ESTADO ACTUAL DE LA DISCRIMINACION ESTADISTICA

T E S I S

QUE PARA OBTENER EL GRADO
DE MAESTRO EN CIENCIAS
PRESENTA EL ACTUARIO

FRANCISCO JAVIER ARANDA ORDAZ

CD. UNIVERSITARIA, D.F. MAYO 1976

RECONOCIMIENTO

Agradezco al Consejo Nacional de Ciencia y Tecnología, al Departamento de Matemáticas de la Facultad de Ciencias y al Instituto de Investigación en Matemáticas y en Sistemas, el apoyo prestado, sin el cual, este trabajo no podría haberse realizado.

Así también, agradezco al Dr. Ignacio Méndez el asesoramiento y aliento proporcionados durante la elaboración de este trabajo.

A mi madre, a quien todo le debo

a mi hermano

y a ϕ

FE DE ERRATAS

Página	Renglón	Dice	Debe decir
80	2	no singulares	no singulares de la forma
142	2	son dicotómicas	sean dicotómicas
143	9	son discretas	sean discretas
144	9	restante no es	restante no sea

I N D I C E

Página

SUMARIO

El objetivo de este estudio es llevar a cabo una revisión del trabajo producido en relación al tema de Discriminación Estadística - (también conocido como Clasificación, Identificación, Asignación, o Análisis Discriminante). Este tema ha interesado a un gran número de - investigadores y está lejano de ser agotado.

En el primer capítulo se presenta el problema y algunos ejemplos de situaciones en las cuales surge la necesidad de desarrollar técnicas de discriminación. Se determina la naturaleza del problema que se desea solucionar y se discuten las diferencias que existen con problemas relacionados. Se efectúa el planteamiento formal de la solución y se presentan diferentes situaciones y una breve idea de las soluciones. Se anexa una nota histórica en relación a la evolución de - las soluciones.

El segundo capítulo comprende la presentación de algunas de las soluciones propuestas (frecuentistas o clásicas, bayesianas y no paramétricas). Se presenta su desarrollo, suposiciones subyacentes, ventajas y desventajas.

En el tercer capítulo se presentan los resultados de trabajos relacionados indirectamente con el problema como son: la estimación de probabilidades de clasificación (tanto correcta como equivocada), criterios de selección de variables, efecto de clasificación equivocada en las muestras de "ensayo", tratamiento de valores faltantes en las observaciones. Además se presentan algunos trabajos que tratan sobre procedimientos iterativos, discriminación secuencial y comparación entre algunos métodos de clasificación.

En el cuarto capítulo se presentan y discuten brevemente ejemplos de aplicación de algunos de los métodos.

Se anexa un apéndice en el que se presentan programas para computadora que pueden utilizarse para efectuar discriminación.

Se incluye una bibliografía que se espera sea útil para aquellos interesados en el tema.

INTRODUCCION. -

El análisis multivariado es una rama de la Estadística que se dedica al estudio simultáneo de variables aleatorias las cuales están correlacionadas entre sí. Es, por así decirlo, el enriquecimiento del análisis convencional de una sola variable permitiendo estudiar los procesos con una "herramienta" más adecuada ya que se gana para el análisis la información que prestan en conjunto las variables estudiadas, lo que en ocasiones no brindan individualmente.

La idea del análisis multivariado no es nueva (en relación a la edad de la estadística moderna) y muchos trabajos técnicos se han desarrollado desde hace tiempo. Sin embargo, hasta época muy reciente la aplicación de resultados del análisis multivariado a datos reales requería de un considerable esfuerzo, largas horas de cálculos manuales o con calculadoras de escritorio y un conocimiento, o mejor -

aún, un sentido "artístico" para escoger niveles de significancia e interpretar resultados. Esta situación se alteró drásticamente con el advenimiento -y cada vez mayor difusión- de los recursos de cálculo de gran velocidad proporcionados por las computadoras digitales. Además, el impacto causado por los paquetes de programas de biblioteca para tales computadoras ha puesto a la disposición de un gran número de personas rutinas de cómputo óptimas (o casi óptimas) para resolver problemas reales que requieren el uso de técnicas del análisis multivariado. Aún cuando se ha de hacer la observación de que las rutinas de cómputo sólo serán realmente útiles para aquellos usuarios que entienden completamente las suposiciones subyacentes en el modelo para el cual la rutina fue diseñada.

Entre los temas del análisis multivariado se encuentra el de discriminación (también conocido como de análisis discriminante, clasificación o identificación) el cual ha servido por muchos años en antropología y taxonomía, y que actualmente está encontrando una creciente aplicación en áreas tan disímbolas como son el reconocimiento de patrones o el mercadeo y la publicidad.

Debido al interés de este tema se han producido (y continúan produciéndose) un gran número de trabajos alrededor de él, lo que hace necesario un trabajo monográfico de revisión de la materia. Este

trabajo tiene esa finalidad y se espera que facilite la labor de futuros interesados en esta área de la Estadística.

La revisión monográfica cubre los artículos aparecidos en relación al tema en las revistas: Annals of Mathematical Statistics (posteriormente con el nombre de Annals of Statistics), a partir de 1943 - Journal of the American Statistical Association a partir de 1943 Biometrics y Technometrics desde su aparición, Biometrika a partir de - 1952, Journal of the Royal Statistical Society series A, B y C desde 1945 aproximadamente, todas estas revistas se encuentran en el acervo de la Biblioteca del IIMAS. Además se revisaron otros artículos a los cuales se hacía referencia en los inicialmente considerados. Esta búsqueda de trabajos relacionados comprende la captación de material en revistas que no existen en México como es Sankhyā, o la búsqueda de artículos en revistas de negocios, economía, etc., como el Journal of Advertising Research, Quaterly Journal of Economics, también se revisó material indispensable como los reportes técnicos de - Fix y Hodges, elaborados para la Fuerza Aérea de los Estados Unidos.

La tarea de recopilación se vio dificultada por el gran número de trabajos respecto al tema. Como es bien sabido la evolución de la investigación principia con un trabajo que abre brecha y que en general se continúa y diversifica por un sinnúmero de investigadores interesa-

dos en el tema, así también, es muy notable en esta rama el interés por épocas y las modalidades que esto causa en los trabajos.

El trabajo trata de presentar la evolución cronológica de las soluciones, homogeneizando la notación (que cambiaba de autor a autor y época a época). Se debe aclarar que no se trataba de verificar rigurosamente los resultados de los trabajos revisados (lo que hubiese sido una labor titánica) sino simplemente ubicarlos, encontrar relaciones entre ellos y clasificarlos en cierta forma.

Como un subproducto de este trabajo quedarán todos los artículos, reportes técnicos, tesis doctorales, folletos conseguidos y revisados, estos trabajos estarán a disposición de los interesados en la Biblioteca del IIMAS, clasificados de acuerdo al tema principal del trabajo.

Otro subproducto es la bibliografía ubicada que aparece al final de esta revisión lo cual presenta de manera casi exhaustiva los trabajos aparecidos con respecto a la materia.

Finalmente he de señalar que aunque no se presenta algún resultado original, este trabajo puede servir como punto inicial para aquellos que se interesen ya sea en la investigación acerca del tema o en -

la aplicación de los métodos ya propuestos.

I.- EL PROBLEMA DE DISCRIMINACION.-

I.1 Presentación y ejemplos de situaciones en las que surge el problema de discriminación.

Entre los problemas que trata el análisis de variables aleatorias multidimensionales se encuentra el de discriminación también conocido como clasificación, o (como propone Rao (1965, 1973)) identificación.[1] Como quiera que sea tal problema ha atraído la atención de un gran número de investigadores de varias disciplinas (no exclusivamente estadísticos). Esta rama de la investigación que se desea identificar un "individuo" en base a un conjunto de mediciones efectuadas sobre él (un animal, un vegetal, un fósil, un proceso, etc.) como perteneciente a alguna de k posibles clases o poblaciones, las cuales se consideran determinadas.

---

1. Aunque en la literatura se encuentran otros nombres asociados a problemas semejantes como: asignación, predicción, clasificación de patrones y selección.

Los siguientes pueden servir como ejemplos de tales situaciones.

1) Determinación del sexo de huesos humanos fósiles basándose en mediciones antropométricas en cráneos (Barnard (1924)).
2) Identificación de un individuo como perteneciente a uno de dos grupor étnicos en base a mediciones de estatura, altura del individuo sentado, longitud de la cabeza, largo de la nariz, perímetro torácico, etc.
3) Clasificación de riesgos -buenos y malos- en préstamos para compra de autos basada en precio del automóvil, enganche, ingreso del individuo que solicita el préstamo, plazo a pagar el préstamo. (Durand (1941)).
4) Clasificación por áreas de estudio de candidatos a ingresar en una universidad en base a las calificaciones obtenidas en diferentes materias en su bachillerato.
5) Clasificación de un individuo como buen o mal agente de ventas, basándose en la puntuación obtenida en diferentes pruebas psicológicas (Wallace y Travers (1938)).
6) Clasificación en bienes de consumo y bienes de producción basándose en características como longitud, amplitud y tasa de cambio en el ciclo de precio de los artículos (Tintner (1946)).

7) Diagnósticos de personas que padecen o no padecen una enfermedad particular (diabéticos y no diabéticos) basados en diferentes pruebas clínicas como cantidad de colesterol en la sangre, de urea, presión sanguínea, etc.

8) Determinar la procedencia de un lote de carbón de entre dos minas posibles en base a mediciones de materia volátil, porcentaje de ceniza del carbón, "carbón fijo", etc. (Baten y Dewitt (1944)).

9) Asignar calificación (MB, B, S o NA) a un individuo en base a sus notas obtenidas en exámenes, ejercicios, participación en clase, etc.

10) Discriminar entre dos especies de vegetales: Iris Setosa e Iris Versicolor (clásico ejemplo debido a Fisher (1936)), basándose en las mediciones obtenidas de longitud de sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo.

11) Pertenencia a una de ciertas clases de neurosis determinada, para un individuo en base a medidas como: estado de ansiedad, sentimiento de culpa, etc.

12) Determinación de tipo de sociedad (edad de bronce o edad de hierro) a la que perteneció un individuo cuyo esqueleto fue encontrado en Inglaterra basándose en características como: altura máxima, anchura máxima, anchura frontal mínima, arco sagital total, arco biporial transverso a través del "bregma", altura basi-bregmática y altura facial superior. (Goodwin y Morant (1940), Rao

(1948)).

13) Discriminación entre individuos con dispepsia no ulcerosa y controles en base a mediciones de ansiedad, dependencia, perfeccionismo, culpa. (Hamilton (1950)).

14) Determinación del autor de un artículo diputado, si fue escrito por Cervantes o no, en base a la frecuencia de diferentes palabras.

15) Determinar si un suelo tiene o no "azotobacteria" en base a las cantidades de fósforo, nitrógeno y P que contiene dicho suelo (Cox y Martin (1937))

16) Discriminar entre pilotos de aviones e ingenieros en base a resultados obtenidos en pruebas de inteligencia, coordinación. (Travers (1939)).

17) Discriminar entre dos clases diferentes de cadáveres de cordero, digamos Royal y Tallarock, en base a mediciones del ancho del hombro, grosor del flanco, ancho del flanco, longitud de la pata, etc.

18) Identificar a un individuo como perteneciente a uno de dos conjuntos de poblaciones en base a mediciones efectuadas en él. (Burnaby (1966), Rao (1966)).

19) Discriminar entre parejas de gemelos (del mismo sexo) monocigóticos y dicigóticos, en base a mediciones antropométricas obtenidas en ellos. (Bartlett y Please (1963)).

I.2 Caracterización del problema y diferencias con problemas afines.

En los ejemplos presentados anteriormente pueden observarse ciertas características comunes que son:

(i) Se desea clasificar a un individuo en una de k poblaciones ($k \geq 2$). Además no se desea identificar a las poblaciones sino discriminar a los individuos, es decir, determinar a qué población (definida a priori) pertenece un individuo dado.

(ii) Debido a la complejidad que implicaría efectuar un gran número de mediciones para caracterizar al individuo en base a ellas se recurre a utilizar unas cuantas, sean estas p, que sirven como indicadores, i.e., mediciones que se espera varien entre las poblaciones y permitan identificar la población de la que proviene el individuo.

(iii) Por las características mismas del problema existe la posibilidad de cometer el error de no clasificar correctamente al individuo, lo que en general acarreará un costo (o pérdida). De lo anterior se desprende la necesidad de encontrar una regla o criterio de clasificación óptimo en cierto sentido, como el de minimizar costos.

