diferencias; por el contrario, creemos que, sin tener el significado estricto: "no hay elementos para rechazar...", una diferencia no significativa indicará que las matrices poblacionales difieren poco, y esa conclusión servirá a los efectos de la confiabilidad que nos merece la clasificación.

En cuanto a los tests realizados con los valores ca-

racterísticos para determinar el número de dimensiones en que las diferencias resultan significativas y los correspondientes gráficos en una o dos dimensiones, son válidos a los efectos de analizar la estructura de los datos, aún cuando se trate de población y no de muestra. Gandesikan y Wilk (1968) hablan, para el empleo de diversos métodos de agrupamiento de datos (incluyendo componentes principales, análisis factorial y función discriminante), de la "colección" de datos o "de una muestra".

Recordamos por último, que las funciones discriminantes pueden ser construídas en base a datos muestrales o poblacionales.

En conclusión, creemos posible en este caso la aplicación de ciertos tests y sobre todo la utilización de las técnicas del análisis multivariado para datos poblacionales. No obstante, todo lo señalado más arriba deberá ser contemplado en cada una de las aplicaciones particulares a fin de no incurrir en este tipo de errores.

5.2.2. El problema del análisis multivariado

Varios autores han señalado críticas y sobre todo limitaciones al uso de técnicas estadísticas multivariadas, especialmente teniendo en cuenta el estado aún primitivo de la teoría correspondiente.

Kowalsky (1972) pasa revista a casi todas estas críticas, al referirse en especial al campo antropológico, llamando la atención sobre el uso generalizado de la estadística multivariada sin reparar en las complicaciones de interpretación (sobre todo en componentes principales), correlaciones canónicas y análisis factorial) ni en el poco conocimiento de la potencia de los tests utilizados así como de su campo de aplicación. Plantea este autor que muchas veces resultaría más claro un análisis univariado o a lo sumo bivariado que el complicado tratamiento de múltiples datos. Compartimos su preocupación por la interpretación de las combinaciones lineales de variables que suponen las

técnicas mencionadas. Transcribimos algunas de sus citas: con respecto a Componentes Principales, cita a Kendall (1972)(*), quien comentando un análisis de componentes principales afirma:"La característica notable del trabajo de Stone, sin embargo, es que pudo interpretar sus componentes. En muchos casos nuestras componentes principales no tienen una existencia separada identificable y deben ser consideradas como entidades matemáticas convenientes. En otros, es discutible si puede darse a las componentes alguna realidad".

Más adelante, y refiriéndose a análisis factorial, cita a Cureton: "La teoría de factores puede ser definida como la racionalización matemática. Un analista de factores es un individuo con una obsesión particular referente a la naturaleza de la capacidad mental o personalidad. Con la aplicación de altas matemática y elaboración personal, él siempre comprueba que su original idea fija o compulsión era correcta o necesaria. En los procesos, generalmente comprueba que todos los otros analistas de factores están peligrosamente insanos y que la única salvación para ellos es soportar su propio tipo de análisis para que la verdadera esencia de sus distintas enfermedades pueda descubrirse..."

Con respecto a las correlaciones canónicas, nuevamente cita a Kendall (1972)(**) "Cuando se llega a la etapa de la interpretación encontramos la misma dificultad que ya habíamos encontrado en el análisis de componentes: la de saber si nuestras funciones lineales corresponden a algo "real" o si son meramente cuestiones de conveniencia matemática".

En efecto, algunas aplicaciones de estas técnicas nos han mostrado la enorme dificultad de interpretación, así como la poca estabilidad de los resultados numéricos, ante pequeños cambios en la tolerancia para las iteraciones o simplemente de métodos de resolución (que son casi todos numéricos)

(*) pág. 26

(**) pág. 82

para las ecuaciones características y la obtención de los vectores.

No compartimos en cambio, la reticencia a la utilización de estos métodos desde el punto de vista puramente estadístico, yaque toda técnica sufre un proceso de depuración y perfeccionamiento, pudiendo en el interin utilizar la misma, proceso que a su vez ayuda a perfeccionarla. No obstante, es necesario tener presentes en todo momento las limitaciones. En este sentido, creemos de mayor importancia las críticas metodológicas generales al uso de la estadística como constructora de teorías, que consideramos en el punto siguiente.

Con respecto a la función discriminante, Kowalsky plantea como limitación la necesidad de suponer que se conoce a priori el número de grupos en los cuales se ejecutará el análisis. Esto, más que una limitación, resultaría una ventaja para su aplicación al campo de las ciencias sociales, donde el modelo previo es no solo necesario sino imprescindible.

5.3. El problema metodológico general

El origen de la discusión se encuentra en la tendencia a utilizar la recolección y tratamiento estadístico de los datos como punto de partida del conocimiento científico.(*). Mario Bunge (1969) crítica seriamente esta orientación, frecuente sobre todo en las ciencias sociales, a lo largo de toda su obra.

Refiriéndose a las líneas de regresión y a los coeficientes de correlación, hace un planteo que muy bien puede aplicarse a las técnicas multivariadas.

(*) Posiblemente los trabajos más ambiciosos de aplicación del análisis multivariado en el ámbito económico son los de Adelman y Morris ya citados. Es por ello que sobre esa base discutimos los aspectos metodológicos más importantes vinculados a estas técnicas. Haan (1972) realiza un correcto resumen de dichos trabajos.

"En cualquier caso, el cálculo de coeficientes de correlación y el trazado de líneas de regresión no debe confundirse con un método para hallar leyes, confusión tan frecuente en las ciencias sociales. Cuando se adopta un modelo de regresión lineal y se calculan los parámetros a partir de los datos, la ley central que se supone rige esa información "ruidosa" (dispersa) no se ha descubierto, sino que se ha supuesto desde el principio. No hay elaboración de datos estadísticos que produzca por sí misma nuevas hipótesis, por no hablar ya de leyes; en general, no hay esfuerzo técnico, por grande que sea, ni empírico ni matemático, que pueda ahorrarnos el trabajo de inventar nuevas ideas, aunque sin duda aquel trabajo técnico puede muy bien disimular la falta de ideas."(**)

Por su parte Haan (1972) critica los métodos multivariados usados por Adelman y Morris porque son más "constructores de teorías" que "prueba de teorías". Más adelante, al extraer conclusiones sobre estos trabajos (los que tratan de clasificar los países subdesarrollados según su potencial de desarrollo a fin de dirigir la "ayuda" de manera que la productividad del capital extranjero invertido en ellos sea mayor)(***), Haan objeta el uso de estos métodos por no considerar importantes aspectos como el tiempo histórico, la realidad mundial cambiante, la ausencia de hechos históricos relevantes. Dice luego:

"No puede dejar de repetirse que las relaciones causales no pueden ser verificadas por el investigador. El análisis estadístico solo muestra interdependencias o asociaciones que pueden ser interpretadas por el analista como relaciones causales.
Como podría esperarse, éste es el aspecto más debil del trabajo. Primero, es dificil, si no imposible,

---

(**) Rag. 347.

(***) A.y Morris (1968-a) En la discusión alrededor de los mismos Rayner (1970) enfoca la crítica desde el punto de vista estadístico, defiende las variables originales en lugar de las combinaciones lineales, las regresiones en lugar de las técnicas multivariadas. En cambio Eckstein (A.yMorris 1970-a; pág.227) va más al fondo de la cuestión:le interesan las variables según el papel que juegan en el desarrollo, y no porque "discriminen mejor".

derivar relaciones causales de los datos "cross-section" porque no pueden usarse rezagos. El uso por A. y M. de cinco relaciones mutuamente causales, ésto es, algunos pares de variables aparecen dos veces en una ecuación, una como variable dependiente y la otra como independiente, es bastante insatisfactoria. Esto una vez más ilustra las limitaciones de la investigación cuantitativa si se conoce poco del campo que se investiga..."(*)

Conclusiones en el mismo sentido se extraen de Castells (1972) cuando, criticando la orientación empirista en la sociología dice:

"En efecto, la perspectiva empirista tradicional, para la cual la "teoría" es el resultado de una interpretación basada en el análisis de datos en su existencia objetiva y autónoma, no requiere sino un proceso de puesta en relación con el fin de organizar la materia prima de esta realidad social directamente aprehendida... Ahora bien, es justamente esta neutralidad teórica de los datos la que se cuestiona totalmente a medida que se profundiza la investigación metodológica.

En un primer nivel de la crítica, se ha podido constatar que la recopilación de los datos descansa siempre sobre una categorización previa de estos datos, ya sea según categorías "ad hoc", de acuerdo con los propósitos de la investigación, o según términos de la práctica administrativa, o más generalmente, del lenguaje común. Es evidente que esta categorización reviste a los datos de un contenido teórico o ideológico, sea por el cuadro conceptual del investigador o por las connotaciones culturales ligadas a los términos del lenguaje. Por lo tanto, sin una reelaboración teórica, se torna imposible controlar los efectos de este contenido introducido en los resultados de la investigación por la codificación previa de los datos".(**)

Esta situación aparece de manera muy concreta en los trabajos de Adelman y Morris, señalada por Haan como un "sesgo occidental" en los mismos, al proponer como modelos de desarrollo a los países actualmente desarrollados. Las diferencias profundas que existen entre los hoy llamados países del "tercer mundo" y los desarrollados han sido señaladas por varios autores. Lebedinsky

---

(*) pág. 27

(**) pág. 140.

(1968), citando a Bethelhein señala algunas de ellas: aquellos países no eran dependientes, su vida autónoma no fué amenazada ni dominada por una potencia externa; su desarrollo era, en general, diversificado, con agricultura y artesanías que más tarde transformaron en industria y con gran auge del comercio; no suffían penetración extranjera en su comercio exterior, ni dependían del exterior para aprovisionarse de maquinarias; no existía competencia internacional en industrias; desarrollaron una política colonial de saqueo hacia numerosos países.

Por ello es que un análisis puramente estadístico, basado en datos como el ingreso "per cápita", la participación de los sectores primario, secundario y terciario en el producto, las características del comercio exterior, ocultan aspectos como la real distribución del ingreso, la propiedad de las empresas, de la tierra, de los resortes del comercio exterior, la deserción escolar, los problemas sanitarios, la dependencia tecnológica, etc. De allí que pueda llegar a decirse que estos métodos reflejan ligazones aparentes entre los fenómenos y no las relaciones internas, profundas, esenciales.(*)

Un ejemplo de las conclusiones equivocadas que pueden extraerse de este manejo mecánico de datos, lo da nuestro trabajo en el crecimiento demográfico reflejado en la zona 3. Ello estaría indicando que se trata de una región en desarrollo. Sin embargo, estudiando más a fondo las características de la misma, resulta que el crecimiento se verifica en la Ciudad de Río Cuarto, mientras las zonas rurales reflejan decremento o tasas muchos menores de crecimiento intercensal. ¿Puede considerarse un índice de desarrollo el despoblamiento del campo? Una respuesta completa requeriría un análisis que sale de los marcos de este trabajo. Pero lo que es evidente, es que la respuesta no puede surgir de los datos numéricos sino de un desarrollo teórico (complementado por cierto con datos).

(*) Lebedinsky (1970.

Nuevamente citamos a Bunge acerca de la importancia de la elaboración teórica para el desarrollo de una ciencia:

> "La dimensión y la adecuación relativas del trabajo teorético, miden, pues, el grado de progreso de una ciencia, al modo como la dimensión y la eficiencia relativas del sistema nervioso son un índice del progreso biológico. Por esta razón la psicología y la sociología, a pesar de su enorme acervo de datos empíricos y generalizaciones de bajo nivel, siguen considerándose aún en un estadio subdesarrollado: porque no abundan en teorías lo suficientemente amplias y profundas como para dar razón del material empírico disponible. Pero en esos como en otros departamentos de la investigación, la teorización se considera frecuentemente como un lujo, y no se admite como ocupación decente más que la recolección de datos, o sea, la descripción. Y esto hasta el punto de que está de moda en esas ciencias oponer la teoría (como especulación) a la investigación (entendida como acarreo de datos). Esta actitud paleocintífica, sostenida por un tipo primitivo de filosofía empirista, es en gran parte la causa del atraso de las ciencias del hombre. En realidad, ese punto de vista ignora que los datos no tienen sentido ni pueden ser relevantes más que en un contexto teorético, y que la acumulación al azar de datos, e incluso las generalizaciones que no son más que condensaciones de datos, son en gran parte pura pérdida de tiempo si no van acompañadas por una elaboración teorética capaz de manipular esos resultados brutos y de orientar la investigación. No se puede saber si un dato es relevante si no se es capaz de interpretarlo; y la interpretación de datos requiere el uso de teorías. Además, solo las teorías pueden sugerir la búsqueda de información no suministrada espontáneamente por los sentidos..."(*)

5.4. Conclusiones

Después de los planteos efectuados en este capítulo, ¿debe concluirse en que los métodos multivariados no resultan de ninguna utilidad para la investigación? Evidentemente esa no es nuestra respuesta. Estos métodos, como toda técnica estadística, son aplicables en determinadas circunstancias y dentro de contex

---

(*) pág. 416. Véase también pag. 754 y sigs.

tos adecuados.

En particular, para el problema de la clasificación de observaciones en grupos homogéneos, más aún, para la zonificación u otras cuestiones similares dentro del campo económico, entendemos que la función discriminantes con los procedimientos iterativos aplicados en este trabajo, son adecuados.

Se requiere, por cierto, el modelo previo de zonificación basado en un análisis teórico-empírico de la realidad; la selección de variables debe ser correctamente efectuada.

La aplicación del método sobre esa base, perfecciona la clasificación original, optimizando la misma desde el punto de vista de maximizar la variabilidad entre los grupos con respecto a la intra grupos, manteniendo al mismo tiempo las características básicas iniciales. Este método no presenta, por otra parte, las cuestiones de interpretación arbitraria de combinaciones lineales de variables que surgen de la aplicación de componentes principales, análisis factorial o correlaciones canónicas.

Sin embargo, no sería correcto descartar totalmente las otras técnicas. Existen ejemplos valiosos de las mismas: una aplicación de componentes principales para zonificar las regiones agrarias de Polonia (Brown y Trott-1968); el trabajo de Naciones Unidas (1972) que aplica métodos numéricos y análisis factorial para clasificar los países de América Latina en grupos homogéneos. Ambos presentan la característica de contar con una buena selección de variables que ha permitido tener en cuenta los aspectos más relevantes para el objetivo de las clasificaciones.