Se encuentra cierto desacuerdo en la literatura respecto no sólo a la denominación para el problema sino también a su definición específica. Así se puede encontrar, por ejemplo, que Kendall usa el término discriminación para el problema del que trata este trabajo mientras Anderson lo designa como clasificación, término que Kendall reseva para otra técnica asociada que, sin embargo, no es la misma. En este trabajo se ha adoptado el término discriminación y con el propósito de puntualizar las diferencias con otras técnicas relacionadas se hace aquí una breve presentación de ellos.

a) Reconocimiento en patrones. Tal técnica tiene como objetivo el desarrollar medios mecánicos de reconocimiento de un patrón predeterminado, por ejemplo: programar una máquina para distinguir entre las letras del alfabeto, o diferentes formas geométricas. Es esencial que la forma del patrón sea determinada de antemano, permitiéndose algún grado de tolerancia.

b) Clasificación. En este caso no hay patrones predeterminados. El problema es el siguiente: dado un conjunto de objetos (lenguajes, organismos biológicos, tipos de cerámica, ciencias, etc.) se debe clasificarlos de alguna forma útil, con cierta jerarquía o estructura. - Esta clasificación no es única pues a posteriori.

c) Agrupamiento (Clustering). Este término incluye una colección de técnicas que son usadas para agrupar entidades multidimensionales de acuerdo a varios criterios de sus grados de homogeneidad y heterogeneidad. El problema central es repartir n individuos en k grupos ($k \geq n$) de tal manera que, en algún sentido, los individuos dentro de un grupo muestren homogeneidad y los individuos de grupos diferentes sean heterogéneos entre sí.

Puede observarse de las definiciones de los problemas relacionados las diferencias que existen con el de discriminación.

I.3 Planteamiento formal y una solución vía teoría de decisiones.

El problema de discriminación se puede plantear de la siguiente manera en términos de funciones de decisión estadística:

A cada individuo, objeto o proceso se le efectúan p mediciones i.e., se observan en él los valores de p v.a. y en base al vector de observaciones se desea determinar a qué población pertenece al individuo, objeto o proceso. Supóngase que se conocen las probabilidades a priori, π_i, de que un individuo pertenezca a la población P_i ($i = 1, \ldots, k$, $k \geq 2$) y además se puede fijar la pérdida c_{ij} asociadas a la asignación de un individuo de la clase i a la j (claramente $c_{ii} = 0$), esta asignación se puede efectuar definiendo primero k regiones $R_1, R_2, \ldots, R_k$, contenidas en el espacio de observaciones $\mathbb{R}^p$ y tales que $\bigcup_{i=1}^{k} R_i = \mathbb{R}^p$ y $R_i \cap R_j = \phi$ $i \neq j$, con las regiones así definidas la discriminación de un individuo en base a un conjunto de mediciones tomadas en él, $\underline{x}$, se lleva a cabo de la siguiente forma: si $\underline{x} \in R_i$ el in

dividuo con vector de mediciones $\underline{x}$ se clasifica como proveniente de la población P_i. Esta será una regla de decisión no aleatorizada y - en este caso el problema se reduce a encontrar la partición $(R_i)_{i=1}^{k}$ óptima, en el sentido de minimizar la pérdida asociada a clasificar algún individuo.

Otra posible regla de decisión consiste en determinar una función vectorial de $\underline{x}$, $\Lambda(\underline{x}) = (\lambda_1(\underline{x}), \ldots, \lambda_k(\underline{x}))$, tal que $\lambda_i(\underline{x}) \geq 0$ y $\sum_{i=1}^{k} \lambda_i(\underline{x}) = 1$, y asignar un individuo con mediciones $\underline{x}$ a la i-ésima población con probabilidad $\lambda_i(\underline{x})$, $i = 1, \ldots, k$. O sea, que una vez observado $\underline{x}$, un experimento aleatorio se realiza para generar una variable aleatoria la cual toma los valores $1, \ldots, k$ con probabilidades $\lambda_1(\underline{x}), \ldots, \lambda_k(\underline{x})$ respectivamente y si el resultado es i el individuo se asigna a la población P_i. Así planteado el problema es encontrar la función Λ óptima en el sentido antes mencionado.

Para encontrar las soluciones, tanto bajo la regla de decisión no aleatorizada como con la aleatorizada, se considerarán: $P_1, \ldots, P_k$ poblaciones con funciones $f_1, \ldots, f_k$ de densidad de probabilidad con respecto a una medida σ - finita en las k poblaciones y c_{ij} la pérdida asociada a la asignación de un individuo de la clase i a la j. En estos términos la pérdida esperada asociada a la aplicación de una re

gla dada para la clasificación de un individuo proveniente de la pobla ción P_i es, bajo la regla no aleatorizada:

$$L_i = \int_{R_1} c_{i1} f_i (\underline{x}) \, dv + \ldots + \int_{R_k} c_{ik} f_i (x) \, dv$$

y para la regla aleatorizada:

$$L_i = \int_{\mathbb{R}^p} (c_{i1} \lambda_1 (\underline{x}) + \ldots + c_{ik} \lambda_k (\underline{x})) f_i \, dv$$

En cualquier caso el vector de pérdida $(L_1, \ldots, L_k)$ caracte riza a una regla de decisión dada y permite llevar a cabo comparacio nes entre diferentes reglas. Así, si hay dos reglas de decisión C1 y C2 son vectores de pérdida asociada $(L_{11}, \ldots, L_{1k})$, $(L_{21}, \ldots, L_{2k})$ se tiene que C1 es mejor que C2 si

$$L_{i1} \leq L_{i2} \quad , \quad i = 1, \ldots, k$$

y $L_{i1} < L_{i2}$ al menos para alguna i. Si la igualdad se tiene para to da i las reglas son equivalentes, en caso de que $L_{i1} > L_{i2}$ para algu nos valores de i y $L_{i1} < L_{i2}$ para los restantes valores, las reglas de decisión no son comparables (se puede comprobar fácilmente que - así definido se induce un orden parcial en el espacio de reglas de de

---

Sin embargo en cualquier situación que ocurra en la práctica se debe considerar la posibilidad de que un individuo observado no pertenez- ca a alguno de los grupos especificados como cuando un biólogo descu bre a un miembro de una nueva especie. Rao (1973) menciona que pa rece que no existe aún una teoría general que incluya tal posibilidad.

cisión). Por lo tanto, se hace necesario introducir un criterio adicional para escoger entre las reglas de decisión, tal criterio esta relacionado con el concepto de reglas de decisión admisibles.

Una regla de decisión C es admisible si no existe otra regla la cual sea mejor que C. Obviamente, si la clase de reglas de decisión admisibles cuenta con un solo miembro se tiene una solución óptima. Sin embargo la clase de decisiones admisibles es amplia lo que motiva la necesidad de otros criterios para seleccionar entre ellas. En el caso de que se conozcan las probabilidades a priori π_i $i=1, \ldots k$ donde

$$\pi_i = P_r \left[\underline{x} \text{ provenga de } P_i \right]$$

el problema admite una solución sencilla pues la pérdida esperada total queda como sigue:

$$L = \pi_1 L_1 + \ldots + \pi_k L_k$$

L_i definidas anteriormente. Por lo tanto para una regla de decisión no aleatorizada, L se puede expresar como:

$$L = \sum_{i=1}^{k} \int_{R_i} (\pi_1 c_{1i} f_i + \ldots + \pi_k c_{ki} f_k) \, dv$$

$$= \int_{R_1} -S_1 \, d \quad + \ldots + \int_{R_k} -S_k \, dv$$

donde $S_i = -(\pi_1 c_{1i} f_1 + \ldots + \pi_k c_{ki} f_k)$ se conoce como el resultado discriminante de un individuo de la i-ésima población. Para una regla de decisión aleatorizada la pérdida esperada total es

$$L = \sum_{i=1}^{k} \pi_i \int_R (c_{i1} \lambda_1(\underline{x}) + \ldots + c_{ik} \lambda_k(\underline{x})) f_i \, dv$$

rearreglando los términos

$$L = \int - \left[\sum_{i=1}^{k} \lambda_i(\underline{x}) S_i(\underline{x}) \right] dv$$

Para encontrar la solución para la regla de decisión no aleatorizada se utilizará el siguiente:

Lema. Sea $G_1, \ldots, G_k$ una división del espacio G en k regiones mutuamente exclusivas. Sea $G_1^*, \ldots, G_k^*$ una división en k regiones mutuamente exclusivas tales que

$$x \in G_i^* \Rightarrow f_i(x) \geq f_j(x) \qquad j = 1, \ldots k$$

con $f_1, \ldots, f_k$ integrables sobre G con respecto a una medida v. entonces

$$\int_{G_1^*} f_1 dv + \ldots + \int_{G_k^*} f_k \, dv \geq \int_{G_1} f_1 dv + \ldots + \int_{G_k} f_k \, dv$$

Aplicando este resultado para encontrar la solución óptima mediante una regla de decisión no aleatorizada se obtiene la partición R_1^*

..., R_k^*

tal que

$$\underline{x} \in R_i^* \Rightarrow \underset{i=1,\ldots,k}{S_i(\underline{x}) \geq} S_j(\underline{x}) \quad \forall_j \ldots \qquad \text{I.3.1}$$

identificando S_i con f_i del lema, R_i^* con G_i^* $i=1, \ldots, k$ se tiene que el valor de $-L$ se maximiza, i.e., el de L se minimiza escogiendo como en I.3.1. a R_i^*. Queda por especificar como actuar en caso de que en I.3.1 se tenga una igualdad para dos o más índices, supóngase por ejemplo que para una $\underline{x}$ se tiene

$$S_{i1}(\underline{x}) = \ldots = S_{im}(\underline{x}) \geq S_{i_{m+1}}(\underline{x}) \geq \ldots \geq S_{i_k}(\underline{x})$$

entonces se puede asignar $\underline{x}$ a cualquiera de las regiones $R_i^*, \ldots R_{i_m}^*$ de una manera arbitraria.

La partición $R_1^*, \ldots, R_k^*$ coincide con la inducida mediante el razonamiento seguido en Anderson (1958) pues éste último propone minimizar la pérdida esperada de asignar $\underline{x}$ a P_j en términos de la función de probabilidades a posteriori

$$h(P_i \mid \underline{x}) = \frac{\pi_i f_i(\underline{x})}{\sum_{L=1}^{k} \pi_L f_L(\underline{x})}$$

obteniendo que la pérdida esperada de asignar $\underline{x}$ a P_j se puede expresar como

$$\sum_{\substack{i=1 \\ i\neq j}}^{k} h\,(P_i \mid \underline{x})\, c_{ij} = \sum_{\substack{i=1 \\ i\neq j}}^{k} \frac{\pi_i f_i(\underline{x})}{\sum_{L=1}^{k} \pi_L f_L(\underline{x})}$$

$$= \frac{1}{M} \sum_{\substack{i=1 \\ i\neq j}}^{k} \pi_i f_i(\underline{x})\, c_{ij} \quad \ldots \text{ I.3.2.}$$

y en base a esto asignar $\underline{x}$ a $P_{j'}$ tal que para $j = j'$ se minimice I.3.2., pero minimizar (2) es equivalente a minimizar $\sum_{\substack{i=1 \\ i\neq j}}^{k} \pi_i f_i(\underline{x})\, c_{ij}$

$$\text{i.e.} \quad \min_j \quad \pi_1 f_1 c_{ij} + \ldots + \pi_k f_k c_{kj}$$

que a su vez es equivalente a

$$\max_j - (\pi_1 f_1 c_{ij} + \ldots + \pi_k f_k c_{kj}) = \max_j S_j(\underline{x})$$

y entonces las regiones de discriminación coinciden, por lo que las soluciones son equivalentes.

En caso de utilizar una regla de decisión aleatorizada se utilizará el resultado del siguiente:

Lema. Sean $f_1, \ldots, f_k$ funciones integrables sobre S con respecto a una medida v y $\phi_i(x) \geq 0 \quad i = 1,..,k$ tal que $(\phi_1 + \ldots + \phi_k) = 1$. Considerar la elección $\phi_{i_{r+1}}^* = \ldots = \phi_{i_k}^* = 0$ y $\phi_{i_1}^*, \ldots, \phi_{i_r}^*$ no negativos pero arbitrarios sujetos a $\phi_1^* + \ldots + \phi_k^* = 1$, cuando $f_i = \ldots = f_{i_r} \geq f_{i_{r+1}} \geq f_{i_{r+2}} \geq \ldots \geq f_{i_k}$, entonces

$$\sum_{1}^{k} \int f_i \, O_i^* \, dv \quad \geq \quad \sum_{1}^{k} \int f_i \, O_i \, dv$$

Identifíquense $S_m(\underline{x}) = f_m$, $m = 1, \ldots k$ y $\lambda_m(\underline{x}) = \phi^*(x)$, en el caso general de que $S_{i1}(\underline{x}) = \ldots = S_{ir}(x) > S_{ir+1}(\underline{x}) \geq \ldots \geq S_{ik}(x)$ por el resultado del lema eligiendo $\lambda_{i1}(\underline{x}) + \ldots + \lambda_{ir}(\underline{x}) = 1$ y $\lambda_{ir+1}(\underline{x}) = \ldots = \lambda_{ik}(\underline{x}) = 0$ se obtiene una solución óptima para la regla de decisión aleatorizada, en particular si $S_i > S_j \ \forall \ j = 1, \ldots, m$ la regla de decisión aleatorizada óptima queda como sigue: $\lambda_i(\underline{x}) = 1$ y $\lambda_j(\underline{x}) = 0 \ \forall_j$.