Respecto de la presente aplicación, resulta limitada fundamentalmente por la falta de datos referidos a las variables necesarias para reflejar la estructura socio-económica de las diferentes regiones de la provincia. Partiendo de un esquema similar, pero contando con aquella información, puede llegarse a resultados más fructíferos.

APENDICE I-A

PROGRAMAS DE COMPUTO

Se trata de siete programas encadenados, en los que frecuentemente se utilizan subrutinas del Scientific Subroutine Package (SSP) preparado para la computadora IBM 1130.

A continuación insertamos un esquema de los pasos seguidos por esos programas.

I. Simbología utilizada:

1. k número de grupos
2. p número de variables
3. n_g número de observaciones en el grupo g.
4. $N = \sum_{g=1}^{k} n_g$ número total de observaciones.
5. x_{ign} = valor de la variable i-ésima en la n-sima observación del grupo g.
6. $\bar{X}_{ig} = \frac{\sum_{n=1}^{g} x_{ign}}{n_g}$ $g = 1, 2, \ldots, k;\ i = 1,2, \ldots, p$

   media de la variable i-ésima en el grupo g.
7. $\bar{X}_i = \frac{\sum_{n=1}^{N} x_{in}}{N}$ $i = 1, 2, \ldots, p$

   media de la variable i-ésima).

II. Programas

1) IGVAG

i) Calcula e imprime la matriz de medias de las observaciones originales en cada grupo, la matriz de productos cruzados de desviaciones promedio intra grupos (matriz D) (o sea la matriz de varianzas-covarianzas intra grupos, que se supone común para todos)(*) y el vector de medias generales media de cada

(*) La matriz D es la que en el trabajo hemos simbolizado con S y D_g corresponde a S_g (pág. 18).

variable para el conjunto de las observaciones). Se utiliza la subrutina DMATW, similar a la DMATX del SSP con una pequeña modificación que permite obtener y grabar un disco para ser utilizada a posteriori, la matriz W (suma de productos cruzados de observaciones entre grupos).

Los elementos de la matriz de medias son de la forma:

$$\bar{X}_{ig} = \frac{\sum_{n=1}^{n_g} x_{ign}}{n_g} \qquad \begin{matrix} i = 1, 2, \ldots, p \\ g = 1, 2, \ldots, k \end{matrix}$$

Los de la matriz W:

a) para cada grupo: $[W_g]$

$$W_{ij}^{(g)} = \sum_{n=1}^{n_g} (x_{ign} - \bar{x}_{ig})(x_{jgn} - \bar{x}_{jg}) \qquad i,j = 1,2,\ldots, p$$

b) para todos los grupos

$$W = \sum_{g=1}^{k} W_g \quad \text{ó} \quad W_{ij} = \sum_{g=1}^{k} W_{ij}^{(g)}$$

Los de la matriz D:

$$d_{ij} = \frac{W_{ij}}{N-k}$$

Los del vector de Medias Generales:

$$\bar{X}_i = \frac{\sum_{n=1}^{N} x_{in}}{N} \quad i=1,2,\ldots, p \quad \text{ó}$$

$$\bar{X}_i = \frac{\sum_{g=1}^{k} n_g \bar{X}_{ig}}{N} \quad i=1,2,\ldots, p$$

ii) Calcula e imprime los determinantes de la matriz D y D_g (g=1,2,...,k) es decir, matrices de dispersión de cada grupo. A fin de efectuar el test sobre igualdad de matrices de varianzas-covarianzas. Para calcular estos determinantes se utiliza la subrutina MINVI (idéntica a MINV pero con precisión simple): mediante esta subrutina también se obtiene la inversa de la matriz D. Todas las matrices y valores necesarios para cálculos posteriores se graban en disco.

2) IGVA2

i) Efectúa el test de igualdad de varianzas, según las fórmulas de páginas 28 /29

ii) Calcula D^2 de Mahalanobis para todos los grupos y para cada par de ellos, incluyendo el estadístico correspondiente para docimar hipótesis sobre igualdad de medias.

Para D^2 general, se utiliza la fórmula (25) de pág.22. Para D^2 entre cada par de grupos, la fórmula (16) de pág. 16.

3) ADISC

Calcula e imprime las funciones discriminantes según la subrutina DISCR del SSP. Se trata de k funciones que tienen una constante y p coeficientes de las variables.

La constante se calcula según:

$$C_{og} = -\frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij} \bar{x}_{ig} \bar{x}_{jg} \quad g=1,2,\ldots, k$$

Siendo d_{ij} los elementos de la matriz D^{-1}

Los coeficientes:

$$C_{ig} = \sum_{j=1}^{p} d_{ij} \bar{x}_{jg} \quad i=1,2,\ldots, p$$

Estas fórmulas son equivalentes a las expresadas en

forma matricial en (23) pág. 21).

Para cada observación de cada grupo se calcula (ver (24) pág. 21).

$$U_{gi} = \sum_{j=1}^{p} C_{jg} x_{ijg} + C_{og} \qquad \begin{matrix} g=1,2,\ldots, k \\ i=1,2,\ldots,n_g \end{matrix}$$

Adjudicándose la i-ésima observación al grupo para el cual el valor de U_{gi} es mayor (sea el grupo M).

Se calcula también la probabilidad de pertenencia a dicho grupo mediante la fórmula:

$$P_M = \frac{1}{\sum_{g=1}^{k} e^{(U_g - U_L)}}$$

(ver pie de pág. 40)

4 y 5) DISCO y DISC2

Programas que, utilizando la subrutina CORRA (CORRE con precisión simple) calculan la matriz suma de productos cruzados de desviaciones total (llamada RX en el programa y T en este trabajo ver (27)) y la matriz de correlaciones entre las variables originales que también se imprime:

$$r_{ij} = \frac{t_{ij}}{\sqrt{t_{ii}}\sqrt{t_{jj}}} \qquad i,j = 1,2,\ldots, p$$

Siendo $t_{ij} = \sum_{n=1}^{N} (x_{in} - \bar{x}_i)(x_{jn} - \bar{x}_j)$

6) VPROP

i) Calcula la matriz suma de productos cruzados de desviaciones entre grupos (28) llamada A en el trabajo y DEG en el programa, efectuando la diferencia:

DEG=RX-W (A=T-W en el trabajo)

ii) Encuentra los valores y vectores característicos de W^{-1}DEG (se utiliza la subrutina NROOT para ello) se estandarizan los vectores propios normalizados, multiplicando cada componente por la raiz cuadrada de los elementos diagonales correspondientes de W (ver 3.6.1. última parte). Se simbolizan con V_i cada uno de esos vectores estandarizados.

iii) Calcula el test Λ de Wilk para diferencia de medias y el porcentaje de cada raíz característica sobre la traza. (ver pág. 27).

7) MEDIE

Calcula matriz de medias y de dispersión en el espacio reducido, la dimensión del cual debe indicarse por consola, una vez analizados los resultados del programa anterior (ver 3.4.2) Por una cuestión de capacidad de máquina, que restringe el programa siguiente, dicha dimensión no podrá ser mayor de 3.

Las fórmulas utilizadas son las indicadas en 3.6.1.

Para la matriz de medias, cada columna se calcula

$$\tilde{\bar{f}}_g = V'\tilde{\bar{X}}_g \qquad g=1,2,\ldots, k$$

Las matrices de dispersión de cada grupo surgen de:

$$D_g = V'S_gV \quad (*) \qquad g = 1,2,\ldots, k$$

8) CLASI

i) Calcula el valor de cada observación en las dos primeras variables canónicas y los imprime, a fin de posibilitar la representación gráfica del conjunto de observaciones.

ii) Clasifica las observaciones según sus valores en

(*) Las S_g se han simbolizado en programas anteriores con D_g y así están grabadas en disco.

el espacio reducido, utilizando para ello el criterio definido en 3.6.2., mediante la fórmula (37).

Es decir, se calculan para cada observación la probabilidad asociada a cada grupo:

$$P(P_g/f) = \frac{\frac{q_g}{|D_g|^{1/2}} \exp - \frac{1}{2}\left[(f-\bar{f}_g)' D_g^{-1} (f-f_g)\right]}{\sum_{i=1}^{k} \frac{q_i}{A^{1/2}} \exp - \frac{1}{2}\left[(f-f_i)' D_i^{-1} (f-f_i)\right]}$$

se imprime el grupo para el cual presenta la probabilidad más elevada. Imprime también el valor de esa probabilidad.

LISTADO DE PROGRAMAS

```
      DIMENSION PR(2),A(1305),N(7),CMEAN(9),IO(145),XBAR(63),D(81),W(81
     *),NI(9),XBAG(9),DET1(7),N5(1)
      COMMON MX,MY
      DEFINE FILE 1(30,200,U,K1),2(100,320,U,K2),3(1,100,U,K3)
C     LAS DIMENSIONES DEBEN SER PR(2) EN LAS SIGUIENTES, EL GUION INDICA
C     'MAYOR O IGUAL QUE'. N-N,CMEAN-N,IO-NT,A-NT*M.XBAR,D,W,RX,R,PROD-M
C     *MXBAG-M,STD-M,B-M,C-M,DET-M,XBAD-K*M,XBAT-M*K,DIST-L*K,SSUM-M.
C     KMEAN/M,DESV-M
C     ESTAS DIMENSIONES VALEN PARA EL PROGRAMA 1-IGVAC Y EL 2-IGVA2-
C     ESTE PROGRAMA CALCULA MATRICES NECESARIAS PARA EFECTUAR EL ANALI
C     SIS DISCRIMINANTE Y EFECTUA TEST DE IGUALDAD DE MATR.DE VAR-COVAR.
C     CALCULA D CUADRADO DE DE MAHALANOBIS PARA TODOS LOS GRUPOS Y PARA
C     CADA PAR DE GRUPOS,LO QUE PERMITE TESTEAR IGUALDAD DE MEDIAS.
C     ANTES DE LOS DATOS SE COLOCARA UNA TARJETA DE PARAMETROS QUE CON
C      TENDRA PR(NOMBRE DEL PROGRAMA),K(NUMERO DE GRUPOS),N(I)(NUMERO DE
C     OBSERVACIONES EN CADA GRUPO),SEGUN FORMATO 1.LOS DATOS SE LEEN SE
C     GUN FORMATO 3.EN LAS ULTIMAS 3 COLUMNAS VA EL NUMERO DE ORDEN DE
C     CADA OBSERVACION.
C     DEBEN CARGARSE LAS SIGUIENTES SUBRUTINAS(DEL MANUAL DE SUBRUTINAS
C     CIENTIFICAS PARA LA IBM 1130),GMSUB,NROOT,EIGEN,DISCR,GMPRD,GMTRA.
C     TODAS CON PRECISION SIMPLE. A LAS QUE SIGUEN SE LES HA MODIFICADO
C     EL NOMBRE PARA SER CARGADS CON PRECISION SIMPLE.MINV1(MINV),MXSAL (MXOUT),
C     LOS(LOC. EN TARJETA NUM.26 DE MXOUT DEBE DECIR CALL LOS)
C     CORRA(CORRE),DATA.DMATW(DMATX, AGREGANDO DIMENSION W(1) Y ANTES DE
C     LA INSTRUCCION NUM. 180, W(I)=D(I)),Y M COMO ARGUMENTO
C     UBICACION DE LAS MATRICES Y VARIABLES EN LOS ARCHIVOS.ARCHIVO K1.REG.1,
C     MATRIZ XBAR,2-D,3-W,4-RX,5-14-D(I),15-XBAT,16-XV,17-DET1,18-XBAG,19-
C     D-INVERSA,20 Y SIGS.-DET.MATR.COV.EN EL ESP.RED.
C     EN EL PROGRAMA MEDIE,SE UBICAN EN K1=1 LA MATR. DE MEDIAS EN EL ESP.
C     REDUCIDO Y EN K1=5 Y SIGS.LAS DE DISPERSION EN ESE ESPACIO.
C     ARCHIVO K2.REG.1,MATRIZ DE DATOS A, 2-IO,ARCHIVO K3-REG.1,K(NUMERO
C     DE GRUPOS,2-M(NUM.DE VARIABLES),3-NT (TOTAL DE OBSERV.),4 Y SIGS.-N(I),
      MX=1
      MY=2
      WRITE(1,800)
  800 FORMAT(////)
      READ(2,1)PR,K,M,(N(I),I=1,K)
    1 FORMAT(2A4,1X,212,1015)
      WRITE(1,2)PR,K,M,(N(I),I=1,K)
    2 FORMAT('ANALISIS DISCRIMINANTE',1X,2A4,1X,'NUMERO DE GRUPOS',1X,I2
     */'NUMERO DE VARIABLES',2X,I3,2X,'NUMERO DE OBSERVACIONES EN CADA
     *GRUPO',1015)
C     LECTURA DE DATOS
      L=0
      NT=0
      DO 110 I=1,K
      N1=N(I)
      NT=NT+N1
      DO 100 J=1,N1
      JK=NT+J-N1
      READ(2,3)(CMEAN(IJ),IJ=1,M),IO(JK)
    3 FORMAT(5F7.3,F7.2,F4.2,F3.0,28X,I3)
```

```
      L=L+1
      N2=L-N1
      DO 100 IJ=1,M
      N2=N2+N1
  100 A(L2)=CMEAN(IJ)
  110 L=N2
C     NT ES EL NUMERO TOTAL DE OBSERVACIONES
      K3=1
       WRITE(3'K3)K,M,NT,(N(I),I=1,K)
C     POR LA SUBR. DMATW SE CALCULA LA MATRIZ DE MEDIAS DE CADA VARIABLE
C     EN CADA GRUPO,LA MATRIZ DE SUMA DE PROD.CRUZ.DE DESVIACIONES Y DE
C     PROMEDIO DE PROD. CRUZ.DE DESVIACIONES INTRA-GRUPOS.POR MINV1 SE
C     OBTIENE EL DET.DE LA ULTIMA MATRIZ MENCIONADA,NECESARIO PARA CALCU
C     LAR D CUADR.Y HACER LOS TEST DE IGUALD.DE MATR.DE VAR-COVAR.
C     GRABAR EN DISCO MATRIZ DE DATOS,DE MEDIAS(XBAR),DE PROD.CRUZ.PROM.
      NTT=NT*M
      KM=K*M
       MM=M*M
       CALL DMATW(K,M,N,A,XBAR,D,W,CMEAN)
      K2=1
       WRITE(2'K2)(A(J),J=1,NTT)
      WRITE(2'K2)(IO(J),J=1,NT)
      K1=1
       WRITE(1'K1)(XBAR(J),J=1,KM)
      WRITE(1,4)
    4 FORMAT(//'MATRIZ DE MEDIAS,CADA COLUMNA INDICA UN GRUPO,CADA FILA
     *INDICA UNA VARIABLE')
       CALL MXSAL(    1,XBAR,M,K,0,60,120,1)
      WRITE(1'K1)(D(J),J=1,MM)
      WRITE(1,5)
    5 FORMAT(//,'MATRIZ DE PROD. CRUZ.DE DES♥.PROMEDIO INTRA GRUPOS-D-')
       CALL MXSAL(    2,D,M,M,0,60,120,1)
      WRITE(1'K1)(W(J),J=1,MM)
       CALL MINV1(D,M,DET,CMEAN,MI)
       K1=19
       WRITE(1'K1)(D(J),J=1,MM)
      DO 101 J=1,M
  101 XBAG(J)=0.0
      DO 103 IM=1,M
      DO 102 J=1,K
      JK=(J-1)*M
      JL=JK+IM
  102 XBAG(IM)=XBAG(IM)+(XBAR(JL)*N(J))
  103 XBAG(IM)=XBAG(IM)/NT
C     EN XBAG ESTA EL VECTOR DE MEDIAS GENERALES,EN STD EL DE DESV.ST.,
      K1=18
       WRITE(1'K1)(XBAG(J),J=1,M)
      WRITE(1,30)
   30 FORMAT(//'VECTOR DE MEDIAS GENERALES')
      WRITE(1,31)(XBAG(J),J=1,M)
   31 FORMAT(7F16.6)
      WRITE(1,7)
    7 FORMAT(//,'DETERMINANTE DE LA MATRIZ D',/)
      WRITE(1,8)DET
    8 FORMAT(E15.9)
C     CALCULO DE LOS DET. DE MATR.DE VAR-COVAR. DE CADA GRUPO
      DET2=DET
       K1=5
      JJ=1
       WRITE(1,9)
    9 FORMAT(//,'DETERMINANTES DE CADA GRUPO',/)
      DO 120 IK=1,K
      K5=1
      N5(1)=N(IK)
      NEL=N(IK)*M
      DO 121 I=1,NEL
      A(I)=A(JJ)
  121 JJ=JJ+1
       CALL DMATW(K5,M,N5,A,XBAR,D,W,CMEAN)
      WRITE(1'K1)(D(J),J=1,MM)
      CALL MINV1(D,M,DET,CMEAN,MI)
      WRITE(1,10)IK,DET
```

```
   10 FORMAT(12,5X,F15.6)
      DET1(IK)=DET
  120 CONTINUE
      K1=17
       WRITE(1'K1)(DET1(IK),IK=1,K),DET2
      CALL LINK(IGVA2)
      END

 FEATURES SUPPORTED
 ONE WORD INTEGERS
 IOCS

 CORE REQUIREMENTS FOR
 COMMON       2  VARIABLES   3288  PROGRAM   1070

  END OF COMPILATION

  // DUP

 *DELETE              IGVAC
CART ID 7777   DB ADDR  4E92   DB CNT   0045

  *STORE      WS  UA  IGVAC
CART ID 7777   DB ADDR  502A   DB CNT   0046

  // FOR
 *IOCS(CARD,DISK,TYPEWRITER,KEYBOARD)
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
*NAME IGVA2
       DIMENSION XBAR(100),XBAD(100),XBAG(10),XBAT(100),D(100),PROD(100),
     *DIST(100),SSUM(10),DESV(10),N(10),DET1(10)
      EQUIVALENCE(DIST(1),XBAT(1)),(PROD(1),XBAR(1))
      DEFINE FILE 1(30,200,U,K1),2(100,320,U,K2),3(1,100,U,K3)
      K5=1
       READ(3'K3)K,M,NT,(N(I),I=1,K)
C     ESTE PROGRAMA ES CONTINUACION DEL IGVA.
      KM=K*M
       MM=M*M
      NTT=NT*M
       K1=17
       READ(1'K1)(DET1(IK),IK=1,K),DET2
      SUM=0
       DO 130 IK=1,K
  130 SUM=SUM+(N(IK)-1)*ALOG(DET1(IK))
      CHIS=(NT-K)*ALOG(DET2)-SUM
      WRITE(1,11)CHIS
   11 FORMAT(//'CHIS=',F10.4)
C     CALCULOS PARA OBTENER LOS GRADOS DE LIBERTAD
      FA1S=0
      GA1S=0
       DO 140 IK=1,K
      FA1=1./(N(IK)-1)
      FA1S=FA1S+FA1
      GA1=1./((N(IK)-1)**2)
  140 GA1S=GA1S+GA1
      F1=0.5*(K-1)*M*(M+1)
      A1=(1.-(1./(NT-K)))*(2*M*M+3*M-1)/(6*(K-1)*(M+1))
      A2=(GA1S-(1./((NT-K)**2)))*((M-1)*(M+2))/(6*(K-1))
      DIF=A2-A1**2
      IF(DIF)141,142,142
  141 F2=(F1+2)/(A1**2-A2)
      BI=F2/(1-A1+(2/F2))
      F=(F2*CHIS)/(F1*(BI-CHIS))
      GO TO 143
  142 F2=(F1+2)/DIF
      BI=F1/(1.-A1-(F1/F2))
      F=CHIS/BI
  143 WRITE(1,12)F,F1,F2
   12 FORMAT('F',F10.4,2X,'CON',F6.0,'Y',F6.0,'GRADOS DE LIBERTAD')
      WRITE(1,13)
   13 FORMAT('UN VALOR SIGNIFICATIVO DE F INDICA QUE DEBE RECHAZARSE LA
```

```
     *HIPOTESIS DE IGUALDAD DE MATR.DE VAR-COVAR')
      K1=18
       READ(1'K1)(XBAG(J),J=1,M)
C     CALCULO DE D CUADR.PARA TODOS LOS GRUPOS Y PARA CADA PAR DE GRUPOS
      L=0
       K1=1
       READ(1'K1)(XBAR(J),J=1,KM)
      DO 150 I=1,K
      XN=N(I)
      DO 150 I1=1,M
      J=(I-1)*M+I1
  150 XBAD(J)=SQRT(XN)*(XBAR(J)-XBAG(I1))
      K1=15
       WRITE(1'K1)(XBAD(J),J=1,KM)
      CALL GMTRA(XBAD,XBAT,M,K)
      K1=19
       READ(1'K1)(D(J),J=1,MM)
      CALL GMPRD(XBAT,D,PROD,K,M,M)
      K1=15
       READ(1'K1)(XBAD(J),J=1,KM)
      CALL GMPRD(PROD,XBAD,DIST,K,M,K)
      SUM=0
      DO 160 I=1,K
      J=(I-1)*K+I
  160 SUM=SUM+DIST(J)
      WRITE(1,14)SUM
C     DIST(J) SON LOS ELEMENTOS DIAGONALES DE DIST
   14 FORMAT(//'D CUADR.DE MAHALANOBIS PARA TODOS LOS GRUPOS',E15.6)
      MK1=M*(K-1)
       WRITE(1,15)MK1
   15 FORMAT(//,'D CUADR. GENERAL TIENE DISTR.APROX. CHI CUADR. CON',I6,
     *1X,'GRADOS DE LIB.')
      WRITE(1,16)
   16 FORMAT(//,'D CUADRADO ENTRE GRUPOS',2X,'D CUADRADO',4X,'ESTAD.F',6
     *X,'GRADOS DE LIBERTAD',/)
      K1=1
       READ(1'K1)(XBAR(J),J=1,KM)
      K1=19
       READ(1'K1)(D(J),J=1,MM)
      K9=K-1
       DO 170 IK=1,K9
      K5=K-IK
       DO 170 IJK=1,K5
      IS=IK+IJK
      IKJ=IJK*M
      DO 171 I1=1,M
      J=IKJ+I1+((IK-1)*M)
      L=J-IKJ
  171 DESV(I1)=XBAR(L)-XBAR(J)
      DO 172 IJ=1,M
  172 SSUM(IJ)=0
      DC=0
       DO 173 IJ=1,M
      DO 173 J=1,M
      JK=(J-1)*M
      JL=IJ+JK
  173 SSUM(IJ)=SSUM(IJ)+DESV(J)*D(JL)
      DO 174 J=1,M
  174 DC=DC+SSUM(J)*DESV(J)
      RN=N(IK)
       RN1=N(IS)
      RI=M
      FEST=((RN*RN1*(RN1+RN-RI-1.))/(RI*(RN+RN1)*(RN+RN1-2.)))*DC
      L2=L(IK)+N(IS)-1-M
  170 WRITE(1,17)IK,IS,DC,FEST,M,L2
   17 FORMAT(10X,I2,1X,'Y',I2,5X,2E15.6,2X,I8,1X,'Y',I4)
      WRITE(1,18)
   18 FORMAT(//'UN VALOR NO SIGNIFIC. DE D CUADR.INDICA QUE LAS MEDIAS D
     *E LOS GRUPOS INVOLUCRADOS NO DIFIEREN SIGNIFICATIVAMENTE')
      CALL LINK(ADISC)
      END
```

```
FEATURES SUPPORTED
ONE WORD INTEGERS
IOCS

CORE REQUIREMENTS FOR IGVA2
COMMON       0  VARIABLES    990  PROGRAM   1454

   END OF COMPILATION

 // DUP

 *DELETE              IGVA2
CART ID 7777   DB ADDR  4D89   DB CNT   0063

  *STORE      WS  UA  IGVA2
CART ID 7777   DB ADDR  500D   DB CNT   0063

  // FOR
 *IOCS(CARD,DISK,TYPEWRITER,KEYBOARD)
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
*NAME ADISC
      DIMENSION XBAR(100),D(100),CMEAN(10),C(100),P(150),LG(150)
     *,N(10),IO(150),A(1305)
      DEFINE FILE 1(30,200,U,K1),2(100,320,U,K2),3(1,100,U,K3)
C     ESTE PROGRAMA CALCULA E IMPRIME LAS FS.DISCR. Y CLASIFICA LAS OBSE
C     RVACIONES SEGUN EL GRUPO CON MAYOR PROBAB.
C     LAS DIMENSIONES PARA ESTE PROGRAMA DEBEN SER- EL GUION INDICA 'MAYOR
C     O IGUAL QUE'-XBAR-M*K,D-M*M,CMEAN-M,C-M*M,P-NT,LG-NT,N-K,IO-NT,A-NT*M
      K3=1
      READ(3'K3)K,M,NT,(N(I),I=1,K)
      NTT=NT*M
      MM=M*M
      KM=K*M
      K2=1
      READ(2'K2)(A(J),J=1,NTT)
      READ(2'K2)(IO(J),J=1,NT)
      K1=1
      READ(1'K1)(XBAR(J),J=1,KM)
      K1=19
      READ(1'K1)(D(J),J=1,MM)
      CALL DISCR(K,M,N,A,XBAR,D,CMEAN,V,C,P,LG)
C     IMPRIMIR FUNCIONES DISCRIMINANTES
      WRITE(1,18)
   18 FORMAT(////)
      N1=1
      N2=M+1
      DO 200 I=1,K
      WRITE(1,21)I,(C(J),J=N1,N2)
   21 FORMAT('FUNCION DISCRIMINANTE',I3//6X,'CONSTANTE*COEFICIENTES',//E
     *14.6,3X,'*',3X,7E14.6//(22X,7E14.6))
      N1=N1+M+1
  200 N2=N2+M+1
C     IMPRIMIR EVALUACION DE LA CLASIFICACION PARA CADA OBSERVACION
      WRITE(1,22)
   22 FORMAT(///,'CLASIFICACION DE CADA OBSERVACION EN EL GRUPO CON MAYO
     *R PROBABILIDAD')
      N1=1
      N2=N(1)
      DO 210 I=1,K
      WRITE(1,23)I
   23 FORMAT(//,'GRUPO',I3/12X,'PROB.ASOC.CON',5X,'MAYOR'4X,'NUMERO'/'OB
     *SERVACION.',1X,'LA MAYOR F.DISC.',2X,'FUNC.NO.',2X,'DE OBSERV.')
      L=0
      DO 211 J=N1,N2
      L=L+1
  211 WRITE(1,24)L,P(J),LG(J),IO(J)
   24 FORMAT(17,9X,F8.5,6X,I6,5X,I5)
      IF(I-K)220,210,210
  220 N1=N1+N(I)
      N2=N2+N(I+1)
  210 CONTINUE
```

```
      WRITE(1,25)
   25 FORMAT(///'SI HAY OBSERVACIONES CLASIFICADAS EL UL GRUPO DISTINTO D
     *EL CRITICAL, CON PROG.ELEVADA,REPETIR TODO EL '/'PROCESO REORDENAND
     *O LAS OBSERVA.Y COMENZANDO POR EL PROGR.1.SI NO LAS HAY, PASAR AL
     *PROGRAMA SIGUIENTE')
      CALL LINK (DISCO)
      END

 FEATURES SUPPORTED
 ONE WORD INTEGERS
 IOCS

CORE REQUIREMENTS FOR ADISC
 COMMON       0  VARIABLES   3880  PROGRAM    640

   END OF COMPILATION

 // DUP

   *DELETE             ADISC
CART ID 7777   DB ADDR  4DD7   DB CNT   003C

   *STORE      WS  UA  ADISC
CART ID 7777   DB ADDR  5034   DB CNT   002A

  // FOR
 *IOCS(CARD,DISK,TYPEWRITER,KEYBOARD)
*ONE WORD INTEGERS
*LIST SOURCE PROGRAM
*NAME DISCO
      DIMENSION A(1305),A1(1305),N(5)
      DEFINE FILE 1(30,200,U,K1),2(100,320,U,K2),3(1,100,U,K3)
C     ESTE PROGRAMA REUBICA LOS DATOS EN LA MATRIZ A1,LUEGO CALCULA
C     MEDIANTE LA SUBRUTINA CORRA, LAS MATRICES DE SUMA DE PRODUCTOS CRU
C     ZADOS DE DESVIACIONES - RX, Y LA DE CORRELACIONES-R.GRABA RX EN
C     DISCO. IMPRIME R
C     LAS DIMENSIONES PARA LOS PROGRAMAS DISCO Y DISC2,DEBEN SER(EL GUION
C     INDICA'MAYOR O IGUAL QUE'.A-N*NT,A1-M*NT,N-K,XBAR-M,STD-M,RX-M*M,
C     R-M*M,B-N,D-M,T-M.
      MM=M*M
       K3=1
       READ(3'K3)K,M,NT,(N(I),I=1,K)
      NTT=NT*M
      K2=1
       READ(2'K2)(A(J),J=1,NTT)
      NN=0
       IM=0
       DO 60 I=1,K
      L=N(I)
       DO 50 J=1,L
      DO 50 KL=1,M
      KS=(KL-1)*N(I)+J+IM
      KW=(KL-1)*NT+J+NN
   50 A1(KW)=A(KS)
      NN=NN+N(I)
   60 IM=NN*M
      K2=1
       WRITE(2'K2)(A1(J),J=1,NTT)
      CALL LINK (DISC2)
      END

  FEATURES SUPPORTED
 ONE WORD INTEGERS
 IOCS

 CORE REQUIREMENTS FOR DISCO
 COMMON       0  VARIABLES   5266  PROGRAM    254

  END OF COMPILATION

 // DUP
```

```
      JL=IDR*(IK-1)+J
      IL=(JJ-1)*IDR+J
      DESV(J)=XD(JK)-C(JL)
  521 CHI(JJ)=CHI(JJ)+(DESV(J)*DD(IL))
      DO 522 J=1,IDR
  522 CHISQ(IK)=CHISQ(IK)+(CHI(J)*DESV(J))
      POT=-CHISQ(IK)/2.
      P1(IK)=COC(IK)*EXP(POT)
  520 P3=P3+P1(IK)
      DO 530 IK=1,K
      IKA=(IND-1)*K+IK
  530 PROB(IKA)=P1(IK)/P3
      M=1
      IKA=(IND-1)*K+1
      PMAY=PROB(IKA)
      DO 540 J=2,K
      IJ=IKA+J-1
      IF(PMAY-PROB(IJ))550,540,540
  550 M=J
        PMAY=PROB(IJ)
  540 CONTINUE
C      EN 'VALOR EN CADA FUNCION', ESCRIBE SOLO LOS VALORES EN LAS DOS PRI
C     MERAS QUE SON LAS UNICAS NECESARIAS PARA GRAFICAR CADA OBSERVACION
C     EN EL ESPACIO REDUCIDO
      J=NT+IND
  510 WRITE(1,51)IND,XD(IND),XD(J),M,PMAY,IO(IND)
   51 FORMAT(10X,I5,12X,2F7.2,10X,I2,11X,F7.5,9X,I3)
      STOP
       END

FEATURES SUPPORTED
ONE WORD INTEGERS
IOCS

CORE REQUIREMENTS FOR CLASI
COMMON      0  VARIABLES   3948  PROGRAM    760

  END OF COMPILATION

 // DUP

 *DELETE              CLASI
CART ID 7777   DB ADDR  4EE5   DB CNT   0034

  *STORE      WS  UA  CLASI
CART ID 7777   DB ADDR  5030   DB CNT   0034

  // XEQ IGVAC    1
*LOCALIGVAC,DMATW,MTRV1,MXSAL
```