De lo anterior se puede observar que las soluciones óptimas para las reglas de decisión aleatorizada y no aleatorizada son esencialmente las mismas. A la decisión aleatorizada definida anteriormente se le llama regla de decisión de Bayes con respecto a la distribución a priori π_i $i = 1, \ldots, k$. Si se denota al vector de pérdida de tal decisión de Bayes mediante $L'_{\pi} = (L_{1\pi}, \ldots, L_{k\pi})$, se puede caracterizar a la decisión de Bayes como aquella tal que, para un vector $\pi = (\pi_1, \ldots, \pi_k)$ de probabilidades a priori dado,

$$\pi' L_{\pi} \leq \pi' L$$

donde L es el vector de pérdida asociado a cualquier otra decisión.

Ahora bien, se debe de considerar un caso que ocurre frecuen

temente en la práctica, cuando se desconocen las probabilidades a priori o estas no son relevantes. Para tratar este problema, se debe de considerar la clase mínima de reglas de decisión para, de tal conjunto, elegir aquella que solucione el problema con algún otro criterio. Con este fin se debe de considerar primero una clase de reglas de decisión completa, se dice que una clase de reglas de decisión C es completa si para cualquier regla de decisión δ fuera de C existe $\delta' \in C$ tal que δ' es mejor o igual que δ. Una clase completa la cual no contiene una subclase que a su vez sea completa se dice que es una clase completa mínima.

Una clase completa mínima puede ser caracterizada (si se desea ver la prueba de las proposiciones ver Rao (1973) y Wald (1950)) como sigue:

i) Una condición necesaria y suficiente para la existencia de una clase completa mínima es que la clase de reglas de decisión admisibles sea completa.

ii) La clase A de todas las reglas de decisión admisibles es una clase completa mínima.

iii) Toda regla de decisión admisible es una regla de Bayes con res

pecto a alguna distribución a priori.

iv) La clase B de todas las reglas de Bayes es una clase completa (puede o no ser mínima).

v) La regla de Bayes con respecto a cualquier vector de probabilidades a priori $\pi = (\pi_1, \ldots, \pi_k)$ es admisible si todos los componentes de π son positivos.

Una posible elección entre las reglas de decisión admisibles es la llamada minimax, que es aquella regla de decisión δ^* tal que

$$\max_{1 \leq i \leq k} (L_1^*, \ldots, L_k^*) = \min_{\delta \in \Delta} \max_{1 \leq i \leq k} (L_1, \ldots, L_k)$$

Δ el espacio de reglas admisibles de decisión, que se puede pensar como aquella que elegiría un estadístico pesimista.

Para la identificación de la regla minimax puede servir el siguiente:

Teorema. Si existe un vector de probabilidades $\bar{\pi}' = (\bar{\pi}_1, \ldots, \bar{\pi}_k)$ tal que para la correspondiente solución de Bayes los componentes del vector de pérdida son todos iguales, entonces $\bar{\pi}$ es llamada la distribución a priori menos favorable y la regla de decisión de Bayes aso-

ciada a $\overline{\pi}$ es una regla minimax.

En este caso si se tiene un procedimiento de clasificación para el cual el vector $L = k\,(1, \ldots, 1)$, k una constante, entonces tal procedimiento es la regla de decisión minimax i.e, minimiza la máxima pérdida posible.

Se ha planteado el problema de discriminación en términos de funciones de decisión estadística, proponiéndose las soluciones óptimas en caso de conocer las probabilidades a priori y ciertas soluciones posibles si tales probabilidades se desconocen.

Tal enfoque, aunque general y analíticamente elegante, tiene la desventaja de que en ciertos casos la función de pérdida es difícil (y aún prácticamente imposible) de establecer por la dificultad de asignar valores de pérdida en situaciones tales como la identificación de un fósil, la clasificación de un candidato a ingresar a una universidad, etc. Además en general las poblaciones no se conocen completamente, por lo que las densidades de probabilidad $f_1, \ldots, f_k$ no están totalmente especificadas, debido a esto es necesario optar por caminos alternativos pues las soluciones anteriores, aunque aplicables, ya no son óptimas. En la siguiente sección se presentan las situaciones generales del problema y una breve exposición de soluciones bajo diferentes enfoques.

I.4 Diferentes formulaciones y breve idea de las soluciones.

S1) Considérense k diferentes poblaciones $P_1, \ldots, P_k$ y una muestra aleatoria (de tamaño $n \geq 1$ ó una sucesión en el caso secuencial) de unidades experimentales que provienen de una población P_o la cual se sabe coincide con alguna de las poblaciones P_i para alguna $i = 1, \ldots, k$, el problema es determinar la i específica. En general para distinguir las poblaciones un individuo será caracterizado por un conjunto de mediciones o atributos que se indican mediante el vector p-dimensional $\underline{x}' = (x_1, x_2, \ldots, x_p)$ y en el caso de que el individuo pertenezca a la i-ésima población (de las k que se consideran como posibles) $\underline{x}$ tiene una densidad f_i con una distribución bien definida $F_i(\cdot)$.

Un gran número de los desarrollos supone que cada $F_i(\cdot)$ es una distribución normal multidimensional[1], aunque esta suposición no

1.- Como se recordará la distribución normal es fundamental en el análisis de variables unidimensionales debido, principalmente, a su papel como distribución límite de sumas de variables aleatorias independientes e idénticamente distribuídas con segundo momento finito (caso especial del teorema central del límite). El análogo multidimensional justifica la importancia de la distribución normal multivariada.

es esencial en el desarrollo de la teoría, simplifica enormemente los cálculos y es útil en el estudio de las propiedades de los procedimientos cuando se desconocen los parámetros de las distribuciones.

Para resolver esta situación se necesita desarrollar una regla tal que después de observar un vector $\underline{x}$ se pueda decidir en qué población clasificar al individuo que tiene asociado tal vector.

S2) Otra posible formulación se tiene en el caso de considerar a P_o como una mezcla de las poblaciones $P_1, \ldots, P_k$. De nuevo un individuo estará caracterizado por un vector $\underline{x}$ de mediciones. Se considera entonces una variable aleatoria I que toma valores 1, 2, ..., k que indica a que población pertenece el individuo a clasificar, i.e. - la población a la que pertenece $\underline{x}$ dado que $I = i$ es P_i.

Como es de suponer para los individuos a clasificar el valor de I no es observable y el problema es decidir sobre el valor de I en base al conocimiento del vector $\underline{x}$. El problema se denominará como "mezcla conocida" o "mezcla desconocida" de acuerdo a si se conoce o desconoce la distribución de I sobre el conjunto $\{1, \ldots, k\}$.

Así planteado el problema presenta variantes ya que la distri

bución de (x, I) se desconoce y para obtener información acerca de - ella se toma una muestra de "ensayo, de tamaño, $m \geq 1$, de P_o. - Tal muestra de ensayo puede ser de tres tipos:

i) "Identificada" o "supervisada" donde para cada elemento muestral $\underline{x}$ e I son observables.

ii) "No identificada" o "no supervisada" donde para cada elemento - muestral solo $\underline{x}$ es observable.

iii) "Post-identificada" o "Post-supervisada", que es el caso cuando se tiene una sucesión de individuos a ser clasificados y después de clasificar al j-ésimo individuo su correspondiente valor asociado de I se conoce exactamente. Entonces para la clasificación del n-ésimo individuo los n-1 individuos previamente clasificados forman una muestra ensayo identificada.

La situación S2 tiene una generalización, la cual es considerar a I en vez de variable artificial, una variable contínua o discreta con significado físico y la población P_i corresponde a $I \in S_i$ donde $\{S_i\}_{i=1}^{k}$ es una partición del espacio de valores de I. Marshall y Olkin tratan un problema bajo esta situación.

Las soluciones "clásicas" conducen a efectuar una partición del

espacio de observaciones posibles en k regiones ajenas E_i $i = 1, \ldots$ k, tales que, si un individuo está caracterizado por un vector de mediciones $\underline{x}$ y $\underline{x} \in E_j$, entonces es clasificado en la población j-ésima. Rao (1950, 1951, 1952) discute de una manera heurística la introducción de regiones de duda, donde se tiene la posibilidad de tomar una decisión provisional y obtener mayor información y también decisiones preferenciales (cuando se deben efectuar las asignaciones en una cierta proporción) además de las k acciones naturales.

Las soluciones no paramétricas se agrupan en tres grandes categorías:

i) Partir de una buena regla (en el sentido de toma de decisiones) suponiendo que las distribuciones se conocen. Reemplazar las funciones de distribución o probabilidad por sus respectivos estimadores muestrales. Por esta característica, tales reglas son conocidas como de inserción.

ii) Usar estadísticas surgidas en el estudio de pruebas para comparar dos muestras o k-muestras.

iii) Trabajar con métodos específicos para el problema de discriminación, como por ejemplo: reglas de "distancia" mínima.

Las soluciones bayesianas al problema parecen ser más simples para ciertas densidades a priori. Estas se basan esencialmente en el cálculo de las probabilidades a posteriori de que el individuo con vector de mediciones $\underline{x}$ pertenezca o provenga de la población P_i, las cuales resultan proporcionales al producto de las probabilidades a priori por la verosimilitud de la observación dados los estimadores (o en caso de conocerlos los parámetros) basados en la muestra inicial o de ensayo de la población P_i.

En el siguiente capítulo se presentan con detalle las soluciones.

I.5 Nota Histórica.

Como menciona Kendall (1957), sin duda la idea de discriminar entre poblaciones podría ser rastreada muy atrás en el pasado (tal vez no de modo explícito semejante a los tratamientos actuales pero la idea subyacente si era la misma). Sin embargo, para marcar un inicio se puede considerar que éste ocurre con el trabajo de Karl Pearson alrededor de 1920 y cuyo objetivo era encontrar un coeficiente el cual "midiese la distancia", en algún sentido aceptable, entre dos poblaciones basándose en datos antropométricos. El primer trabajo publicado sobre el coeficiente de Pearson de semejanza racial, denotado por C^2, fue el de M. L. Tildesley (1921) sobre cráneos de Birmania.

Alrededor de la misma época P. C. Mahalanobis se interesó en la materia y llegó a la conclusión de que la solución propuesta por Pearson no era adecuada (el valor de C^2 depende en gran medida del

tamaño muestral, etc.). Mahalanobis presentó una medida alternativa llamada por él D^2, la cual usó en 1925 para discutir las mezclas raciales en Bengala. Kendall considera que este trabajo visionario fue el punto inicial de investigación de la "robusta" escuela hindú.

En 1926, Pearson publicó el primer trabajo teórico acerca del coeficiente C^2 donde sugirió una forma para calcular el coeficiente cuando las variables son dependientes.

Entre 1927 y 1930 Mahalanobis y Pearson mantuvieron su controversia. Mahalanobis (1930) continuó su investigación práctica y teórica sobre D^2. Tal controversia continuó hasta la muerte de K. Pearson en 1936. Más tarde se encontraría la relación existente entre la D^2 de Mahalanobis y la T^2 de Hotelling propuesta por éste en 1931.

Parece que existe un receso en la investigación alrededor del problema y es hasta 1936 en que el gran R. A. Fisher publicó su primer artículo sobre funciones discriminantes el cual se originó debido a que E. S. Martín (1936), deseaba clasificar huesos de la mandíbula recuperados de un sepulcro, como pertenecientes a individuos de sexo masculino o femenino. Para resolver el problema Fisher sugirió calcular una función lineal de las mediciones hechas al hueso de la mandíbu

la la cual ofreciese una separación máxima entre las distribuciones de las mediciones para los dos sexos en términos del cociente de la diferencia en los valores medios entre la desviación estandar común, supuesta para simplificar el problema , y resolvió el problema de prueba de significancia. Hay que hacer notar que todo el trabajo desarrollado hasta aquí se enfocaba a diferenciar entre dos poblaciones simplemente. La diferencia principal entre el enfoque de Mahalanobis y el de Fisher es que mientras Mahalanobis estaba interesado en medir distancia entre dos poblaciones, a Fisher sólo le interesaba dividir el espacio muestral en dos regiones y asignar un valor muestral a una población o a otra de acuerdo a en cual región quedaba tal valor muestral, éste último enfoque está muy cercano al de teoría de decisiones que se discutirá en el capítulo II de este trabajo.

Bose (1936), Bose y Roy (1938), y el mismo Fisher (1938), trabajaron sobre la D^2 de Mahalanobis. Welch (1939) unifica la teoría de funciones discriminantes y la de pruebas estadísticas de una manera simple. Este último demostró que la solución en base a tales funciones es equivalente a la obtenida mediante el cociente de verosimilitudes utilizado para probar que las mediciones en las que se basa el criterio pertenecen a un miembro de una de dos poblaciones propuestas.

Después de ésto, el tratamiento del problema estadísticamente sufre una interrupción en cuanto a la continuidad en la investigación - (hay que recordar que se presentaron años de guerra), que se debe además a un cambio de intereses de investigación de la escuela inglesa. - Sin embargo, en los Estados Unidos la idea había empezado a interesar a estadísticos como Wald (1944) que propuso una solución muy interesante, Cochran (1943) el cual toca un punto sumamente atrayente, Von Mises (1945) que encuentra la solución para clasificar en un caso teórico a un individuo entre k poblaciones, Cochran y Bliss (1948) continuan la investigación, generalizando el problema para la clasificación en dos poblaciones mismo que es tratado por Anderson (1951).

S. N. Roy y principalmente C. R. Rao (1946 a 1948, 1949, 1950) llevaron a cabo otros desarrollos teóricos, Kendall (1951) señala atinadamente que Rao hizo algunas sugerencias sobre el problema general de medir distancias, aunque tal camino no parece haber sido seguido, posiblemente porque el análisis tensorial y la geometría diferencial son campos no muy familiares para la mayoría de los estadísticos matemáticos.

Rao (1952) presentó la extensión a más de dos alternativas y la aplicación práctica de los resultados. La importancia de la función

discriminante se basa en un teorema debido a C. A. B. Smith (1947) que establece su suficiencia con respecto a las dos poblaciones alternativas bajo consideración, así que ninguna pérdida de información resulta de la reducción de las mediciones múltiples a una función discriminante lineal individual. Rao (1942) generalizó el resultado de Smith para establecer la suficiencia de la función discriminante derivada de dos poblaciones para el conjunto más amplio de poblaciones alternativas definidas por las medias vectoriales sobre el segmento de línea que une las posiciones de las medias de las dos poblaciones originales. Este resultado es útil en la prueba de la propiedad (o suficiencia) de la función discriminante en la clasificación de un individuo cuando se admite la posibilidad de que pertenezca a un tercer grupo desconocido.

Rao (1966, 1973) presenta una posible generalización al problema originada por la necesidad de determinar si el Australopithecus africanus, un fósil descubierto en el desfiladero de Olduvai (Africa) por Leakey, era más homínido que antropoide. Tal determinación se basó en la comparación de las mediciones en el A. africanus, el vector $\underline{x}$, con aquellas de grupos conocidos de fósiles de homínidos y de antropoides para identificar sus afinidades.

En esta situación la suposición de que el fósil observado es un

miembro de uno de dos grupos conocidos de fósiles no es plausible, se debe de tener en cuenta la posibilidad de que el fósil puede pertenecer a un grupo desconocido cuya existencia no ha sido establecida todavía. Así también, aunque es posible obtener estimadores burdos de las medias y matrices de covarianza de las mediciones para cada grupo conocido, no parece existir ningún método para obtener las probabilidades a priori pues los datos referentes al número de fósiles en los diferentes grupos no son necesariamente buenos indicadores de la abundancia relativa en la población de fósiles. Además no hay posibilidad de asignar probabilidades a priori a los grupos no descubiertos.

La situación así planteada, requiere una solución especial, pues las propuestas para el problema no son aplicables en este caso. Este es un problema que requiere mayor investigación aunque Rao (1973) propone una solución bajo ciertas restricciones la cual se discutirá más adelante.

B I B L I O G R A F I A

CAPITULO I.-

I.1

Rao (1948), (1965), (1966), (1973)

Durand (1941)

Travers (1938), (1939)

Tintner (1946)

Baten y Dewitt (1944)

Fisher (1936)

Hamilton (1950)

Cox y Martin (1937)

Barnaby (1966)

Bartlett y Please (1963)

I.3

Rao (1973)

I.4

Rao (1950), (1951), (1952)

DasGupta (1973)

I.5

Kendall (1957)

Tildlesley (1921)

Mahalanobis (1930)

Martin (1936)

Fisher (1936), (1938)

Bose (1936)

Bose y

Roy (1938)

Welch (1939)

Wald (1944)

Von Mises (1945)

Cochran y

Bliss (1948)

Anderson (1951), (1958)

Rao (1946), (1948), (1949), (1950),

(1962), (1966), (1973)

Smith (1947)

II. ALGUNAS SOLUCIONES PROPUESTAS.-

En este capítulo se presentan algunas soluciones al problema. La intención es observar la evolución que han tenido las mismas, así como presentar alternativas para una situación específica.

II.1 Soluciones clásicas para poblaciones normales.

II.1.1 El caso de dos poblaciones con homoscedasticidad.

Se tratará inicialmente el caso de discriminación para dos poblaciones. La primera solución que se presenta es la propuesta por Fisher (1936), cuyo planteamiento es el siguiente:

Supóngase que se tiene un vector de observaciones $\underline{x}' = (x_1, \ldots, x_p)$ el cual se sabe que pertenece a una de las poblaciones P_1, P_2, que se distribuyen mediante una normal multidimensional con medias $\underline{\mu}_1$ y $\underline{\mu}_2$ respectivamente y la misma matriz de covarianza Σ. Se propone resolver el problema de discriminación en base a una función lineal de las componentes de $\underline{x}$ digamos $\underline{\lambda}' \underline{x}$, y dependiendo del valor que tome esta función considerar la observación como proveniente de la población P_1 o de la población P_2. El vector $\underline{\lambda}$ debe tener características óptimas en el sentido de que maximice el cociente entre la diferen

cia absoluta de las medias muestrales de los valores de $\underline{\lambda}' \underline{x}$ y la desviación estandar dentro de las muestras de los valores de $\underline{\lambda}' \underline{x}$, i.e., se desea encontrar $\underline{\lambda}$ tal que

$$\frac{|\text{Media de } \underline{\lambda}' \underline{x} \text{ en } P_1 - \text{Media de } \underline{\lambda}' \underline{x} \text{ en } P_2|}{\text{Desviación estándar de } \underline{\lambda}' x \text{ dentro de poblaciones}} \qquad \text{(II.1)}$$

se maximice. El procedimiento así planteado tiene la ventaja de disminuir la dimensión del problema, de uno p-dimensional a otro uni-dimensional.

Analíticamente el valor óptimo de $\underline{\lambda}$ se encuentra al maximizar

$$\frac{|\underline{\lambda}' \underline{\mu}_1 - \underline{\lambda}' \underline{\mu}_2|}{(\underline{\lambda}' \Sigma \underline{\lambda})^2} = \frac{|\underline{\lambda}' \underline{\delta}|}{(\underline{\lambda}' \Sigma \underline{\lambda})^{1/2}} \qquad \text{(II.2)}$$

$$\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$$

con respecto a $\underline{\lambda}$. Es más conveniente maximizar el cuadrado de II.2 o sea trabajar con

$$\frac{\underline{\lambda}' \underline{\delta}\, \underline{\delta}' \underline{\lambda}}{\underline{\lambda}' \Sigma \underline{\lambda}} \qquad \text{(II.3)}$$

maximizando el numerador de la expresión II.3 manteniendo el denominador constante. Si se denota por η a un multiplicador de Lagrange

se deberá maximizar

$$L = \underline{\lambda}' \underline{\delta}\, \underline{\delta}' \underline{\lambda} - \eta (\underline{\lambda}' \Sigma \underline{\lambda} - 1)$$

derivando e igualando a cero se obtiene

$$\frac{\partial L}{\partial \underline{\lambda}} = \underline{\delta}\underline{\delta}' \underline{\lambda} - \eta \Sigma \underline{\lambda} = 0 \qquad \text{(II.4)}$$

$$\frac{\partial L}{\partial \eta} = \underline{\lambda}' \Sigma \underline{\lambda} - 1 = 0 \qquad \text{(II.5)}$$

del conjunto de ecuaciones II.4, premultiplicando por $\underline{\lambda}'$ se tiene

$$\underline{\lambda}' \underline{\delta}\, \underline{\delta}' \underline{\lambda} = \eta$$

i.e. el valor del multiplicador es igual al del numerador de II.3, como II.4 es equivalente a

$$(\underline{\delta}' \underline{\delta} - \eta \Sigma) \underline{\lambda} = 0 \qquad \text{(II.6)}$$

que tiene solución distinta de la trivial sí y sólo si $| \underline{\delta}\underline{\delta}' - \Sigma | = 0$ de donde se obtiene que η es una raíz característica de $\underline{\delta}\underline{\delta}'$ en la métrica de Σ y $\underline{\lambda}$ es su vector característico asociado, por lo tanto para maximizar L hay que escoger como valor de η la máxima raíz de $| \underline{\delta}\, \underline{\delta}' - \eta \Sigma | = 0$ y encontrar $\underline{\lambda}$, su vector asociado. Puede observarse que $\tau = \underline{\delta}' \underline{\lambda}$ es un escalar por lo que II.4 puede escribirse como

$$\Sigma \underline{\lambda} = \underline{\delta}\,\underline{\delta}'\underline{\lambda} = \underline{\delta}\, c$$

y de aquí $$\Sigma \underline{\lambda} = \delta c$$

$$\Rightarrow \quad \frac{\eta}{c}\Sigma \underline{\lambda} = \underline{\delta}$$

i.e. la solución es proporcional a $\underline{\delta}$, y finalmente se obtiene que*

$$\underline{\lambda} = \frac{c}{\eta}\Sigma^{-1}\underline{\delta}$$

obteniendo la función discriminante lineal de Fisher como

$$\underline{\lambda}'\underline{x} = \frac{c}{\eta}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x} \qquad \text{(II.7)}$$

En caso de desconocer los parámetros $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ se propone substituir sus valores por el de sus respectivos estimadores muestrales $\bar{\underline{x}}_1$, $\bar{\underline{x}}_2$ y S . Aunque no se sabe exactamente que efectos tenga esto, se espera que para muestras grandes las alteraciones no sean grandes.

Wald (1944), en relación al mismo problema atacado por Fisher y considerando que se tiene una muestra aleatoria de N_1 unidades pertenecientes a la población P_1 y otra muestra aleatoria independiente de la anterior que consta de N_2 unidades pertenecientes a la población P_2, presenta una solución que se deriva de la prueba de la hi

* recordar que Σ es definida positiva.

pótesis H_1, que la observación a discriminar, $\underline{x}$, fue extraída de P_1 frente a la alternativa H_2 de que fue extraída de P_2.

Si se conociesen $\underline{\mu}_1$ y $\underline{\mu}_2$ y Σ la región se encontraría fácilmente mediante la aplicación del lema de Neyman - Pearson, obteniéndose el siguiente procedimiento:

Considerar a $\underline{x}$ como proveniente de P_2 si

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \leq k \qquad \text{(II.8)}$$

donde k es tal que $P\left[\frac{f_1}{f_2} \leq k \mid f_1\right] = \alpha$, α nivel de significancia de la prueba y $f_i(\underline{x})$ denota la densidad* para P_i, $i = 1, 2$.

En otro caso considerar a $\underline{x}$ como proveniente de P_1.

De la aplicación del lema en la situación que interesa y efectuando posteriormente algunas transformaciones se obtiene la siguiente desigualdad

$$-\frac{1}{2}\left[(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\right] \leq \ln k \qquad \text{(II.9)}$$

De una manera directa (ver Anderson (1958)) puede demostrarse que II.2 es equivalente a

* la cual se considera que existe

$$\underline{x}' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) - \frac{1}{2} (\underline{\mu}_1 + \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \geq \ln k \quad \text{(II.10)}$$

Wald (1944) trabaja con la desigualdad

$$\underline{x}' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \geq k' \quad \text{(II.11)}$$

que puede obtenerse fácilmente de (II.10). En vista de que el valor de los parámetros $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ se desconoce, Wald (1944) propone usar la estadística $U = \underline{x}' S^{-1} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$ para llevar a cabo la discrimina - ción.