APENDICE I-B

CLASIFICACION INICIAL

(Primera Iteración)

ANALISIS DISCRIMINANTE CORDOBA NUMERO DE GRUPOS 7
NUMERO DE VARIABLES 8 NUMERO DE OBSERVACIONES EN CADA GRUPO 19 14 11 15 6 6 73

MATRIZ DE MEDIAS,CADA COLUMNA INDICA UN GRUPO,CADA FILA INDICA UNA VARIABLE

MATRIZ 1 8 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 1	2	3	4	5	6
FILA 1	40.103645	70.161041	57.493331	78.167083	26.976158	35.447326
FILA 2	25.588298	0.186857	0.782363	3.707131	1.363999	0.106166
FILA 3	2.462523	1.460142	1.782271	2.795798	0.197833	3.085497
FILA 4	9.049257	4.215997	4.164362	2.106864	12.015331	91.760818
FILA 5	5.284153	22.432628	4.585907	5.910329	1.476332	0.101499
FILA 6	122.384643	198.844879	200.251739	223.141815	212.671600	148.431671
FILA 7	0.665789	0.367857	0.807272	0.715999	15.725000	1.679999
FILA 8	101.789474	100.785721	108.545456	104.533340	121.666671	106.666671

MATRIZ 1 8 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 7
FILA 1	28.667400
FILA 2	0.529738
FILA 3	1.804613
FILA 4	6.910631
FILA 5	0.328698
FILA 6	59.926216
FILA 7	0.753012
FILA 8	96.506851

VECTOR DE MEDIAS GENERALES

41.780662 4.170126 1.945892 9.512548 4.076508 119.446243 1.366804
100.847183

DETERMINANTE DE LA MATRIZ D

0.285788869E 15

DETERMINANTES DE CADA GRUPO

1 0.736977E 14
2 0.112821E 07
3 0.204994E 10
4 0.177071E 10
5 -0.106931E-07
6 -0.119727E-09
7 0.318322E 11

CHIS= 1756.5212
F 4.6895 CON 216.Y 2342.GRADOS DE LIBERTAD
UN VALOR SIGNIFICATIVO DE F INDICA QUE DEBE RECHAZARSE LA HIPOTESIS DE IGUALDAD DE MATR.DE VAR-COVAR

D CUADR.DE MAHALANOBIS PARA TODOS LOS GRUPOS 0.241534E 04

D CUADR. GENERAL TIENE DISTR.APROX. CHI CUADR. CON 48 GRADOS DE LIB.

D CUADRADO ENTRE GRUPOS D CUADRADO ESTAD.F. GRADOS DE LIBERTAD

1 Y 2 0.915255E 02 0.733052E 02 8 Y 21

1	Y	3	0.520769E 02	0.340127E 02	8 Y 21
1	Y	4	0.476187E 02	0.398801E 02	8 Y 25
1	Y	5	0.936659E 02	0.371405E 02	8 Y 16
1	Y	6	0.111261E 03	0.441177E 02	8 Y 16
1	Y	7	0.551103E 02	0.957783E 02	8 Y 83
2	Y	3	0.335018E 02	0.179453E 02	8 Y 16
2	Y	4	0.259852E 02	0.174239E 02	8 Y 20
2	Y	5	0.113046E 03	0.362690E 02	8 Y 11
2	Y	6	0.122894E 03	0.394285E 02	8 Y 11
2	Y	7	0.664782E 02	0.895770E 02	8 Y 78
3	Y	4	0.472251E 01	0.265357E 01	8 Y 17
3	Y	5	0.480604E 02	0.124391E 02	8 Y 8
3	Y	6	0.629877E 02	0.163927E 02	8 Y 8
3	Y	7	0.110525E 02	0.120796E 02	8 Y 75
4	Y	5	0.664327E 02	0.224772E 02	8 Y 12
4	Y	6	0.829488E 02	0.280653E 02	8 Y 12
4	Y	7	0.272124E 02	0.388809E 02	8 Y 79
5	Y	6	0.742381E 02	0.835178E 01	8 Y 3
5	Y	7	0.361676E 02	0.227868E 02	8 Y 70
6	Y	7	0.504405E 02	0.317793E 02	8 Y 70

UN VALOR NO SIGNIFIC. DE D CUADR.INDICA QUE LAS MEDIAS DE LOS GRUPOS INVOLUCRADOS NO DIFIEREN SIGNIFICAT IVAME NTE

FUNCION DISCRIMINANTE 1

CONSTANTE*COEFICIENTES

-0.433229E 02 * 0.265814E 00 0.206150E 01 -0.662984E 00 -0.647117E-01 0.351879E 00 -0.107873E-01 0.877439E-02
0.245496E 00

FUNCION DISCRIMINANTE 2

CONSTANTE*COEFICIENTES

-0.505876E 02 * 0.559967E 00 0.230271E-01 -0.178996E 01 -0.708759E-01 0.207094E 01 -0.987117E-02 -0.299155E 00
0.202529E 00

FUNCION DISCRIMINANTE 3

CONSTANTE*COEFICIENTES

-0.244527E 02 * 0.324142E 00 0.697069E-01 -0.104804E 01 -0.671283E-01 0.410536E 00 0.539177E-02 -0.657967E-01
0.265816E 00

FUNCION DISCRIMINANTE 4

CONSTANTE*COEFICIENTES

-0.321015E 02 * 0.500760E 00 0.379854E 00 -0.142781E 01 -0.909482E-01 0.609471E 00 0.705444E-02 -0.119377E 00
0.217584E 00

FUNCION DISCRIMINANTE 5

CONSTANTE*COEFICIENTES

-0.407618E 02 * 0.655833E-01 -0.168568E-01 -0.320598E 00 0.194651E-02 -0.175359E 00 -0.134466E-01 0.248108E 01
0.360993E 00

FUNCION DISCRIMINANTE 6

CONSTANTE*COEFICIENTES

-0.385301E 02 * 0.693186E-01 -0.165234E 00 -0.665519E 00 0.544971E 00 -0.152213E 00 0.962506E-03 0.279543E 00
0.244407E 00

FUNCION DISCRIMINANTE 7

CONSTANTE*COEFICIENTES

-0.144695E 02 * 0.664716E-01 -0.655132E-02 -0.173468E 00 -0.422799E-01 0.774819E-02 -0.985302E-02 0.165663E 00

0.290832E 00

CLASIFICACION DE CADA OBSERVACION EN EL GRUPO CON MAYOR PROBABILIDAD

GRUPO 1

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.98432	1	7
2	0.63849	1	26
3	1.00000	1	35
4	0.70906	1	62
5	0.99999	1	78
6	1.00000	1	85
7	1.00000	1	86
8	1.00000	1	87
9	0.92644	1	88
10	0.99998	1	91
11	1.00000	1	92
12	1.00000	1	116
13	1.00000	1	119
14	1.00000	1	125
15	1.00000	1	126
16	0.99999	1	127
17	1.00000	1	128
18	1.00000	1	129
19	1.00000	1	130

GRUPO 2

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.99993	2	39
2	1.00000	2	40
3	1.00000	2	41
4	1.00000	2	42
5	1.00000	2	43
6	0.99999	2	44
7	0.80117	2	45
8	0.95659	2	53
9	0.81795	4	54
10	0.92486	2	55
11	0.65073	2	142
12	1.00000	2	143
13	1.00000	2	144
14	0.85631	4	145

GRUPO 3

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.75682	7	19
2	0.84493	3	20
3	0.95368	3	22
4	0.69083	3	23
5	0.78322	4	56
6	0.95922	3	63
7	0.49951	3	64
8	0.83403	3	65
9	0.88490	3	66
10	0.88897	3	67
11	0.55957	3	68

GRUPO 4

PROB.ASOC.CON MAYOR NUMERO

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.91738	4	24
2	0.94569	4	25
3	0.83075	4	27
4	0.85849	4	28
5	0.98753	4	29
6	0.99084	4	36
7	0.87346	4	37
8	0.82095	4	38
9	0.78360	4	90
10	0.98911	4	93
11	0.70469	3	108
12	0.98959	4	109
13	0.75288	4	110
14	0.76912	4	111
15	0.91567	4	141

GRUPO 5

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	1.00000	5	5
2	0.58893	7	12
3	0.60261	5	57
4	0.99547	5	58
5	1.00000	5	60
6	1.00000	5	97

GRUPO 6

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	1.00000	6	9
2	1.00000	6	10
3	0.75577	7	11
4	0.96892	6	101
5	1.00000	6	102
6	0.99847	6	105

GRUPO 7

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.96622	7	2
2	0.92874	7	3
3	0.98702	7	4
4	0.95155	7	6
5	0.98058	7	8
6	0.70336	7	13
7	0.99829	7	14
8	0.99984	7	15
9	0.99944	7	16
10	0.99951	7	17
11	0.99958	7	18
12	0.97265	7	21
13	0.99986	7	30
14	0.99318	7	31
15	0.99973	7	32
16	0.99845	7	33
17	0.99308	7	34
18	0.99871	7	46
19	0.99851	7	47
20	0.99717	7	48
21	0.99321	7	49
22	0.99993	7	50
23	0.99854	7	51
24	0.99679	7	52
25	0.96781	7	59
26	0.77374	5	61
27	0.99349	7	69
28	0.99461	7	70
29	0.99469	7	71

30	0.98737	7	72
31	0.99047	7	73
32	0.99015	7	74
33	0.99749	7	75
34	0.99908	7	76
35	0.99064	7	77
36	0.97370	7	79
37	0.99901	7	80
38	0.99888	7	81
39	0.99929	7	82
40	0.99972	7	83
41	0.99990	7	84
42	0.70437	7	89
43	0.99918	7	94
44	0.99991	7	95
45	0.99299	7	96
46	0.99968	7	98
47	0.99642	7	99
48	0.99351	7	100
49	0.99804	7	103
50	0.96670	7	104
51	0.76913	7	106
52	0.85467	3	107
53	0.94423	3	112
54	0.64856	7	113
55	0.99299	7	114
56	0.99628	7	115
57	0.99365	7	117
58	0.99782	7	118
59	0.99919	7	120
60	0.99797	7	121
61	0.99989	7	122
62	0.95648	7	123
63	0.99653	7	124
64	0.99485	7	131
65	0.99653	7	132
66	0.99395	7	133
67	0.99977	7	134
68	0.92899	7	135
69	0.99983	7	136
70	0.99953	7	137
71	0.99416	7	138
72	0.99831	7	139
73	0.99967	7	140

SI HAY OBSERVACIONES CLASIFICADAS EN UN GRUPO DISTINTO DEL ORIGINAL,CON PROB.ELEVADA,REPETIR TODO EL PROCESO REORDENANDO LAS OBSERVA.Y COMENZANDO POR EL PROGR.1.SI NO LAS HAY, PASAR AL PROGRAMA SIGUIENTE