Como puede observarse U esta ligada íntimamente con la fun ción discriminante lineal de Fisher expresada en II.7, pues si se con sidera el transpuesto de esta última, que por ser un número real no sufre alteración en su valor, se tiene que

$$\underline{x}' \underline{\lambda} = \frac{2}{\eta} \underline{x}' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

de donde se puede observar que U es proporcional a la función discri minante lineal de Fisher. Dado que se desconocen $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ pa rece intuitivamente razonable usar en vez de los parámetros los co - rrespondientes estimadores insesgados y consistentes.

$$\hat{\underline{\mu}}_1 = \bar{\underline{x}}_1 \quad , \quad \hat{\underline{\mu}}_2 = \bar{\underline{x}}_2 \qquad \text{y}$$

$$\hat{\Sigma} = S = \frac{1}{N_1 + N_2 - 2} \left[\sum_{j=1}^{N_1} (\underline{x}_{1j} - \bar{\underline{x}}_1)(\underline{x}_{1j} - \bar{\underline{x}}_1)' + \sum_{j=1}^{N_2} (\underline{x}_{2j} - \bar{\underline{x}}_2)(\underline{x}_{2j} - \bar{\underline{x}}_2)' \right]$$

donde $\bar{x}_i = \frac{1}{N_j} \sum_{j=1}^{N_i} x_{ij}$ $\quad i = 1, 2$

por lo que, substituyendo en la expresión de U

$$U = x' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

Debido a su carácter aleatorio es preciso analizar la distribución de la estadística U. Wald (1944) presenta la solución en el caso en que N_1 y N_2 son "grandes". En esta situación y como los estimadores substituídos en vez de los parámetros convergen estocásticamente a éstos últimos, puede demostrarse que la distribución asintótica de U es normal con media α_i y varianza σ^2 si x proviene de P_i $\quad i = 1, 2$ donde

$$\alpha_i = \mu_i' \Sigma^{-1} (\mu_2 - \mu_1) \quad y \quad \sigma^2 = (\mu_2 - \mu_1)' \Sigma^{-1} (\mu_2 - \mu_1)$$

utilizando lo anterior la región de rechazo para la hipótesis H_i anteriormente será

$$\frac{U - \hat{\alpha}_1}{\hat{\sigma}} \geq d$$

que bajo H_1 se distribuye asintóticamente como N (0, 1) si N_1 , N_2 son grandes, donde d se escoge tal que $\int_d^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \gamma$

y γ es el tamaño de la región crítica.

En caso de que se opte por utilizar la solución minimax (ver sección I.2 de este trabajo) d ha de escogerse de manera que

$$c_{12}\ \frac{1}{\sqrt{2\pi}}\int_{d}^{\infty} e^{-z^2/2}\ dz = c_{21}\ \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{((\hat{a}_1 - \hat{a}_2)/\hat{\sigma}\ +\ d)} e^{-z^2/2}\ dz$$

c_{12} , c_{21} las pérdidas asociadas a los resultados equivocados.

Sin embargo, si N_1 y N_2 no son grandes se necesita derivar la distribución muestral exacta de U lo cual se trata de manera general más adelante.

En relación al mismo problema Anderson (1951), sigue un procedimiento semejante al de Wald (1944) pero propone usar la estadística

$$W = \underline{x}'\ S^{-1}\ (\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - \frac{1}{2}\ (\bar{\underline{x}}_1 + \bar{\underline{x}}_2)'\ \Sigma^{-1}\ (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

que puede observarse que es el primer miembro completo de la desigualdad (II.10).

El mismo Anderson (1951), trata la distribución de tal estadística bajo la suposición de que las medias $\underline{\mu}_1$ y $\underline{\mu}_2$ son proporcio

nales. Sitgreaves (1952) trata la distribución general de una matriz aleatoria simétrica, M_{2x2}, con la característica de que las estadísticas U de Wald y W de Anderson pueden ser escritas como funciones de los elementos de M.

Como Wald y Sitgreaves indican la distribución muestral de U, está contenida como caso particular de la distribución muestral de la variable

$$V = k\, Y_1\, A^{-1}\, Y_2 \qquad \text{(notación de Sitgreaves)}$$

k una constante conocida, Y_1 y Y_2 variables aleatorias p - dimensionales con vectores de medias ζ y ξ respectivamente y A_{pxp} una matriz simétrica con distribución Wishart con $n = N_1 + N_2 - 2$ grados de libertad y los tres conjuntos de variables distribuídas independientemente con la misma matriz de covarianza.

La estadística W puede expresarse como:

$$W = \left[\underline{x} - (1/N_1 + N_2)\,(N_1\, \underline{\bar{x}}_1 + N_2\, \underline{\bar{x}}_2) \right]' S^{-1} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$
$$+ \left[(N_1 - N_2)/2\,(N_1 + N_2) \right] (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)'\, S^{-1} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

y la distribución muestral de W, ya sea que $\underline{x}$ provenga de P_1 o de P_2,

es un caso particular de

$$W^* = aY_1' A^{-1} Y_2 + bY_1' A^{-1} Y_2$$

donde a y b son escalares conocidos, Y_1 , Y_2 y A definidos como antes.

Wald (1944) llega a expresar la distribución muestral de V como una función de 3 variables m_1 , m_2 y m_3 que Sitgreaves denota como m_{11} , m_{22}, m_{12} elementos de la matriz simétrica M_{2x2}, donde

$$M = Y' B Y$$

$$Y = (Y_1 , Y_2)^{(*)} \qquad y \qquad B = A + Y Y'$$

en términos de los elementos de M queda como

$$V = k\, m_{12} \left[(1 - m_{11})(1 - m_{22}) - m_{12}^2 \right]^{-1}$$

Sitgreaves presenta una derivación analítica de la distribución de M en el caso en que ζ y ξ sean proporcionales a $(\underline{\mu}_1 - \underline{\mu}_2)$, obteniendo la constante de la distribución y además la distribución de la matriz asociada $M^* = Y' A^{-1} Y$, proporcionando la función de densidad de M que es:

(*) Notar que Y es una matriz de dimensiones px2

$$f(M) = \frac{\Gamma\left(\frac{n+1}{2}\right) e^{-\frac{1}{2}\lambda^2(k_1^2+k_2^2)} |M|^{\frac{1}{2}(p-3)} |I-M|^{\frac{1}{2}(n-p-1)}}{\Gamma\left(\frac{n-p+2}{2}\right) \Gamma\left(\frac{n-p+1}{2}\right) \Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\frac{1}{2}\right)}$$

$$\times \sum_{j=0}^{\infty} \frac{\Gamma\left(\frac{n+2}{2}+j\right)}{\Gamma\left(\frac{1}{2}p+j\right)\Gamma(j+1)} \left(\tfrac{1}{2}\lambda^2\right)^j \left(k_1^2 m_{11} + 2k_1k_2 m_{12} + k_2^2 m_{22}\right)^j$$

con $0 \leq m_{11} \leq 1$, $0 \leq m_{22} \leq 1$ $\quad |M| \geq 0 \quad |I-M| \geq 0$

y la distribución de M^* es tal que su función de densidad es:

$$f(M^*) = \frac{\Gamma\left(\frac{n+1}{2}\right) e^{-\frac{1}{2}\lambda^2(k_1^2+k_2^2)} |M^*|^{\frac{1}{2}(p-3)}}{\Gamma\left(\frac{n-p+2}{2}\right) \Gamma\left(\frac{n-p+1}{2}\right) \Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\frac{1}{2}\right)}$$

$$\times \sum_{j=0}^{\infty} \left\{ \frac{\Gamma\left(\frac{1}{2}(n+2)+j\right)}{\Gamma\left(\frac{1}{2}p+j\right)\Gamma(j+1)} \left(\tfrac{1}{2}\lambda^2\right)^j \right.$$

$$\left. \times \frac{k_1^2 m_{11}^* + 2k_1k_2 m_{12}^* + k_2 m_{22}^* + (k_1^2+k_2^2)|M^*|]^j}{|I+M^*|^{\frac{1}{2}(n+2)j}} \right\}$$

Las distribuciones de M y de M^* son útiles debido al interés en las distribuciones exactas de las estadísticas U y W, ya que -

$M^* = Y' A^{-1} Y$ efectuando este producto se obtiene:

$$M^* = \begin{pmatrix} Y_1' \\ \hdashline Y_2' \end{pmatrix} A^{-1} \left(Y_1 \vdots Y_2 \right) = \begin{bmatrix} Y_1' A^{-1} Y_1 & Y_1' A^{-1} Y_2 \\ Y_1' A^{-1} Y_2 & Y_2' A^{-1} Y_2 \end{bmatrix}$$

$$= \begin{bmatrix} m_{11}^* & m_{12}^* \\ m_{12}^* & m_{22}^* \end{bmatrix}$$

de donde, como $V = k\, Y_1' A^{-1} Y_2$ se tiene que $V \propto m_{12}^*$ por lo que se puede observar que la distribución de V estará relacionada con la de m_{12}^*, así también, como $W^* = aY_1' A^{-1} Y_2 + bY_1' A^{-1} Y_2$ la distribución de W^* puede obtenerse del estudio de $am_{12}^* + bm_{22}^*$. Esto hace interesante el estudio de la distribución de la matriz M^*.

Kabe (1963) trata el mismo problema, pero sin suponer la proporcionalidad de ζ y ξ con respecto a $\underline{\delta} = (\underline{\mu}_1 - \underline{\mu}_2)$, obteniendo los siguientes resultados:

$$f(M) = C^{-1} \left| I - M \right|^{(n-p-1)/2} |M|^{(p-3)/2} G(\phi_1^2, \phi_2^2) \qquad \text{II.12}$$

donde

$$G_1(k_1, k_2) = \sum_{\alpha,\beta_1,\beta_2=0}^{\infty} \frac{(k_1^2)^{\alpha+\beta_1}(k_2^2)^{\alpha+\beta_2}\,\Gamma\left[\frac{1}{2}(n+2)+\alpha+\beta_1\right]}{2^{2\alpha+\beta_1+\beta_2}\,\alpha!\,\beta_1!\,\beta_2!\,\Gamma\left[\frac{1}{2}(p-1)+\alpha\right]}$$

$$\times \frac{\Gamma\left[\frac{1}{2}(n+2)+\alpha+\beta_2\right]\Gamma\left[\frac{1}{2}(n+1)+\alpha\right]}{\Gamma\left[\frac{1}{2}p+2\alpha+\beta_1+\beta_2\right]\Gamma\left[\frac{1}{2}(n+2)+\alpha\right]} \qquad \text{(II.13)}$$

ϕ_1^2 y ϕ_2^2 son las raíces de la ecuación

$$\left| \psi'\Sigma^{-1}\psi - \lambda M^{-1} \right| = 0 \qquad \text{(II.14)}$$

donde $\psi = (\zeta, \xi)$

$$C = \exp\left\{\frac{1}{2}\,\mathrm{tr}\,\psi\psi'\right\}\Gamma\left(\tfrac{1}{2}\right)\Gamma\left[\tfrac{1}{2}(n+1-p)\right]\Gamma\left[\tfrac{1}{2}(n+2-p)\right]$$

si ζ y ξ son proporcionales como en el caso tratado tanto por Anderson (1951) y Sitgreaves (1952) sea $\zeta = k_1\delta$ y $\xi = k_2\delta$ entonces (II.14) tiene solamente una raíz*

$$\phi_1^2 = (\underline{\delta}'\Sigma^{-1}\underline{\delta})(k_1^2 m_{11} + 2k_1k_2m_{12} + k_2^2 m_{22})$$

m_{ij} el elemento ij de la matriz M, de donde el resultado de Sitgreaves ((1952), ecuación 21) se tiene substituyendo en (II.12) $\alpha = 0$, $\beta_2 = 0$, $\phi_2 = 0$, $\phi_1 = \phi_1$ de (II.13)

---

* lo cual se puede verificar fácilmente substituyendo valores en (II.14)

Se recordará que se obtuvo también la distribución de $M^* = M(I + M)^{-1}$, Kabe a su vez encuentra que

$$p(M^*) = C^{-1} \, |M^*|^{(p-3)/2} \, |I + M|^{-(n+2)/2} \, G(\chi_1^2, \chi_2^2)$$

donde χ_1^2 y χ_2^2 son las raíces de

$$|\psi' \Sigma^{-1} \psi - \lambda(I + M^*) M^{*-1}| = 0$$

En relación al mismo problema puede utilizarse el criterio de cociente de verosimilitudes, el cual puede plantearse de la siguiente manera: considerar que se está probando la hipótesis compuesta H_1 : $\underline{x}_1^{(1)}, \ldots, \underline{x}_{N1}^{(1)}, \underline{x}$ (la observación por discriminar) han sido extraídos de una $N(\underline{\mu}_1, \Sigma)$ y que $\underline{x}_1^{(2)}, \ldots, \underline{x}_{N2}^{(2)}$ han sido extraídos de $N(\underline{\mu}_2, \Sigma)$ contra la alternativa H_2 : $\underline{x}_1^{(1)}, \ldots, \underline{x}_{N1}^{(1)}$ han sido extraídos de $N(\underline{\mu}_1, \Sigma)$ y $\underline{x}_1^{(2)}, \ldots, \underline{x}_{N2}^{(2)}, \underline{x}$ de $N(\underline{\mu}_2, \Sigma)$ con $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ no conocidos. Desarrollando la prueba se obtiene el cociente (para más detalle en esta derivación ver Anderson (1958)).

$$\lambda = \frac{|\hat{\Sigma}_2|^u}{|\hat{\Sigma}_1|^u} = \frac{\left| C + \frac{N_2}{N_2+1} (\underline{x} - \bar{\underline{x}}_2)(\underline{x} - \bar{\underline{x}}_2)' \right|^u}{\left| C + \frac{N_1}{N_1+1} (\underline{x} - \bar{\underline{x}}_1)(\underline{x} - \bar{\underline{x}}_1)' \right|^u} \qquad \text{(II.15)}$$

donde $u = N_1 + N_2 + 1$

$$\hat{\Sigma}_i = \frac{1}{u}\left[C + \frac{N_i}{N_i+1}(\underline{x} - \bar{\underline{x}}_i)(\underline{x} - \bar{\underline{x}}_i)' \right]$$

$$C = \sum_{i=1}^{2} \sum_{\alpha=1}^{N_i} (\underline{x}_\alpha^{(i)} - \bar{\underline{x}}_i)(\underline{x}_\alpha^{(i)} - \bar{\underline{x}}_i)'$$

$\hat{\Sigma}_i$ $i = 1, 2$, es el estimador máximo verosímil de Σ bajo la hipótesis respectiva.

$$\bar{\underline{x}}_i = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \underline{x}_\alpha^{(i)}$$

la expresión (II.15) se puede escribir también como

$$\lambda = \left[\frac{1 + \frac{N_2}{N_2+1}(\underline{x} - \bar{\underline{x}}_2)' C^{-1} (\underline{x} - \bar{\underline{x}}_2)}{1 + \frac{N_1}{N_1+1}(\underline{x} - \bar{\underline{x}}_1)' C^{-1} (\underline{x} - \bar{\underline{x}}_1)} \right]^u \qquad \text{(II.16)}$$

obteniéndose como región de discriminación de $\underline{x}$ como de P_1 todos los puntos para los cuales el cociente λ o una función monótona creciente de él, sea mayor que una constante determinada por el tamaño especificado para la prueba de hipótesis efectuada. Aun cuando este resultado luce muy elegante ya que tanto el numerador como el denominador son transformaciones simples de variables T^2 de Hotelling, sin embar

go, existe el problema de que no son independientes por lo que se sabe poco acerca de la distribución de tal cociente.

En este punto cabe mencionar la interesante analogía que existe entre discriminación y análisis de regresión.

Si se considera la siguiente ecuación de regresión

$$E(\underline{x}) = \alpha + \underline{\beta}\,\xi \tag{II.17}$$

donde

$$\xi = \begin{cases} \lambda_1 \text{ si } \underline{x} \text{ proviene de } P_1 \\ \lambda_2 \text{ si } \underline{x} \text{ proviene de } P_2 \end{cases}$$

ξ una variable ficticia, λ_1 , λ_2 constantes

$\underline{x}$ el vector de mediciones o atributos para un individuo y si

$$\alpha = \frac{1}{\lambda_1 - \lambda_2}\,(\lambda_1 \underline{\mu}_2 - \lambda_2 \underline{\mu}_1) \quad , \quad \underline{\beta} = \frac{1}{\lambda_1 - \lambda_2}\,(\underline{\mu}_1 - \underline{\mu}_2)$$

mediante substitución directa puede verificarse que el miembro derecho de (II.17) se reduce a $\underline{\mu}_1$ cuando $\xi = \lambda_1$ y a $\underline{\mu}_2$ cuando $\xi = \lambda_2$. Entonces la ecuación que aparece en (II.17) es formalmente la regresión de $\underline{x}$ en o sobre ξ (ver Cramér (1946)). Dado que en este caso lo que se desea es decidir si un individuo pertenece a P_1 o a P_2 en base a las mediciones efectuadas, esto conduce a considerar la regre

sión inversa, i.e., la regresión de $\underline{\xi}$ en o sobre $\underline{x}$; obteniendo

$$E\left[\underline{\xi}\right] = X^* \underline{b}$$

donde $X^* = (\underline{e}, X')$ de dimensiones $(n_1 + n_2) \times (p + 1)$

$\underline{e} = (1, \ldots, 1)'$ con $n_1 + n_2$ elementos

$X = (\underline{x}_1^{(1)}, \underline{x}_2^{(1)}, \ldots \underline{x}_{n1}^{(1)}, \underline{x}_1^{(2)}, \ldots, \underline{x}_{n2}^{(2)})$ una matriz de $p \times (n_1 + n_2)$ componentes.

El vector $\underline{b}$ se obtiene de las ecuaciones normales

$$(X^{*\prime} X^*)\, \underline{b}^* = X^{*\prime} \underline{\xi} \qquad \text{(II.18)}$$

de donde $\underline{b} = (X^{*\prime} X^*)^{-1} X^{*\prime} \underline{\xi}$

efectuando el producto $X^{*\prime} \underline{\xi}$ se tiene

$$X^{*\prime} \underline{\xi} = \begin{bmatrix} n_1 \lambda_1 + n_2 \lambda_2 \\ \frac{n_1 n_2}{n_1 + n_2} (\bar{\xi}_1 - \bar{\xi}_2) \end{bmatrix}$$

por comodidad se eligen λ_1 y λ_2 tales que $\lambda_1 n_1 + \lambda_2 n_2 = 0$

p. ej. $\lambda_1 = \dfrac{n_2}{n_1 + n_2}$, $\lambda_2 = \dfrac{-n_1}{n_1 + n_2}$ (*). Así elegidos λ_1 y λ_2 y a partir de las ecuaciones normales (II.18) se tiene

(*)Valores utilizados por Fisher, los cuales facilidan los cálculos y como se observará los valores de λ_1, λ_2 aún cuando conducen a diferentes valores del vector $\underline{b}$ no tienen verdadera importancia pues todos los vectores $\underline{b}$ y por lo tanto a la función discriminante lineal, que se obtengan serán proporcionales, pues la función discriminante que se obtenga ha de estandarizarse antes de usarla, con lo cual la proporcionalidad no afecta en los resultados finales.

$$\underline{b}^{*\prime}\, X^{*\prime} X^{*}\, \underline{b}^{*} = \underline{b}^{*\prime} X^{*\prime}\, \underline{z}$$

$$= \frac{n_1\, n_2}{n_1 + n_2}\; \underline{b}'\, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

donde $\underline{b}^{*} = \begin{pmatrix} b_o \\ \underline{b} \end{pmatrix}$ i.e. $\underline{b}$ es el vector de los coeficientes de regresión de los elementos de $\underline{x}$ y b_o es el término constante de la regresión. Además, es necesaria una expresión de $\underline{b}$ en términos de la información muestral. Efectuando el producto $X^{*\prime} X^{*}\, \underline{b}$ se tiene

$$X^{*\prime} X^{*}\, \underline{b} = \begin{bmatrix} \underline{e}'\, \underline{e}\; b_o + \underline{e}'\, X'\, \underline{b} \\ X_a\; b_o \quad + X X'\, \underline{b} \end{bmatrix} \qquad \text{(II.20)}$$

entonces de (II.19) y (II.20) se tiene

$$b_o = -\frac{1}{\underline{e}'\, \underline{e}}\; \underline{e}'\, X'\, \underline{b}$$

y entonces puede demostrarse que

$$X \left(I - \frac{1}{\underline{e}'\underline{e}}\, \underline{e}\underline{e}' \right) X'\, \underline{b} = \frac{n_1 n_2}{n_1 + n_2}\, (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

de donde mediante un poco de álgebra se obtiene que

$$\left[X_1 \left(I_{n_1} - \frac{1}{n_1}\, \underline{e}_{n_1} \underline{e}_{n_1}' \right) X_1' + X_2 \left(I_{n_2} - \frac{1}{n_2}\, \underline{e}_{n_2} \underline{e}_{n_2}' \right) X_2' + \frac{n_1 n_2}{n_1 + n_2}\, \underline{d}\, \underline{d}' \right] \underline{b} = \frac{n_1 n_2}{n_1 + n_2}\, \underline{d}$$

i.e. $$\left(S_e + c^2\, \underline{d}\, \underline{d}' \right) \underline{b} = c^2\, \underline{d} \qquad \text{(II.21)}$$

donde $X = \left[X_1 \vdots X_2 \right]$, X_1 la matriz de observaciones correspondientes a la muestra de P_1, X_2 la respectiva para la población P_2, $\underline{d} = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$, $e_{n_i} = (1, \ldots, 1)'$ vector con n_i elementos $i = 1, 2.$, I_{n_i} la matriz identidad de orden n_i.

$$S_o = X_1 \left(I_{n_1} - \frac{1}{n_1} \underline{e}_{n_1} \underline{e}_{n_1}' \right) X_1' + X_2 \left(I_{n_2} - \frac{1}{n_2} \underline{e}_{n_2} \underline{e}_{n_2}' \right) X_2'$$

$$c^2 = \frac{n_1 n_2}{n_1 + n_2}$$

Las ecuaciones normales, expresión (II.21), pueden escribirse en forma concisa utilizando el resultado que aparece en Press ((1972), pag. 23).

$$\underline{b} = \frac{c^2}{1 + c^2 \underline{d}' S_o^{-1} \underline{d}} S_o^{-1} \underline{d}$$

de donde se observa que $\underline{b}$ es proporcional a $S^{-1} \underline{d}$ y $\underline{b}' x$ es proporcional a la función discriminante lineal de Fisher (II.7), habiendo substituído los parámetros por los respectivos estimadores y tomando en cuenta que $S_o = (n_1 + n_2 - 2) S$. Por lo anterior y excepto por una constante de proporcionalidad la función discriminante lineal y la función de regresión son iguales por lo que este enfoque del problema conduce al mismo procedimiento de discriminación que el de la función lineal de Fisher.

Kshirsagar (1966) presenta una interesante discusión acerca de lo apropiado que puede ser el aplicar los métodos del análisis de regresión a este caso, ya que no se cumplen las suposiciones de dicho análi

sis. Sin embargo, la justificación para usar la prueba de F para probar que los coeficientes de la ecuación de regresión verdaderos son iguales a cero (lo cual es equivalente a que $\underline{\mu}_1 = \underline{\mu}_2$ i.e. las poblaciones coinciden) proviene de la distribución de la D_p^2 ya que el cociente de sumas de cuadrados que se debe calcular es

$$F = \frac{(\text{Suma de Cuadrados debido a la regresión})/\text{g.l.}}{(\text{Suma de Cuadrados del error})/\text{g.l.}}$$

$$= \frac{n_1 + n_2 - p - 1}{p} \quad \frac{c^2}{n_1 + n_2 - 2} \quad D_p^2$$

donde $D_p^2 = (n_1 + n_2 - 2)\, \underline{d}'\, S_o^{-1}\, \underline{d}$ es un múltiplo de la distancia muestral de Mahalanobis.

Se desprende de lo anterior que los programas de computadora escritos para resolver problemas de regresión pueden ser usados también para resolver problemas de discriminación.

Es común que la función discriminante lineal sea interpretada en forma subjetiva, i.e., si en la función $\underline{\lambda}'\, \underline{x}$ el coeficiente de una variable es "grande" se considera que esa variable tiene valor discriminatorio, por lo contrario, si el coeficiente de una variable es "pequeño" (cercano a cero) se considera que dicha variable no tiene valor discriminante.

II.1.2 El caso de dos poblaciones con heterocedasticidad.

Una generalización al problema planteado es considerar dos poblaciones con distribución normal multivariada con vectores de medias diferentes así como matrices de covarianza diferentes. Anderson y Bahadur (1962) presentan una solución (suponiendo conocidos los parámetros) para esta faceta del problema, la cual está basada en el uso de un procedimiento lineal.

Se define un procedimiento lineal como aquel que asigna una observación $\underline{x}$ a la población P_1, si la combinación lineal $\underline{b}'\underline{x} \leq c$ - ($\underline{b} \neq \underline{0}$ un vector p - dimensional y c un escalar) en otro caso asigna $\underline{x}$ a la población P_2, para una elección adecuada de $\underline{b}$ y c.