APENDICE I-C

CLASIFICACION FINAL

(Ultima Iteración)

ANALISIS DISCRIMINANTE CORDOBA NUMERO DE GRUPOS 7
NUMERO DE VARIABLES 8 NUMERO DE OBSERVACIONES EN CADA GRUPO 19 8 19 23 6 5 64

MATRIZ DE MEDIAS,CADA COLUMNA INDICA UN GRUPO,CADA FILA INDICA UNA VARIABLE

MATRIZ 1 8 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 1	2	3	4	5	6
FILA 1	40.103645	63.286987	55.238121	77.512359	24.449993	35.877107
FILA 2	25.568298	0.118625	1.117683	2.492998	1.498166	0.127499
FILA 3	2.462523	1.204624	5.323259	2.442065	0.114409	3.533197
FILA 4	0.049257	5.890374	5.210100	2.112127	10.832830	101.951000
FILA 5	5.284153	29.338241	2.536261	7.934515	1.401333	0.103099
FILA 6	122.384643	231.956183	160.260437	222.231964	101.144958	135.925987
FILA 7	0.665789	0.406249	0.739525	0.487390	16.233333	0.583999
FILA 8	101.789474	103.625015	113.315796	102.173019	127.000015	100.000015

MATRIZ 1 8 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 7
FILA 1	24.839973
FILA 2	0.401733
FILA 3	0.751839
FILA 4	6.694077
FILA 5	0.192218
FILA 6	47.628700
FILA 7	0.867342
FILA 8	92.955140

VECTOR DE MEDIAS GENERALES
41.730662 4.170127 1.945893 9.512559 4.076508 119.446273 1.360804
100.847213

DETERMINANTE DE LA MATRIZ D

0.758439729E 14

DETERMINANTES DE CADA GRUPO

1 0.736977E 14
2 0.225025E-01
3 0.676473E 12
4 0.198210E 10
5 -0.975219E-09
6 0.334359E-17
7 0.311449E 10

CHIS= 1756.5559
F 4.2070 CON 216.Y 1669.GRADOS DE LIBERTAD
UN VALOR SIGNIFICATIVO DE F INDICA QUE DEBE RECHAZARSE LA HIPOTESIS DE IGUALDAD DE MATR.DE VAR-COVAR

D CUADR.DE MAHALANOBIS PARA TODOS LOS GRUPOS 0.319004E 04

D CUADR. GENERAL TIENE DISTR.APROX. CHI CUADR. CON 48 GRADOS DE LIB.

D CUADRADO ENTRE GRUPOS	D CUADRADO	ESTAD.F	GRADOS DE LIBERTAD
1 Y 2	0.132122E 03	0.669420E 02	8 Y 18
1 Y 3	0.500201E 02	0.478491E 02	8 Y 20
1 Y 4	0.538123E 02	0.577401E 02	8 Y 33
1 Y 5	0.122242E 03	0.484719E 02	8 Y 10
1 Y 6	0.138720E 03	0.468015E 02	8 Y 15
1 Y 7	0.690262E 02	0.100427E 03	8 Y 74
2 Y 3	0.944356E 02	0.478473E 02	8 Y 18
2 Y 4	0.575939E 02	0.326184E 02	8 Y 22
2 Y 5	0.229743E 03	0.388135E 02	8 Y 5
2 Y 6	0.208559E 03	0.291663E 02	8 Y 4
2 Y 7	0.132106E 03	0.105685E 03	8 Y 63
3 Y 4	0.156235E 02	0.167639E 02	8 Y 33
3 Y 5	0.726541E 02	0.288059E 02	8 Y 10
3 Y 6	0.963849E 02	0.325702E 02	8 Y 15
3 Y 7	0.156816E 02	0.262029E 02	8 Y 74
4 Y 5	0.121221E 03	0.534118E 02	8 Y 20
4 Y 6	0.120388E 03	0.487684E 02	8 Y 19
4 Y 7	0.452444E 02	0.878090E 02	8 Y 78
5 Y 6	0.131981E 03	0.980554E 01	8 Y 2
5 Y 7	0.511819E 02	0.314833E 02	8 Y 61
6 Y 7	0.681578E 02	0.353837E 02	8 Y 69

UN VALOR NO SIGNIFIC. DE D CUADR.INDICA QUE LAS MEDIAS DE LOS GRUPOS INVOLUCRADOS NO DIFIEREN SIGNIFICA TIVA MENTE

FUNCION DISCRIMINANTE 1

CONSTANTE*COEFICIENTES

-0.498114E 02 * 0.327870E 00 0.209681E 01 0.432627E 00 -0.171055E 00 0.675187E 00 0.662563E-02 -0.154575E 00
0.295279E 00

FUNCION DISCRIMINANTE 2

CONSTANTE*COEFICIENTES

-0.386286E 02 * 0.514747E 00 0.613297E-01 0.302626E 00 -0.206049E 00 0.374127E 01 0.319903E-01 -0.131248E 01
0.278391E 00

FUNCION DISCRIMINANTE 3

CONSTANTE*COEFICIENTES

-0.331380E 02 * 0.302922E 00 0.147786E 00 0.103067E 01 -0.224098E 00 0.470222E 00 0.179135E-01 -0.586446E-01
0.362188E 00

FUNCION DISCRIMINANTE 4

CONSTANTE*COEFICIENTES

-0.452542E 02 * 0.596285E 00 0.379957E 00 0.165910E 00 -0.255318E 00 0.113815E 01 0.361167E-01 -0.721432E 00
0.257661E 00

FUNCION DISCRIMINANTE 5

CONSTANTE*COEFICIENTES

-0.576163E 02 * -0.365049E-01 -0.537328E-01 0.423933E 00 -0.133950E 00 -0.280884E 00 -0.363905E-01 0.363482E 01
0.519339E 00

FUNCION DISCRIMINANTE 6

CONSTANTE*COEFICIENTES

-0.408661E 02 * 0.121048E-01 -0.194963E 00 -0.328977E 00 0.685182E 00 0.246254E-01 -0.736577E-02 -0.123821E 00

0.207433E 00

FUNCION DISCRIMINANTE 7

CONSTANTE*COEFICIENTES

-0.161067E 02 * 0.768090E-01 -0.459639E-02 0.206205E 00 -0.771742E-01 0.562961E-01 -0.863759E-02 0.413431E 00

0.330473E 00

CLASIFICACION DE CADA OBSERVACION EN EL GRUPO CON MAYOR PROBABILIDAD

GRUPO 1

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.95889	1	7
2	0.88756	1	26
3	1.00000	1	35
4	0.98071	1	62
5	0.99998	1	78
6	1.00000	1	85
7	1.00000	1	86
8	1.00000	1	87
9	0.62835	3	88
10	0.99999	1	91
11	1.00000	1	92
12	1.00000	1	116
13	1.00000	1	119
14	1.00000	1	125
15	1.00000	1	126
16	0.99999	1	127
17	1.00000	1	128
18	1.00000	1	129
19	1.00000	1	130

GRUPO 2

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.77941	2	39
2	1.00000	2	40
3	1.00000	2	41
4	1.00000	2	42
5	1.00000	2	43
6	0.99989	2	44
7	1.00000	2	143
8	1.00000	2	144

GRUPO 3

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.80675	3	13
2	0.82874	3	19
3	0.93229	3	20
4	0.95866	3	22
5	0.90085	3	63
6	0.61146	4	65
7	0.74044	3	66
8	0.51000	3	67
9	0.86005	3	71
10	0.99912	3	73
11	0.99879	3	74
12	0.99993	3	76
13	0.99999	3	77
14	0.68871	3	79

15	0.98814	5	79
16	0.99954	5	106
17	0.99875	5	107
18	0.74092	5	108
19	0.99838	5	112

GRUPO 4

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.80163	4	23
2	0.99992	4	24
3	0.99995	4	25
4	0.99967	4	27
5	0.78513	4	28
6	0.99916	4	29
7	0.99998	4	36
8	0.99926	4	37
9	0.99526	4	38
10	0.99999	4	45
11	0.99126	4	53
12	0.99993	4	54
13	0.99999	4	55
14	0.99999	4	56
15	0.92772	4	68
16	0.98131	4	90
17	0.99965	4	93
18	0.99999	4	109
19	0.98740	4	110
20	0.99779	4	111
21	0.99977	4	141
22	0.99996	4	142
23	0.99999	4	145

GRUPO 5

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	1.00000	5	5
2	0.92702	7	57
3	0.99869	5	58
4	1.00000	5	60
5	0.99879	5	61
6	1.00000	5	97

GRUPO 6

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	1.00000	6	9
2	1.00000	6	10
3	0.56009	7	101
4	1.00000	6	102
5	0.98250	6	105

GRUPO 7

OBSERVACION	PROB.ASOC.CON LA MAYOR F.DISC.	MAYOR FUNC.NO.	NUMERO DE OBSERV.
1	0.98832	7	2
2	0.98541	7	3
3	0.90873	7	4
4	0.90878	7	6
5	0.80913	7	8
6	0.99990	7	11
7	0.91036	7	12
8	0.99955	7	14
9	0.99999	7	15
10	0.99995	7	16
11	0.99996	7	17
12	0.99995	7	18
13	0.96176	7	91

14	0.99857	7	81
15	0.99869	7	81
16	0.99999	7	82
17	0.99998	7	83
18	0.99963	7	84
19	0.99982	7	86
20	0.99997	7	87
21	0.99997	7	88
22	0.99990	7	89
23	0.99969	7	90
24	0.99997	7	91
25	0.99996	7	92
26	0.97326	7	99
27	0.98722	7	94
28	0.99989	7	99
29	0.99967	7	70
30	0.97430	7	72
31	0.99984	3	75
32	0.99997	7	80
33	0.99924	7	81
34	0.99968	7	82
35	0.99996	7	83
36	0.99999	7	84
37	0.99997	7	94
38	0.99993	7	95
39	0.99990	7	96
40	0.99998	7	98
41	0.99929	7	99
42	0.99995	7	100
43	0.99950	7	103
44	0.99985	7	104
45	0.99171	7	113
46	0.99810	7	114
47	0.99991	3	115
48	0.99994	7	117
49	0.99921	7	118
50	0.99997	7	120
51	0.99989	7	121
52	0.99999	7	122
53	0.97816	7	123
54	0.99990	7	124
55	0.99214	7	131
56	0.98275	7	132
57	0.99437	7	133
58	0.99787	3	134
59	0.99751	7	135
60	0.99998	7	136
61	0.99999	7	137
62	0.99949	7	138
63	0.99987	7	139
64	0.99998	7	140

SI HAY OBSERVACIONES CLASIFICADAS EN UN GRUPO DISTINTO DEL ORIGINAL, CON PROB. ELEVADA, REPETIR TODO EL PROCESO REORDENANDO LAS OBSERVA. Y COMENZANDO POR EL PROGR. 1. SI NO LAS HAY, PASAR AL PROGRAMA SIGUIENTE

MATRIZ DE CORRELACIONES ENTRE VARIABLES ORIGINALES

MATRIZ 3 6 FILAS 6 COLUMNAS MODO DE ALM 1

COLUMNA	1	2	3	4	5	6
FILA 1	1.000000	0.000256	-0.421596	-0.037587	0.457080	0.442672
FILA 2	0.000256	1.000000	0.074857	-0.013754	0.077891	0.073903
FILA 3	-0.421596	0.074857	1.000000	0.202765	-0.000076	0.110733
FILA 4	-0.037587	-0.013754	0.202765	1.000000	-0.105141	0.062586
FILA 5	0.457080	0.077891	-0.000076	-0.105141	1.000000	0.442403
FILA 6	0.442672	0.073903	0.110733	0.062586	0.442403	1.000000

FILA 7	-0.142458	-0.060730	-0.141522	0.049011	-0.035857	0.200931
FILA 8	0.246885	0.054899	0.126572	0.165074	0.082777	0.340101

MATRIZ 3 8 FILAS 8 COLUMNAS MODO DE ALM 1

COLUMNA	7	8
FILA 1	-0.142458	0.246885
FILA 2	-0.060730	0.054899
FILA 3	-0.141522	0.126572
FILA 4	0.049011	0.165074
FILA 5	-0.035857	0.082777
FILA 6	0.200931	0.340101
FILA 7	1.000000	0.196175
FILA 8	0.196175	1.000000

	VALOR PROPIO	PORC.DE LA TRAZA
1	11.226	48.11
2	5.655	24.23
3	2.786	11.94
4	1.987	8.51
5	1.300	5.57
6	0.378	1.61
7	0.000	0.00
8	-0.000	-0.00

VECTORES PROPIOS
CADA COLUMNA ES UN VECTOR PROPIO

MATRIZ 4 8 FILAS 8 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3	4	5	6
FILA 1	0.175615	0.000201	-0.057107	-0.097778	0.102239	-0.001703
FILA 2	0.352594	-0.028226	0.008554	0.054068	0.032307	-0.030948
FILA 3	0.050058	-0.056351	-0.328010	-0.031210	0.101446	0.974440
FILA 4	-0.082250	0.020626	0.313248	0.019701	0.144089	-0.024934
FILA 5	0.711125	0.348665	0.315071	0.585022	-0.305978	0.114251
FILA 6	0.216301	0.004789	-0.001064	-0.015398	0.014021	0.000590
FILA 7	-0.623303	-0.121217	-0.813961	0.814912	0.800204	-0.147792
FILA 8	-0.030097	0.003051	-0.057295	0.050837	0.051500	0.070174

MATRIZ 4 8 FILAS 8 COLUMNAS MODO DE ALM 0

COLUMNA	7	8
FILA 1	-0.304004	0.010048
FILA 2	-0.057814	0.002218
FILA 3	0.250929	0.044134
FILA 4	-0.056524	0.011685
FILA 5	-0.269965	0.074485
FILA 6	0.137760	-0.004872
FILA 7	-0.761662	0.404963
FILA 8	-0.204235	-0.100591

F 54.14437 CON 48.00000 Y 345.71679 GRADOS DE LIB. LAMDA= 0.00034

ESTE TEST ES EQUIVALENTE AL EFECTUADO EN EL PROGRAMA 1 CON D CUADRADO

VECTORES PROPIOS ESTANDARIZADOS

MATRIZ 5 8 FILAS 8 COLUMNAS MODO DE ALM 0

	COLUMNA 1	2	3	4	5	6
FILA 1	25.018512	0.895200	-8.134052	-13.913826	25.032036	-13.049518
FILA 2	10.665744	-30.194137	2.894685	2.283037	1.367962	-1.688927
FILA 3	1.534161	-1.542188	-0.822675	-0.854406	2.777420	20.[illegible]037
FILA 4	-11.562195	2.884476	43.895427	2.755117	20.229835	-3.403864
FILA 5	24.543096	13.957666	10.874462	19.591361	-13.066927	5.943324
FILA 6	10.670658	4.818409	-1.088445	-15.746610	14.338523	0.399201
FILA 7	-17.859294	-3.473190	-23.322139	23.349388	25.221803	-4.234653
FILA 8	-0.033050	0.611708	-11.484941	10.190399	10.324646	14.966652

MATRIZ 5 8 FILAS 8 COLUMNAS MODO DE ALM 0

	COLUMNA 7	8
FILA 1	-56.169033	2.283605
FILA 2	-2.441192	0.093696
FILA 3	7.031409	23.101754
FILA 4	-7.904036	1.654079
FILA 5	-14.007959	2.570812
FILA 6	140.876251	-4.982267
FILA 7	-22.663813	14.182010
FILA 8	-40.939361	-38.284668

3

MATRIZ DE MEDIAS DE CADA GRUPO EN EL ESPACIO REDUCIDO
CADA FILA ES UNA VARIABLE,CADA COL. UN GRUPO

MATRIZ 6 3 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 1	2	3	4	5	6
FILA 1	15.182202	31.997650	10.414206	20.107105	-0.052285	-3.217076
FILA 2	-20.014441	11.861309	1.008333	2.127930	-1.130000	2.945106
FILA 3	-5.427000	0.532744	-0.818334	-8.462589	-18.200079	21.794212

MATRIZ 6 3 FILAS 7 COLUMNAS MODO DE ALM 0

	COLUMNA 7
FILA 1	1.635181
FILA 2	0.348535
FILA 3	-5.587356

MATRIZ DE DISPERSION DEL GRUPO 1 EN EL ESPACIO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

	COLUMNA 1	2	3
FILA 1	23.808708	-7.940645	-1.066709
FILA 2	-7.940642	54.709583	-12.535562
FILA 3	-1.066861	-12.532594	14.042962

MATRIZ DE DISPERSION DEL GRUPO 2 EN EL ESPAC IO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	24.824432	12.866945	11.021089
FILA 2	12.866939	7.245349	6.789002
FILA 3	11.021085	6.789003	7.524467

MATRIZ DE DISPERSION DEL GRUPO 3 EN EL ESPAC IO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	6.752897	1.230798	-0.035439
FILA 2	1.230798	3.882971	-0.246611
FILA 3	-0.035440	-0.246611	6.706613

MATRIZ DE DISPERSION DEL GRUPO 4 EN EL ESPAC IO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	7.161455	2.898064	2.719694
FILA 2	2.898066	18.491539	2.938091
FILA 3	2.719694	2.938091	3.479087

MATRIZ DE DISPERSION DEL GRUPO 5 EN EL ESPAC IO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	8.694095	1.892338	15.188776
FILA 2	1.892337	2.473163	7.386033
FILA 3	15.188766	7.386032	43.116111

MATRIZ DE DISPERSION DEL GRUPO 6 EN EL ESPAC IO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	[illegible]	-3.3[illegible]	-34.5[illegible]

FILA 2	-3.382501	2.804062	24.899662
FILA 3	-34.581778	24.899665	251.938751

MATRIZ DE DISPERSION DEL GRUPO 7 EN EL ESPACIO REDUCIDO

MATRIZ 7 3 FILAS 3 COLUMNAS MODO DE ALM 0

COLUMNA	1	2	3
FILA 1	3.984474	-0.457430	-0.865969
FILA 2	-0.457430	1.572372	-0.290248
FILA 3	-0.865969	-0.290248	4.452538

OBSERV.	VALOR EN FUNC.1	FUNC.2	GRUPO CON MAYOR PROB	PROBAB.	NUM. DE OBSERV.
1	7.00	-13.06	1	1.00000	7
2	19.32	-12.79	1	0.99995	26
3	22.37	-16.76	1	0.99999	35
4	15.12	-8.59	1	0.99999	62
5	8.66	-16.74	1	1.00000	76
6	12.80	-31.28	1	1.00000	85
7	14.74	-30.52	1	1.00000	86
8	13.39	-19.36	1	1.00000	87
9	14.37	-11.30	1	0.98859	88
10	12.62	-15.13	1	0.99976	91
11	16.78	-32.02	1	1.00000	92
12	6.44	-20.40	1	1.00000	116
13	12.82	-24.79	1	1.00000	119
14	17.30	-18.45	1	1.00000	125
15	20.11	-32.76	1	1.00000	126
16	18.99	-19.19	1	0.99999	127
17	12.75	-16.66	1	1.00000	128
18	17.80	-28.66	1	1.00000	129
19	24.03	-22.84	1	1.00000	130
20	23.51	7.53	2	0.74687	39
21	38.40	14.96	2	1.00000	40
22	35.38	14.84	2	1.00000	41
23	34.81	13.70	2	1.00000	42
24	34.40	11.69	2	0.99999	43
25	28.13	9.37	2	0.99952	44
26	33.42	12.87	2	1.00000	143
27	26.10	10.06	2	0.99996	144
28	5.81	2.38	7	0.49962	13
29	6.12	1.19	3	0.81209	19
30	11.31	2.06	3	0.99626	20
31	12.44	3.40	3	0.99961	22
32	12.45	-3.34	3	0.80450	63
33	14.40	3.78	3	0.95867	65
34	8.85	1.61	3	0.98102	66
35	12.67	2.07	3	0.97239	67
36	6.04	1.16	3	0.54379	71
37	7.09	-0.53	3	0.95819	73
38	8.36	-0.10	3	0.99499	74
39	10.30	-0.93	3	0.99582	76
40	11.87	1.31	3	0.98152	77
41	9.11	-3.40	3	0.89664	79
42	11.72	1.27	3	0.98732	89
43	12.17	0.69	3	0.98977	100
44	11.88	2.02	3	0.99178	107
45	13.56	3.12	3	0.99571	108
46	11.52	1.35	3	0.99021	112
47	16.52	4.03	4	0.84123	23
48	15.89	-0.14	4	0.61115	24
49	20.83	1.19	4	0.99874	25
50	16.58	1.63	4	0.99750	27

52	20.61	-7.95	4	0.87966	29
53	22.99	-2.98	4	0.99662	36
54	99.60	4.14	4	0.97429	37
55	18.25	3.46	4	0.98091	38
56	24.49	6.47	4	0.74378	45
57	24.79	7.03	4	0.99828	53
58	21.30	5.57	4	0.84543	54
59	24.26	4.83	4	0.99178	55
60	23.47	6.03	4	0.99068	56
61	15.85	3.45	4	0.70566	68
62	20.66	-3.39	4	0.97945	90
63	20.98	-7.00	4	0.91339	93
64	21.59	4.14	4	0.99871	109
65	17.12	3.02	4	0.92537	110
66	19.49	4.24	4	0.99421	111
67	19.69	4.30	4	0.96824	141
68	22.32	4.99	4	0.99054	142
69	23.01	5.60	4	0.98604	145
70	-8.21	-4.24	5	1.00000	5
71	-2.97	-0.10	5	0.93114	57
72	-5.62	-1.05	5	0.99999	58
73	-7.21	-0.15	5	1.00000	60
74	-4.83	-0.27	5	0.99999	61
75	-10.29	-0.69	5	1.00000	97
76	-6.39	5.85	6	1.00000	9
77	-2.01	1.85	6	1.00000	10
78	-0.96	2.35	6	0.99998	101
79	-5.21	2.82	6	1.00000	102
80	-1.48	1.34	6	1.00000	103
81	1.39	0.83	7	0.97585	2
82	0.09	-2.23	7	0.99940	3
83	0.73	0.52	7	0.98987	4
84	3.67	-0.89	7	0.99829	6
85	3.56	1.30	7	0.99814	8
86	-1.29	1.47	7	0.89217	11
87	3.29	1.22	7	0.99109	12
88	1.32	1.20	7	0.99534	14
89	-0.48	0.40	7	0.98102	15
90	0.39	0.57	7	0.98658	16
91	-0.77	0.75	7	0.98243	17
92	0.03	0.49	7	0.98854	18
93	2.92	0.86	7	0.99284	21
94	-0.24	0.64	7	0.98540	30
95	2.28	0.94	7	0.99767	31
96	-0.21	0.53	7	0.98088	32
97	0.45	0.59	7	0.95602	33
98	1.57	0.73	7	0.98984	34
99	2.05	1.00	7	0.99844	46
100	0.51	0.56	7	0.97942	47
101	1.30	0.47	7	0.98745	48
102	0.85	0.83	7	0.98028	49
103	-1.08	0.31	7	0.98252	50
104	0.52	-0.08	7	0.98567	51
105	-0.58	0.66	7	0.97898	52
106	2.57	-0.07	7	0.97240	59
107	7.08	1.91	3	0.70604	64
108	2.90	0.44	7	0.98501	69
109	2.58	0.12	7	0.99758	70
110	3.94	0.26	7	0.99432	72
111	0.26	-0.32	3	0.57339	75
112	0.73	0.39	7	0.98605	80
113	0.69	1.04	7	0.98116	81
114	0.04	0.46	7	0.97848	82
115	-0.28	0.40	7	0.98078	83
116	-0.76	0.24	7	0.98147	84
117	-0.65	0.84	7	0.94666	94
118	-1.13	0.39	7	0.96871	95
119	1.61	0.58	7	0.98125	96
120	-0.21	0.42	7	0.98575	98
121	1.11	0.80	7	0.98763	99
122	1.19	0.81	7	0.99006	100
123					

124	2.23	[illegible]	7	0.99379	[illegible]
125	4.03	0.47	7	0.97780	113
126	3.85	-8.01	1	0.99992	114
127	3.75	0.22	7	0.98473	115
128	-0.42	-0.05	7	0.84845	117
129	1.20	-2.34	7	0.99300	118
130	0.70	0.35	7	0.98778	120
131	1.16	0.42	7	0.99963	121
132	-0.81	0.35	7	0.98105	122
133	4.29	1.29	7	0.98706	123
134	1.26	0.52	7	0.99130	124
135	1.95	1.29	7	0.99766	131
136	3.11	0.62	7	0.99649	132
137	3.30	0.40	7	0.99815	133
138	6.49	0.01	7	0.53368	134
139	3.52	0.68	7	0.99833	135
140	-0.28	0.45	7	0.98395	136
141	0.31	0.35	7	0.97296	137
142	1.64	0.86	7	0.99570	138
143	[illegible]	0.22	7	0.98612	139
144	-0.23	0.22	7	0.98288	140

APENDICE II

UN METODO PARA DIBUJAR LAS ELIPSES DE EQUIPROBABILIDAD EN UNA DISTRIBUCIÓN NORMAL BIVARIADA

Transcribimos este método(*) porque resulta el indicado para dibujar las elipses de probabilidad de cada grupo en el espacio bidimensional. Bajo el supuesto de normalidad multivariada de cada población original, la distribución en dos dimensiones será normal bivariada (las variables canónicas son una combinación lineal de las variables originales), luego, pueden dibujarse elipses que correspondan a una probabilidad de 1-α con centro en la media del grupo en dicho espacio.

Las elipses se refieren a ejes que pasan por las medias de los grupos.

Sus ecuaciones son:

$$\frac{x_1^2}{\sigma_1^2} - 2\,\frac{\rho x_1 x_2}{\sigma_1 \sigma_2} + \frac{x_2^2}{\sigma_2^2} = C^2(1-\rho^2)$$

ρ es el coeficiente de correlación.

C^2 es el valor de 1-α en la distribución χ^2 con dos grados de libertad (ya que esa es la distribución de la forma cuadrática).

Deben trazarse seis pares de tangentes a la elipse, marcando los puntos para luego dibujarla.

1) Se trazan las líneas de regresión

$$\text{i)}\quad x_2 = \left(\rho\frac{\sigma_2}{\sigma_1}\right)x_1 \quad \text{y} \quad \text{ii)}\quad x_1 = \left(\rho\,\frac{\sigma_1}{\sigma_2}\right)x_2$$

las tangentes en sus extremos son paralelas a los ejes x_2 y x_1 respectivamente; los puntos de contacto con la elipse son

$$(\pm C\sigma_1, \pm C\rho\sigma_2) \quad \text{(para i)}$$

$$(\pm C\rho\sigma_1, \pm C\sigma_2) \quad \text{(para ii)}$$

(*) El método está descripto en M. y R. Healy (1972)

2) Las intersecciones con el eje x_1 son $(\pm C\xi\sigma_1, 0)$ (siendo $\xi = \sqrt{1-\rho^2}$) y las tangentes pasan por los puntos

$$\left(0, \pm C\xi \frac{\sigma_2}{\rho}\right)$$

Las intersecciones con el eje x_2 son $(0, \pm C\xi\sigma_2)$ y y las tangentes pasan por los puntos

$$\left(\pm C\xi \frac{\sigma_1}{\rho}, 0\right)$$

3) Haciendo $a = \sigma_2^2 - \sigma_1^2$, $B = 2\rho\sigma_1\sigma_2$ los ejes son $x_2 = kx_1$, $x_1 = -kx_2$ con $k = (A + \sqrt{A^2+B^2})/B$

Las longitudes de los semiejes son: $C\sqrt{\frac{1}{2}(\sigma_1^2+\sigma_2^2 \pm \sqrt{A^2+B^2})}$ y las tangentes en sus extremos son perpendicualres a los ejes.