Suponiendo entonces $\underline{\mu}_1 \neq \underline{\mu}_2$, $\Sigma_1 \neq \Sigma_2$ ($\underline{\mu}_i$ y Σ_i denotan el vector de medias y la matriz de covarianza en la población P_i $i = 1, 2$) Σ_1 y Σ_2 no singulares, Anderson y Bahadur (1962) se avocan al estudio de procedimientos lineales óptimos. En los términos en que se ha planteado el problema puede observarse fácilmente que las - probabilidades de clasificación equivocada dependen de las cantidades

$$y_1 = \frac{c - \underline{b}'\underline{\mu}_1}{(\underline{b}'\Sigma_1\underline{b})^{\frac{1}{2}}} \qquad y \qquad y_2 = \frac{\underline{b}'\underline{\mu}_2 - c}{(\underline{b}'\Sigma_2\underline{b})^{\frac{1}{2}}}$$

los cuales se obtienen al calcular las probabilidades de clasificación equivocada bajo cada población y son

$$p_{1|2} = P\left[\underline{b}'\underline{x} \leq c \mid \underline{x} \in P_2\right] = 1 - P\left[\underline{b}'\underline{x} > c \mid \underline{x} \in P_2\right]$$

$$= 1 - \underline{\Phi}(y_2) \quad *$$

$$p_{2|1} = P\left[\underline{b}'\underline{x} > c \mid \underline{x} \in P_1\right] = 1 - P\left[\underline{b}'\underline{x} \leq c \mid \underline{x} \in P_1\right]$$

$$= 1 - \underline{\Phi}(y_1) \quad *$$

donde $\underline{\Phi}(z)$ denota la distribución normal evaluada en z, y se plantea la solución a los siguientes problemas:

i) encontrar el procedimiento que minimice la probabilidad de un error de clasificación cuando el otro sea especificado.

ii) encontrar el procedimiento que minimice la máxima probabilidad de error

iii) encontrar el procedimiento que minimice la probabilidad de error cuando se han especificado probabilidades a priori para ambas - poblaciones.

La solución a estos problemas se encuentra dentro del conjunto de procedimientos lineales admisibles. Un procedimiento l_1 es me

* hay que observar que las probabilidades de clasificación equivocada son funciones decrecientes de y_1 y de y_2

jor que otro l_2 si cada probabilidad de clasificación equivocada para l_1 es menor o igual que para l_2 y al menos una es estrictamente menor. Un procedimiento es admisible si no hay otro procedimiento mejor que él.

Para caracterizar analíticamente los procedimientos lineales admisibles, lo cual significa encontrar el vector $\underline{b}$ que maximiza y_1 para cada y_2, obteniendo la constante c correspondiente de las fórmulas para y_1 ó y_2 se tiene que

$$\underline{b} = (t_1 \Sigma_1 + t_2 \Sigma_2)^{-1} \underline{\delta} \qquad \text{(II.22)}$$

t_1 , t_2 escalares y $\underline{\delta} = \underline{\mu}_2 - \underline{\mu}_1$

y entonces

$$c = \underline{b}' \underline{\mu}_1 + t_1 \underline{b}' \Sigma_1 \underline{b} = \underline{b}' \underline{\mu}_2 - t_2 \underline{b}' \Sigma_2 \underline{b} \qquad \text{(II.23)}$$

Anderson y Bahadur (1962) prueban que el procedimiento definido por (II.22) y (II.23) es admisible para cualquier t_1 y t_2 tales que $t_1 \Sigma_1 + t_2 \Sigma_2$ sea definida positiva (Teorema 2) y a menos de que $\underline{\delta}$ tenga una relación especial con Σ_1 y Σ_2 todos los procedimientos admisibles quedan definidos por las condiciones antes citadas. Además se espera que estos resultados cubran los casos de interés práctico.

Entonces dados t_1 y t_2 puede obtenerse el vector $\underline{b}$ óptimo y a su vez la constante c. A menudo t_1 y t_2 no son especificados de antemano, y la solución se determina de otra manera. Ahora - bien, retornando a los problemas i), ii) y iii) se tiene:

i) minimizar una probabilidad de clasificación errónea habiendo es pecificado la otra. Suponiendo que se da y_2 (que equivale a fi - jar la probabilidad de discriminación equivocada de un individuo de la población P_2) y se desea maximizar y_1, i.e. minimizar $P_{2|1}$.

La solución para este problema la proporcionan Anderson y Baha dur (1962) y consiste en que si $y_1 > 0$ (i.e. $P_{2|1} < 1/2$) si max $y_1 \geq 0$ se desea encontrar $t_2 = 1 - t_1$ tal que $y_2 = t_2 (\underline{b}' \Sigma_2 \underline{b})^{1/2}$ donde

$$\underline{b} = \left[t_1 \Sigma_1 + t_2 \Sigma_2 \right]^{-1} \underline{\delta} \qquad \text{(II.24)}$$

la solución puede hallarse vía ensayo y error. Para $t_2 = 0$ $y_2 = 0$ y para $t_2 = 1$ $y_2 = (\underline{b}' \Sigma_2 \underline{b})^{1/2} = (\underline{b}' \underline{\delta})^{1/2} = (\underline{\delta}' \Sigma_2^{-1} \underline{\delta})^{1/2}$

donde $\Sigma_2 \underline{b} = \underline{\delta}$

Para $t_2 > 0$, $t_1 < 0$ y $t_2 - t_1 = 1$, y_2 es una fun - ción decreciente de t_2 y en $t_2 = 1$ $y_2 = (\underline{\delta} \Sigma_2^{-1} \underline{\delta})^{1/2}$.

Si $y_2 > (\underline{\delta}' \Sigma_2^{-1} \underline{\delta})^{1/2} \Rightarrow \quad y_1 < 0$ y se busca t_2 tal que $y_2 = t_2 (\underline{b}' \Sigma_2^{-1} \underline{b})^{1/2}$ (i.e. proporcional al módulo de $\underline{b}$ en la métrica de Σ_2). Cuando $t_2 < 0$, $t_1 > 0$ y $t_1 - t_2 = 1$ se tiene $y_2 < 0$ (i.e. $P_{1|2} > 1/2$), en este caso y_2 es una función creciente de t_2, de nuevo se busca el valor de t_2 tal que $y_2 = t_2 (\underline{b}' \Sigma_2^{-1} \underline{b})^{1/2}$.

ii) minimizar la máxima probabilidad de error.

La solución es el procedimiento minimax, i.e., el procedimiento para el cual $y_1 = y_2$ *. Ya que para este procedimiento $P_{1|2}$, $P_{2|1} > 0$ $y_1 = y_2 > 0$ y $t_1 > 0$, $t_2 > 0$, se requiere encontrar $t = t_1 = 1 - t_2$

tal que

$$0 = y_1^2 - y_2^2 = t^2 \underline{b}' \Sigma_1 \underline{b} - (1-t)^2 \underline{b}' \Sigma_2 \underline{b}$$
$$= \underline{b}' \left[t^2 \Sigma_1 - (1-t)^2 \Sigma_2 \right] \underline{b}$$

existiendo una solución para lo anterior, la cual puede aproximarse por ensayo y error. Un enfoque alternativo es resolver para c en las expresiones de y_1 y y_2 obteniendo

$$y_1 = y_2 = \frac{\underline{b}' \underline{\delta}}{(\underline{b}' \Sigma_1 \underline{b})^{1/2} + (\underline{b}' \Sigma_2 \underline{b})^{1/2}} \qquad \text{(II.25)}$$

buscando el vector $\underline{b}$ que maximice (II.25) donde $\underline{b}$ es de la forma $\underline{b} = \left[t \Sigma_1 + (1 - t) \Sigma_2 \right] \quad 0 < t < 1$ (II.26)

* ver sección I.3 de este trabajo

Cuando $\Sigma_1 = \Sigma_2$ al doble del máximo de (II.25) se le conoce como la distancia entre las poblaciones P_1 y P_2 lo que sugiere la idea de extender la denominación a este caso.

iii) minimizar la probabilidad de error cuando han sido especificadas probabilidades a priori.

Suponiendo que tales probabilidades son q_1 de que provenga de P_1 y q_2 de que provenga de P_2, se tiene que la probabilidad de clasificación errónea es

$$q_1[1 - \Phi(y_1)] + q_2[1 - \Phi(y_2)] = q_1 + q_2 - [q_1\Phi(y_1) + q_2\Phi(y_2)]$$
$$= 1 - [q_1\Phi(y_1) + q_2\Phi(y_2)]$$

el objetivo es entonces, minimizar esta última expresión.

Como este procedimiento involucra a y_1 y a y_2 si se sabe que $y_1, y_2 \geq 0$ (i.e. $P_{2|1}, P_{1|2} < 1/2$) se puede substituir $y_1 = t\,(\underline{b}'\Sigma_1\underline{b})^{1/2}$, $y_2 = (1-t)\,(\underline{b}'\Sigma_2\underline{b})^{1/2}$ donde $\underline{b}$ esta dada por (II.23) entonces, maximizando con respecto a t se obtiene

$$q_1\,\phi(y_1)\frac{dy_1}{dt} + q_2\,\phi(y_2)\frac{dy_2}{dt} = 0 \qquad \text{(II.27)}$$

donde $\phi(u) = 2\pi e^{-1/2u^2}$

la expresión (II.27) no es fácil de resolver para t pero puede modificarse obteniendo

$$\frac{q_1}{(\underline{b}'\Sigma_1\underline{b})^{\frac{1}{2}}}\,\phi(q_1) = \frac{q_2}{(\underline{b}'\Sigma_2\underline{b})^{\frac{1}{2}}}\,\phi(q_2) \qquad \text{(II.28)}$$

los autores aconsejan graficar la curva de soluciones admisibles y tratar valores de t en (II.28) para encontrar una solución operativa.

Por último señalan que para todo procedimiento lineal admisible $\underline{b}'\underline{\delta} \neq 0$, i.e., $\underline{b}'\underline{\mu}_1 - \underline{b}\,\underline{\mu}_2 \neq 0$, lo cual significa que todo procedimiento lineal admisible hace uso del hecho de que $\underline{\mu}_1 \neq \underline{\mu}_2$.

II.1.3 El caso de k poblaciones.

El problema de discriminación entre k ($k > 2$) poblaciones es la siguiente etapa natural en la evolución del problema de discriminación. Von Mises (1945) presenta una solución bajo la suposición de que se tienen k poblaciones cada una con función de densidad P_j $j=1, 2, \ldots, k$ y se desea discriminar a N nuevos individuos, la solución consiste en subdividir el espacio p - dimensional de las observaciones en k regiones $R_1, R_2, \ldots, R_k$, asignando a un individuo que caiga en la región R_j a la población P_j, el objetivo es minimizar el riesgo de efectuar una decisión equivocada, que es el complemento a 1 de la probabilidad de decisión correcta que para un individuo de la población j' estará dado por

$$\Pi_{j'} = \int_{R_{j'}} P_{j'}(\underline{x})\, d\,\underline{x} \qquad j' = 1, 2, \ldots, k$$

Las características* que debe tener la partición óptima son:

i) para todas las k regiones R_l $l = 1, \ldots, k$ el valor de Π_l es constante.

ii) a lo largo de la frontera entre dos regiones sean estas R_l y $R_{l'}$

* las cuales en el caso de $k = 2$ coinciden con las propuestas por Welch (1939).

el cociente $\frac{\pi_l(x)}{\pi_{l'}(x)}$ es constante, puede observarse fácilmente que en el caso unidimensional sólo la primera característica es relevante. En todo caso la probabilidad de decisión correcta es igual al valor común π_l.

En cuanto a este problema tratando exclusivamente con poblaciones normales multidimensionales la mayoría de las soluciones para 2 poblaciones pueden extenderse de una manera directa. Se presenta aquí el tratamiento dado por Anderson (1958) al problema de discriminar entre k poblaciones normales con vectores de medias distintos y la misma matriz de covarianza bajo la suposición de que los parámetros se conocen. Para costos (o pérdidas) arbitrarias y probabilidades a priori conocidas pueden formarse las k funciones definidas en términos de la expresión I.3.2.

Si los costos de discriminación equivocada son iguales se usan las funciones

$$u_{jl} = \ln \frac{p_j(x)}{p_l(x)} = \left[x - \tfrac{1}{2}(\mu_j - \mu_l) \right]' \Sigma^{-1} (\mu_j - \mu_l)$$

y conociéndose las probabilidades a priori se define la región R_j (recordar desarrollo en la sección I.3) como el conjunto de las x's tales que

$$u_{jl} > \ln \frac{q_l}{q_j} \qquad l = 1, \ldots, k \qquad l \neq j$$

q_l $l = 1, \ldots, k$, probabilidad a priori de provenir de P_l.