APENDICE III-A

LISTADO DE PEDANIAS DE LA PROVINCIA DE CORDOBA, POR DEPARTAMENTO: (Corresponde a Mapa I)

DEPARTAMENTOS	PEDANIAS
Capital	1. Capital
Calamuchita	2. Los Reartes
	3. Los Molinos
	4. Santa Rosa
	5. Mansalvo
	6. Cañada de Alvarez
	7. Los Condores
	8. Río de los Sauces
Colón	9. Cañas
	10. Constitución
	11. San Vicente
	12. Río Ceballos
	13. Calera Norte
Cruz del Eje	14. Cruz del Eje
	15. Pichanas
	16. Higueras
	17. San Marcos
	18. Candelaria
General Roca	19. Sarmiento
	20. Necochea
	21. El cuero
	22. Jagueles
	23. Italo
General San Martin	24. Algodón
	25. Las Mojarras
	26. Yucat
	27. Villa María
	28. Villa Nueva
	29. Chazón
Ischilín	30. Quilino
	31. Toyos
	32. Copacabana
	33. Parroquia
	34. Manzanas
Juarez Celman	35. Carnerillo
	36. Chucul
	37. Reducción
	38. Carlota
Marcos Juarez	39. Colonia
	40. Espinillo
	41. Saladillos
	42. Cruz Alta
	43. Liniers
	44. Caldera
	45. Las Tunas

Minas	46. Guasapampa
	47. Argentina
	48. Ciénaga del Coro
	49. San Carlos
Pocho	50. Chancaní y República
	51. Parroquia
	52. Salsacate
Presidente Roque S.Peña	53. La Amarga
	54. San Martín
	55. La Paz
	56. Independencia
Punilla	57. Dolores
	58. San Antonio
	59. Rosario
	60. San Roque
	61. Santiago
Rio Cuarto	62. Las Peñas
	63. Tegua
	64. San Bartoloméé
	65. Rio Cuarto
	66. Achiras
	67. 3 de Febrero
	68. La Cautiva
Río Primero	69. Chalacea
	70. Timón Cruz
	71. Castaños
	72. Esquina
	73. Tala
	74. Santa Rosa
	75. Suburbios
	76. Quebracho
	77. Yegua Muerta
	78. Remedios
	79. Villa Norte
Río Seco	80. Higuerillas
	81. Villa de María
	82. Estancia
	83. Candelaria Sur
	84. Candelaria Norte
Rio Segundo	85. Pilar
	86. San José´
	87. Suburbios
	88. Villa del Rosario
	89. Oratorio de Peralta
	90. Arroyo de Alvarez
	91. Matorrales
	92. Impira
	93. Calchin

San Alberto	94. Ambul
	95. El Carmen
	96. Panaholma
	97. Tránsito
	98. Toscas
	99. San Pedro
	100. Nono
San Javier	101. Dolores
	102. Las Rosas
	103. San Javier
	104. Luyaba
	105. La Paz (talas)
San Justo	106. San Francisco
	107. Arroyito
	108. Concepción
	109. Libertad
	110. Sacanta
	111. Juarez Celman
Santa María	112. Calera Sur
	113. Lagunilla
	114. Caseros
	115. Alta Gracia
	116. San Antonio
	117. Potrero de Garay
	118. San Isidro
	119. Cosme
Sobremonte	120. Aguada del Monte
	121. Cerrillos
	122. Chuñahuasi
	123. San Francisco
	124. Caminiaga
Tercero Arriba	125. Salto
	126. Capilla de Rodriguez
	127. Los zorros
	128. Pampayasta Norte
	129. Pampayasta Sur
	130. Punta del Agua
Totoral	131. Totoral
	132. Macha
	133. Candelaria
	134. Sinsacate
	135. Río Pinto
Tulumba	136. San Pedro
	137. Intihasi
	138. Parroquia
	139. San José (Dormida)
	140. Mercedes

APENDICE III-B

Clasificación inicial (corresponde a Mapa II)

Zona 1 (19 pedanías)

7. Los Cóndores
26. Yucat
35. Carrerillo
62. Las Peñas
78. Remedios
85. Pilar
86. San José
87. Suburbios
88. V.Del Rosario
91. Matorrales
92. Impira
116. S. Antonio
119. Cosme
125. Salto
126. Cap. de Rodriguez
127. Los Zorros
128. Papayasta NOrte
129. Pampayasta Sur
130. Punta del Agua

Zona 2 (14 pedanías)

39. Colonias
40. Espinillos
41. Saladillo
42. Cruz Alta
43. Liniers
44. Calderas
45. Las Tunas
53. La Amarga
54. San Martín
55. La Paz
142. Ballesteros
143. B.Ville
144. Ascasubi
145. Loboy.

Zona 3 (11 pedanías)

119. Sarmiento
20. Necochea
22. Jaguales
23. Italo
56. Independencia
63. Tegua
64. S.Bartolomé
65. Río Cuarto
66. Achira
67. 3 de Febrero
68. Cautiva

Zona 4 (15 pedanías)

24. Algodón
25. Las Mojarras
27. Villa Maria
28. Villa Nueva
29. Chazón
36. Chucul
37. Reducción
38. Carlota
90. Arroyo de Alvarez
93. Calchín
108. Concepción
109. Libertad
110. Sacanta
111. Juarez Celman
141. Litin

Zona 5 (6 pedanías)

5. Monsalvo
12. Río Ceballos
57. Dolores
58. San Antonio
60. San Roque
97. Tránsito

Zona 6 (6 pedanías)

9. Cañas
10. Constitución
11. San Vicente
101. Dolores
102. Las Rosas
105' La Paz

Zona 7 (73 pedanías)

2. Los Reartes
3. Los Molinos
4. Santa Rosa
6. Cda. de Alvarez
8. Rio de los Sauces
13. Calera N.
14. Cruz del Eje
15. Pichanas
16. Higueras
17. San Marcos
18. Candelaria
21. El Cuero
30. Quilino
31. Tohos
32. Copacabana
33. Parroqui
34. Manzanas
46. Guasapampa
47. Argentina
48. Ciénaga del Coro
49. San Carlos
50. Chancaní y República
51. Parroquia
52. Salsacate
59. Rosario
61. Santiago
69. Chalacea
70. Timón Cruz
71. Castaños
72. Esquina
73. Tala
74. Santa Rosa
75. Suburbios
76. Quebracho
77. Yegua Muerta
79. Villa Norte
80. Higuerillas
81. V.de María
82. Estancia
83. Candelaría
84. Candelaria N.
89. Oratorio de Peralta
94. Ambul
95. El Carmen
96. Panaholma
98. Toscas
99. San Pedro
100. Nono
103. San Javier
104. Luyaba
106. San Francisco
107. Arroyito
112. Calera Sur
113. Lagunilla
114. Caseros
115. Alta Gracia
117. Potrero de Garay
118. San Isidro
120. Aguada del Monto
121. Cerrillos
122. Chuñauasi
123. San Francisco
124. Caminiaga
131. Totoral
132. Macha
133. Candelaria
134. Sinsacate
135. Río Pinto
136. San Pedro
137. Intihasi
138. Parroquia
139. San José
140. Mercedes

APENDICE III - C

CLASIFICACION FINAL (Corresponde a Mapa III)

Zona 1 (19 pedanías)

7. Los Cóndores
26. Yucat
35. Carrerillo
62. Las Peñas
78. Remedios
85. Pilar
86. San José
87. Suburbios
88. V.del Rosario
91. Matorrales
92. Impira
116. S.Antonio
119. Cosme
125. Salto
126. Cap.de Rodriguez
127. Los Zorros
128. Pampayasta Norte
129. Pampayasta Sur
130. Punta del Agua

Zona 4 (23 pedanías)

23. Italo
24. Algodón
25. Las Mojarras
27. Villa María
28. Villa Nueva
29. Chazón
36. Chucul
37. Reducción
38. Carlota
45. Las Tunas
53. La Amarga
54. San Martín
55. La Paz
56. Independencia
68. La Cautiva
90. Arroyo de Alvarez
93. Calchin
109. Libertad
110. Sacanta
111. Juarez Celman
141. Litin
142. Ballesteros
145. Loboy

Zona 2 (8 pedanías)

39. Coloniss
40. Espinillos
41. Saladillo
42. Cruz Alta
43. Liniers
44. Calderas
143. B.Ville
144. Ascasubi

Zona 3 (20 pedanías)

13. Calera Norte
19. Sarmiento
20. Necochea
22 Jaguales
63. Tegua
65. Rio Cuarto
66. Achira
67. 3 de Febrero
71. Castaños
73. Tala
74. Santa Rosa
75. Suburbios
76. Quebracho
77. Yegua Muerta
79. Villa Norte
89. Oratorio de Peralta
106. San Francisco
107. Arroyito
108. Concepción
112. Calera Sur

Zona 5 (6 pedanías)

5. Monsalvo
57. Dolores
58. San Antonio
60. San Roque
61. Santiago
97. Tránsito

Zona 6 (5 pedanías)

9. Cañas
10. Constitución
101. Dolores
102. Las Rosas
105. La Paz

Zona 7 (63 pedanías)

2. Los Reartes
3. Los Molinos
4. Santa Rosa
6. Cda. de Alvarez
8. Río de los Sauces
11. San Vicente
12. Río Ceballos
14. Cruz del Eje
15. Pichanas
16. Higueras
17. San Marcos
18. Candelaria
21. El Cuero
30. Quilino
31. Toyos.
32. Copacabana
33. Parroquia
34. Manzanas
46. Guasapampa
47. Argentina
48. Ciénaga del Coro
49. San Carlos
50. Chancaní y República
51. Parroquia
52. Salsacate
59. Rosario
64. San Bartolomé
69. Chalacea
70. Timón Cruz
72. Esquina
80. Higuerillas
81. Villa de María
82. Estancia
83. Candelaria Sur
84. Candelaria Norte
94. Ambul
95. El Carmen
96. Panaholma
98. Toscas
99. San Pedro
100. Nono
103. San Javier
104. Luyaba
113. Lagunilla
114. Caseores
115. Alta Gracia
117. Potrero de Garay
118. San Isidro
120. Aguada del Monte
121. Cerrillos
122. Chuñahuasi
123. San Francisco
124. Caminiaga
131. Totoral
132. Macha
133. Candelaria
134. Sinsacate
135. Río Pinto
136. San Pedro
137. Intihuasi
138. Parroquia
139. San José (dormida)
140. Mercedes

APENDICE IV

PRUEBAS DE NORMALIDAD

****
CALCULO DE CONSTANTES ESTADISTICAS Y PRUEBAS DE NORMALIDAD
****

Los programas fueron facilitados por el Lic. José Cocilovo (Inst. de Antropología- Fac. de Filos. y Humanidades -UNC). La letra R junto a los valores de las pruebas W y U indica rechazo de la hipótesis para la variable indicada, según las tablas correspondientes de Shapiro (1965) y Pearson (1954)

ZONA 1

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	19	360.72	761.96	6492.96	0.474	-1.129	0.945	3.240
2	19	67.51	486.17	1215.22	0.735	-1.162	0.944	3.281
3	19	6.44	46.78	115.98	5.061	7.299	0.690 R	4.190
4	19	83.66	171.93	1505.88	4.847	6.907	0.717 R	4.074
5	19	15.06	109.39	271.23	1.004	-1.390	0.806	2.958 R
6	15	2350.11	2325.30	32901.61	0.846	1.256	0.961	4.090
7	13	2.78	12.64	33.47	6.273	12.268	0.466 R	3.759 R
8	19	237.61	1934.00	4277.15	-1.581	0.304	0.930	3.632

ZONA 2

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	8	149.58	596.29	1047.11	1.307	-0.307	0.895	2.841
2	2	0.07	0.94	0.07	0.000	-17.748	0.999 R	1.414 R
3	8	0.11	9.63	0.82	1.523	-0.508	0.875	2.788
4	8	6.91	47.12	48.39	1.024	-0.197	0.948	2.904
5	8	46.09	234.70	322.67	-0.557	-0.963	0.945	2.769
6	8	9145.61	1648.44	64019.28	-0.058	0.438	0.987 R	3.391
7	8	0.01	3.24	0.13	-0.198	-1.405	0.934	2.748
8	8	42.33	829.00	299.87	1.120	0.920	0.953	3.208

ZONA 3

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	19	147.32	1049.52	2651.90	-0.880	0.956	0.973	4.346
2	14	4.91	21.23	63.90	4.160	4.871	0.688 R	3.491
3	18	10.54	101.14	281.25	1.146	-1.042	0.903	3.029
4	19	29.23	98.99	526.29	3.070	2.056	0.815 R	3.373
5	17	5.96	48.18	95.46	1.388	-1.003	0.884	2.938 R
6	17	5197.46	3044.94	83159.38	2.879	2.595	0.873	3.900
7	16	1.05	13.87	15.78	7.396	15.785	0.445 R	4.298
8	19	609.33	2153.00	10968.10	0.955	-0.372	0.970	3.645

ZONA 4

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	23	96.21	1782.78	2116.70	-0.469	-0.078	0.983	3.990
2	19	17.27	57.33	310.94	2.709	0.158	0.729 R	2.872 R
3	23	1.04	56.16	23.95	1.243	0.065	0.904	4.093
4	23	3.47	48.57	76.49	4.189	3.117	0.719 R	3.590
5	23	18.27	182.49	402.09	0.382	-0.995	0.972	3.597
6	22	11177.61	5111.33	234729.99	5.771	0.138	0.676	4.326
7	20	0.17	11.20	3.41	7.026	14.968	0.617	4.789 R
8	23	125.05	2350.00	2751.30	0.738	-0.010	0.969	4.023

ZONA 5

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	6	20.63	146.69	103.15	1.013	0.436	0.975	2.874
2	5	6.53	8.44	26.13	3.462	5.456	0.666 R	2.320
3	6	0.01	0.68	0.07	2.330	2.192	0.857	2.668
4	6	44.24	64.99	221.20	2.193	2.291	0.888	2.838
5	2	35.29	8.40	35.29	0.000	-17.748	0.999 R	1.414
6	5	19244.10	1146.87	70976.43	2.858	3.526	0.771	2.364
7	6	126.55	97.40	632.77	2.682	3.022	0.828	2.688
8	6	1133.60	762.00	5668.00	1.603	-0.845	0.848	2.435

ZONA 6

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	5	213.32	179.38	853.31	0.779	-3.556	0.845	2.140
2	2	0.19	0.63	0.19	0.000	-17.748	0.999	1.414
3	5	20.90	17.66	83.63	1.093	-3.566	0.742	1.992
4	5	3593.63	509.75	14374.53	2.354	2.730	0.839	2.483
5	4	0.01	0.52	0.04	0.501	-6.367	0.891	2.099
6	4	11246.63	675.12	33739.90	1.784	-0.466	0.879	2.129
7	4	0.08	2.91	0.26	2.675	4.009	0.860	2.247
8	5	600.00	545.00	2400.00	-0.977	0.249	0.973	2.612

ZONA 7

VARIABLE	NRO.OBSERV.	VARIANZA	SUMA SIMP.	SUMA CUADR.	ASIMETRIA	CURTOSIS	PRUEBA W	PRUEBA U
1	64	110.85	1589.76	6984.12	1.634	-0.100	0.000	4.511
2	25	4.53	25.71	108.94	7.605	14.509	0.000	4.549
3	48	4.30	48.12	202.17	11.305	24.889	0.578	5.620 R
4	61	44.37	428.42	2662.44	8.053	16.881	0.578	6.123 R
5	19	1.16	12.30	20.88	4.966	5.564	0.000	3.430
6	30	3590.23	3048.23	104116.89	0.870	-1.227	0.000	3.540
7	39	2.53	55.50	73.52	4.241	4.155	0.000	4.270
8	64	219.31	5949.00	13816.85	1.474	0.609	0.000	4.456

EN ESTA ZONA NO CORRESPONDE LA PRUEBA W POR SER MAYOR QUE 50 EL NUMERO DE OBSERVACIONES

B I B L I O G R A F I A (*)

A. Básica

* Anderson T.W. (1958): Introduction to Multivariate Statistical Analysis (Wiley-New York)

* Hope, Keith (1968): Methods of Multivariate Analysis (Unibooks-London).

* Kendall M.G. (1972): A Course in Multivariate Analysis (Briffin-London)

* Kendall M.G. (1965): Discrimination and Classification (en "Proceedings of the first international symposium in multivariate analysis"-Dayton-Ohio- 1965 - Academic Press - New York p.165/185)

* Kendall M.G. and Stuart (1967): The Advanced Theory of Statistics (Vol. II y III) (Griffin-London).

* King, Leslie J. (1969): Statistical Analysis in Geography (Prentice Hall - New Jersey).

* Kullback, Solomon (1959): Information Theory and Statistics (Wiley - New York).

* Rao C.R. (1952): Advanced Statistical Methods in Biometric Research (Wiley - New York)

* Sokal R.R. and Sneath P.H. (1963): Principles of Numerical Taxonomy (Freeman and Co. - London).

B. Adicional

* Adelman, Irma and Morris, Cynthia T. (1968a): Performance criteria for evaluating economic development potential: and operational approach. (The Quarterly Journal of Economics - V. LXXXII - N°2 - p. 260/280)

* Adelman, Irma and Morris, Cynthia T. (1968a): An econometric model of socio-economic and political change in underdeveloped countries - (The American Economic Review - V. LVIII - N°5 - Part.1 - Pag. 1184/1218)

* Adelman, Irma and Morris, Cynthia T. (1970a): Comentario y réplica sobre el artículo (1968b). Comentan: Sara S. Berry (p.222); Peter Eckstein (p.227); Répiica de los autores (p.236) - (The American Economic Review - V. LX, N°1).

(*) El material señalado con asterisco es el citado en el trabajo; el resto ha sido consultado.

* Adelman, Irma and Morris, Cynthia T. (1970b): Factor analysis and Gross National Product: a reply. (The Quartely Journal of Economics - V. LXXXIV - N°4 p. 651/662)

Anderson M.W. and Benning R. (1970): A distribution free discrimination procedurs based on clustering (Trans.of inf.theory-V.16 p.541/548 N.York).

Anderson T.W. and Bahadur R.R. (1972): Clasification into two multivariate.normal distribution with different covariance matrices (The Annals of Mathematical Statistics V. 33 - p. 420/431).

Anderson J.A. (1969): Discrimination between k populations with constraints on the probabilities of misclassification (Journal of the Royal Statistical Society - Serie B - V.XXXI N°1 p. 123/139)

* Araoz Julián (1968): Asociación en Taxonomía Numérica - (Univ. Central de Venezuela - Facultad de Ciencias-Depto. de Computación-Publicación 68-12-Caracas).

Banco de la Provincia de Córdoba (1968) Provincia de Córdoba Esquema Económico General. (Memoria y Balance de 1968 - pág. 33)

Banco de la Provincia de Córdoba (1975): La Participación en la tenencia de la tierra en la Provincia de Córdoba (oficina de Estadística).

* Berry B.G.L. (1965): Patrones básicos del desarrollo económico (en Atlas del Desarrollo Económico - N.Guinsbourg Eudeba - Bs.As.).

Bolch and Huang (1974): Multivariate Statistical Methods for Business and Economics (Prentice Hall - Englewood Cliffs - New Jersey)

Bolshev (1969): Cluster Analysis (Bulletin of the JSI V.43 - p. 411/425).

* Boulton D.M. and Wallace C.S. (1970): A program for numerical classification (The computer journal -V.13 N°1 p. 63/69).

* Box G.E.(1949): A general distribution theory for a class of likelihood criteria (Biometrika - V.36 p.317/346)

* Box G.E.(1953): Nonnormality and tests of variances (Biometrika V. 40 p.318/335).

* Brown S. and Trott Charles (1968): Grouping tendencies in an economic regionalization of Poland (Annals Assoc. American Geographers 1968 pag.327/342).

Bryan J. (1951): The generalizad discriminant function: mathematical. foundations and computational routine (Harvard educational review V.21 N°2-p.90/95).

* Bunge Mario (1969): La investigación científica (Ariel-Barcelona).

* Casetti E. (1964): Multiple Discriminnnt functions.(Technical Report N°11, Computer Applications in the Earth

Sciences Project, Department of Geography, Northwestern University).

* Castells M. (1972): Las nuevas fronteras de la metodología sociológica (Revista Latinoamericana de Ciencias Sociales - Flacso - N°3- p. 143/170).

* Cooley W. and Lohnes P.R.(1962): Multivariate Procedures for the Behavioral Sciences (John Wiley-New York).

Chien Pai,H (1968):A note on discrimination in the case of unequal covariance matrices (Biometrika V.55 p. 586/587).

Defrise-Grussenhoven E. (1952): Discrimination des populations voisines (Buletin Inst.Roy.Sc.Mat. - Belgique V.28-N°46 p. 1/34)

Defrise-Grussenhoven E. (1966):Escala de masculinidad-feminei-dad basada en una función discriminante (Acta Genética-V.16 - p.198-208).

Fortier and Solomon (1965): Clustering Procedures (en Procce@ dings of the first international symposium in multivariate analysis Dayton-Ohio-Academic Press N. YOrk - p. 493/506).

Friedman H.P. and Rubin,J. (1967): On some invariant criteria for grouping data (Journal of the American Statistical Association V.62-N°320 p. 1159/78)

* Gales, Kathleen (1957): Discriminant functions of socio-economic class (Applied Statistics-V.VJ-N°2 pag. 123/132 Publisher of the Royal Statistical Society).

* Galtung,Johan (1966): Teoría y métodos de la Investigación Social (Tomos I y II) EUDEBA-Bs.As.

Georgescu H. (1968): Sur quelques problemes de classification mathematique (Bull.Math. de Roumanie -T.12-XX (60) N°3 - p. 32/37).

Geisser, S.(1965): Predictive discrimination (de Proceedings of the first international symposium in multi-variate analysis-Dayton; Ohio-Academic Press-N. York p.149/168).

Gessanan M.P. and Gessanan P.H.(1972): A comparison of some multivariate discrimination procedures (Journal of the American Statistical Association-V.67 N°338, p. 469-477).

* Gilbert E.(1969): The effect of unequal var-covar-matrices on Fisher's linear discriminant function (Biometrics V.25 - p.505/515).

* Gnadesikan R.and Wilk (1968): Data analytic methods in multivariate statistical analysis (en "Multivariate Analysis II" Proceedings of the second international symposium in multivariate analysis - Dayton Ohio - p.593/636).

* Goldenhersch de Roitter, Hebe (1973):Métodos para clasificar observaciones multivariantes-un caso de aplicación al campo económico.(Inst.deMat. y Est. FCE UNC. Trabajo no publicado).

Gower J.C.(1966): Some distance properties of latent root and vector methods used in multivariate analysis (Biometrika-V.53-N°3 y 4 - p.325/338).

Gupta Shanti (1965): On some selection and ranking procedures for multivariate normal populations using distance functions (en Proceedings of the first international symposium in multivariate analysis Dayton Ohio-pág. 457/475).

* Haan, H.H.de (1972): Multivariate Analysis and Development: en actual review of the work of Adelman and Morris (Netherlands School of Economics Centre of Development Planning-Rotterdam - Discussión paper N°17).

Hartigan J.A. (1972): Direct clustering of a data matrix (Journal of the American Statistical Association V67 N°337 - p.123/129).

* Healy M.J.R. (1972): Drawing a probability ellipse (Applied Statistics - V.21 N°2 - p.202/203 Ed. by the Royal Stat. Soc.)

* Holloway L.N. and Dunn. O.J.(1967): The robustness of Hotelling's T^2 (Journal of the American Statistical Association - V.62 N°317 - p.124/136).

* Ito, Koichi (1968): On the effect of heteroscedasticity an nonormality upon some multivariate test procedures (en Multivariate Analysis II-Proceedings of the second international symposium in multivariate analysis-Dayton-Ohio - pág.87/119).

James G.(1954): Tests of linear hipotesis in univariate and multivariate analysis when the ratios of the population variances are unknown (Biometrika V.41 - p.19/43).

Jardine and Sibson (1971): Mathematical Taxonomy (Wiley-New York)

* Kaminsky Mario (1971) The structure of multiple output dairy farms in the centro santafesino region of Argentina; a multivariate analysis. (Tesis en University of Wisconsin)

* Kendall M.G.(1939): The geographical distribution of crop productivity in England (Journal of the Royal Statistical Society CII-p.21/48).

Kendall M.G.(1961): A course in the geometry of n-dimensions Griffin-London).

* Kowalsky Ch.J.(1972):A comentary on the use of multivariate statistical methods in anthropometric research (American Journal of Physical Anthropology V.36 N°1 - p.119/132)

Kullback S.(1952): An application of information theory to Multivariate analysis (The annals of mathematical statistics-V.23-p.88/102).

* Lebedinsky,Mauricio(1968): Del Sub-desarrollo al desarrollo (Ed. Quipo-Bs.As.)

* Lebedinsky, Mauricio (1970): Argentina: bases económicas de la política (E. Quipo, Bs.As.)

* Lockhart R.S.(1967) The assumption of multivariate normality (The British journal of mathematical and statistical psychology- V.20 - p.63-69).

Matusita, Kameo (1965) A distance and relate statistics in multivariate analysis (en Proceedings of the first international symposium in multivariate analysis Dayton-Ohio- p.187-200)

Marshall A.W. and Olkin,I.(1968): A general approach to some screening and classification problems (Journal of the Royal Statistical Society-SerieB-V.30p.407/43)

Mc.Henry H.andGiles E.(1971): Morphological variation and heritability in three melanesian populations. A multivariate approach. (American JOurnal of Phisical Anthropology-V.35-N°2).

* Naciones Unidas (1972): Estudio sobre la clasificación económica y social de las países de América Latina Bol Econ. de América Latina. V.XVII-N°2- p.155-218)

Neymark Y.(1970): A linear minimax classification algorithm : (Enginmering cybernetics-N°2 - p.328/336).

* Pearson E.S. David H.A. and Hartley H.O. (1954) The distribution of the ratio, in a single normal sample, of range to standard deviation (Biometrika-V.41 - p.482/493).

Pearson E.S. and Stephens M.A. (1965): The rationof range to standard deviation in the same normal sample (Biometrika-V.51 - p.484/486).

* Porebski, O.(1966): On the interrelated nature of the multivariate statistics used in discriminatory analysis (The British Journal of mathematical and statistical psychology-V.19 - p. 197/214).

Rao,R.C.(1969): Recent advances in discriminatory analysis (Journal of the Indian society of agricultural statistics -V.21 p 1/15)

* Rayner, A.C.(1970) A critique to the approach adopted by Adelman and Morris (The quartely journal of economics. V. LXXXIV N°4 - p.639/650).

* Ríos, Raúl Arturo (1968): Regiones socio-económicas de la Provincia de Córdoba (Seminario realizado por los estudiantes Rosselot y Abraham en el Instituto de Economía, FCE. UNC.)

* Rogers, Andrei (1971): Matrix Methods in Urban and Regional Analysis (Holden Day - San Fco.)

Roulon, Ph. (1951): Distinctions between discriminant and regression analysis and a geometric interpretation of the discriminant function (Harvard Educational Review - V.31 - N°2 -p.80/90).

* Secretaría Ministerio de Desarrollo de la Provincia de Córdoba (1969/70): Preinformes de los sectores directamente productivos; Tomos I a V.

* Serrato, Esmeralda (1969): Bases metodológicas para la determinación de regiones homogéneas mediante técnicas estadísticas de análisis multivariado. Aplicación de la función discriminante (5° reunión de centros de investigación económica - La Plata)

* Shapiro S.S. and Wilk M.B.(1965): An analysis of Variance Test for Normality (Biometrika-V.52 p.591/612)

* Tintner G.(1946): Some applications of multivariate analysis into economic data (Journal of the American Statistical Association-V.41 p.472/500).

Varsavsky, Oscar (1969): Entropía y Taxonomía Numérica (U.C.V. Fac.de Ciencias-Dpto. de Computación-Pubic. 69/01-Caracas).

* Varsavsky, Oscar(1971): América Latina; Modelos Matemáticos (Cap.II)(Ed.Universitaria-Santiago-Chile)

Wagle,B.(1968): Multivariate B distribution and a test for multivariate normality (Journal of the Royal Statistical Society Serie B-V.30-p.511/516)

* Wallace C.S. and Boulton D.M. (1968): An information measure for classification (The computer journal-V. 11-p.185/195).

* Williams, E.J.(1952) Some exact tests in multivariate analysis (Biometrika - V.39 - p.17/31)

* Williams,E.J.(1955): Significance tests form discriminant functions and linear functional relationships (Biometrika-V.42-p.360/381)

Yaglom A.M. et Yaglom I.M.(1959): Probabilité et Information (Dunod, Paris)

* Zhezhel, Yun (1968): The efficiency of a linear discriminant function for arbitrary distributions (Engineering Cybernetics-N°6 - p.107/111)