$R_1, \ldots, R_k$ definidas en (II.28) por las constantes d_j, las cuales son determinadas de manera que las siguientes integrales:

$$P(j \mid j, \{R\}) = \int_{d_j - d_k}^{\infty} \cdots \int_{d_j - d_1}^{\infty} f_j \, du_{j1} \cdots du_{j,j-1} \, du_{j,j+1} \cdots du_{jk}$$

donde f_j es la densidad de la variable U_{ji} $(i = 1, \ldots, k)$ $i \neq k$

$$U_{ji} = \left[\bar{x} - \tfrac{1}{2} (\bar{\mu}_i - \bar{\mu}_j) \right]' \Sigma^{-1} (\bar{\mu}_j - \bar{\mu}_i)$$

y $\{R\}$ denota la partición del espacio p - dimensional en R_1, R_2, ..., R_k.

tengan un valor igual, minimizan la máxima pérdida (o costo) esperada condicional.

Anderson ((1958) teorema 6. 7. 1) afirma que así definidas R_1, ..., R_k, se minimiza el costo esperado de asignación equivocada. Si se desconocen las probabilidades a priori R_j queda definida por desigualdades de la forma

$$u_{jl} \geq d_j - d_l \qquad l = 1, \ldots, k \quad l \neq j \quad d_l \geq 0 \quad l = 1, \ldots, k \qquad \text{(II.28')}$$

del teorema 6. 7. 2 Anderson (1958) se tiene que bajo los supuestos establecidos anteriormente las regiones $R_1, \ldots, R_k$ definidas en - (II.28')con las constantes d_j determinadas de manera que las integrales:

$$P(j|j, \{R\}) = \int_{d_j - d_k}^{\infty} \cdots \int_{d_j - d_1}^{\infty} f_j \, du_{j1} \cdots du_{jj-1} \, du_{jj+1} \cdots du_{jk}$$

donde f_j es la densidad de la variable U_{ji} $(i = 1, \ldots, k)$ $i \neq k$

y

$$U_{ji} = \left[\underline{x} - 1/2 \, (\underline{u}_i - \underline{\mu}_j) \right]' \Sigma^{-1} (\underline{\mu}_j - \underline{\mu}_i)$$

$\{R\}$ denota la partición del espacio p - dimensional en $R_1, \ldots, R_k$ tengan un valor igual, minimizan la máxima pérdida (o costo) esperada condicional.

Si los parámetros se desconocen, suponiendo que se cuenta con una muestra de cada población, en vez de los parámetros, pueden substituirse en la expresión para u_{ij} $(\underline{x})$, los estimadores muestrales que son

$$\underline{x}_i = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} \underline{x}_{\alpha}^{(i)} \qquad i = 1, \ldots, k \quad \text{para} \quad \underline{\mu}_i$$

$\underline{x}_{\alpha}^{(i)}$ una observación de la muestra de N_i unidades de la población P_i

$$S = \frac{1}{\left(\sum_{i}^{k} N_i - k\right)} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (x_j^{(i)} - \bar{x}_i)(\bar{x}_j^{(i)} - \bar{x}_i)'$$

sin embargo, sólo se pueden utilizar los resultados antes citados si se cuenta con muestras "suficientemente" grandes.

Se ha propuesto para resolver la misma situación un procedimiento que sigue un razonamiento análogo al de la función discriminante lineal de Fisher. Si se denota por E a la matriz de cuadrados medios y productos cruzados entre las poblaciones y por D la matriz de cuadrados medios y productos cruzados dentro de las poblaciones para el caso de dos poblaciones, Fisher sugirió encontrar la combinación $\underline{\lambda}' \underline{x}$ (ver sección II.1) que maximizase al cociente $\phi = \underline{\lambda}' E \underline{\lambda} / \underline{\lambda}' D \underline{\lambda}$ del cuadrado medio entre poblaciones entre el cuadrado medio dentro de poblaciones de la combinación $\underline{\lambda}' \underline{x}$. Para más de dos poblaciones los vectores $\underline{\lambda}$ que maximizan ϕ corresponden a soluciones de la ecuación.

$$(E - \lambda D)\, \underline{\lambda} = \underline{0} \qquad \text{(II.29)}$$

Esta ecuación tiene soluciones que son los vectores característicos de $D^{-1} E$ y existen a lo más $f = \min (k - 1, p)$, donde k es el número de poblaciones y p es la dimensión del espacio de observaciones. A las variables generadas por combinaciones lineales entre estos vectores característicos y el vector de observaciones, i.e., a $v_1 = \underline{\lambda}_1'$

$\underline{x}$, $v_2 = \underline{\lambda}_2' \underline{x}$, ..., $v_r = \underline{\lambda}_r' \underline{x}$ se les llama variables canónicas. Estas variables también se interpretan en forma intuitiva al igual que la función discriminante lineal individual para el caso de dos poblaciones. De hecho se puede pensar que, en este caso, se tienen r funciones discriminantes lineales.

Es útil hacer una gráfica con los valores para las primeras variables canónicas (funciones discriminantes) en los puntos muestrales, de acuerdo a la dispersión de esos puntos se pueden inferir analogías entre poblaciones o grupos considerados. Para ejemplos de ésto ver Blackith y Reyment (1971) y Seal (1964).

Ya que $\left\{\underline{\lambda}_i' \underline{x}\right\}_{i=1}^{r}$ son transformaciones lineales de variables normales, ellas son normales a su vez (dado el valor de $\underline{\lambda}_i$) y entonces la regla de decisión, tomando en cuenta r vectores propios de $D^{-1} E$, es la siguiente: asignar la "nueva" observación $\underline{x}$ a P_i si

$$\sum_{l=1}^{r} \left(\underline{\lambda}_l' (\underline{x} - \hat{\underline{\mu}}_i) \right)^2 = \min_j \sum_{l=1}^{r} \left(\underline{\lambda}_l' (\underline{x} - \hat{\underline{\mu}}_j) \right)^2 \qquad \text{(II.30)}$$

en otros términos asignar $\underline{x}$ a P_i si

$$\left(\underline{y} - \tfrac{1}{2} \underline{v}_i \right)' \underline{v}_i = \max_j \left(\underline{y} - \tfrac{1}{2} \underline{v}_j \right)' \underline{v}_j$$

donde $\underline{y}' = (\underline{\lambda}_1' \underline{x}, \underline{\lambda}_2' \underline{x}, \ldots, \underline{\lambda}_r' \underline{x})$

y $\underline{y}_c' = (\underline{\lambda}_1' \hat{\underline{\mu}}_1, \underline{\lambda}_2' \hat{\underline{\mu}}_2, \ldots, \underline{\lambda}_r' \hat{\underline{\mu}}_r)$

El desarrollo que conduce a la regla de decisión anterior es análogo al seguido en la sección II.1 que es el considerar el proble - ma como uno de regresión habiendo definido k-1 variables ficticias.

$$\underline{y}_\alpha = \begin{cases} 1 & \text{si una observación } \underline{x} \text{ proviene de } P \\ 0 & \text{de otra manera } (\alpha = 1, \ldots, k-1) \end{cases}$$

y entonces

$$E(\underline{x}) = \underline{\mu}_k + (\underline{\mu}_1 - \underline{\mu}_k) y_1 + \cdots + (\underline{\mu}_{k-1} - \underline{\mu}_k) y_{k-1} \qquad \text{(II.31)}$$

la expresión anterior se reduce a $\underline{\mu}_j$ $(j = 1, \ldots, k)$ cuando $\underline{x}$ proviene de P_j y se puede escribir como

$$E[\underline{x}] = \underline{\mu}_k + \beta \underline{y} \qquad \text{(II.32)}$$

$\beta = \left[\underline{\mu}_1 - \underline{\mu}_k \,\middle|\, \underline{\mu}_2 - \underline{\mu}_k \,\middle|\, \cdots \,\middle|\, \underline{\mu}_{k-1} - \underline{\mu}_k \right]$ una matriz de $p \times (k-1)$

$\underline{y} = (y_1, y_2, \ldots, y_{k-1})'$ vector de $k-1$ componentes.

El objetivo será entonces maximizar el cociente de la suma de cuadrados entre grupos entre la suma de cuadrados dentro de grupos en

el análisis de varianza de la función lineal $\underline{\lambda}' \underline{x}$ i.e. maximizar el cociente

$$\phi = \frac{\underline{\lambda}' E \underline{\lambda}}{\underline{\lambda}' D \underline{\lambda}}$$

que es equivalente a maximizar *

$$\frac{\underline{\lambda}' E \underline{\lambda}}{\underline{\lambda}' (E + D) \underline{\lambda}}$$

llevando a cabo la diferenciación se obtiene que $\underline{\lambda}$ debe satisfacer el conjunto de ecuaciones

$$\left[E - r^2 (D + E)\right] \underline{\lambda} = \underline{0} \qquad \text{(II.33)}$$

donde r^2 es una raíz de la ecuación del determinante

$$\left|E - r^2 (D + E)\right| = 0 \qquad \text{(II.34)}$$

que es una condición necesaria y suficiente para que (II.33) tenga solución diferente de la trivial.

Se tiene la ventaja de que la expresión (II.34) tiene como raíces $r_1^2 > r_2^2 > \ldots > r_f^2$ ($f = \min(p, k-1) = \text{rango}(\beta)$) los cuadrados de las correlaciones canónicas muestrales entre la variable $\underline{x}$ y el vector de variables ficticias $\underline{y}$. Además ya que generalmente

* es preferible trabajar con esta expresión por motivos de cálculo.

se desconocen tanto $\underline{\mu}_i$ ($i = 1, \ldots, k$) como Σ se pueden utilizar las pruebas de significancia de las correlaciones canónicas para determinar cuáles son significativamente diferentes de cero, supóngase que son r, y así determinar cuantas funciones discriminantes utilizar. Por esta razón se llama variables canónicas a estas funciones discriminantes.

La razón de utilizar la expresión (II.30) se deriva del hecho de que mediante las r funciones discriminantes basadas en $\underline{\lambda}_1, \ldots, \underline{\lambda}_r$ se pueden definir "nuevas coordenadas" con respecto a los ejes $\underline{\lambda}_1, \ldots, \underline{\lambda}_r$ para el valor $\underline{x}$, que se desea asignar a una de las k poblaciones, i.e., definiendo

$$U_{01} = \underline{\lambda}_1' \underline{x} \quad , \quad U_{02} = \underline{\lambda}_2' \underline{x} \quad , \quad \ldots \quad , \quad U_{or} = \underline{\lambda}_r' \underline{x}$$

que equivale a disminuir la dimensionalidad del problema considerando solamente el subespacio afín en el que están las k medias poblacionales, así también las nuevas coordenadas de la media estimada $\bar{\underline{x}}_i$ de P_i son:

$$U_{i1} = \underline{\lambda}_1' \bar{\underline{x}}_i \quad , \quad U_{i2} = \underline{\lambda}_2' \bar{\underline{x}}_i \quad , \quad \ldots \quad , \quad U_{ir} = \underline{\lambda}_r' \bar{\underline{x}}_i$$

y la distancia entre el punto $\underline{x}$ y la media muestral $\bar{\underline{x}}_j$ en base a las nuevas coordenadas es

$$\sum_{\ell=1}^{r} (U_{0\ell} - U_{j\ell})^2 = \sum_{\ell=1}^{r} \left(\underline{\lambda}_\ell' (\underline{x} - \bar{\underline{x}}_j) \right)^2$$

de donde la elección de la regla de decisión de la expresión (II.30) es

intuitivamente clara. En cuanto a la posibilidad de aplicación de las pruebas de significancia de las correlaciones canónicas entre $\underline{x}$ y $\underline{y}$ no existe ningún impedimento teórico pues es suficiente que una de las variables tenga distribución normal para poder aplicar tales pruebas.

En el caso de conocer la forma funcional de las densidades $f_i(\underline{x})$ $i = 1, \ldots, k$. aún cuando se desconozcan los parámetros de ellas, puede utilizarse otra solución la cual maximiza la probabilidad media de clasificación correcta y consiste en asignar la observación "nueva" $\underline{x}$ a P_i si

$$q_i f_i(\underline{x}) = \max_j q_j f_j(\underline{x})$$

donde q_i es la probabilidad a priori de que un individuo pertenezca a P_i. Tal regla, en el caso de probabilidades a priori iguales, se reduce a asignar $\underline{x}$ a P_i si $f_i(\underline{x}) = \max_j f_j(\underline{x})$. Los parámetros pueden estimarse por algún método v.g., máxima verosimilitud. Si las densidades involucradas pertenecen a poblaciones normales con la misma matriz de covarianza la regla anterior conduce a asignar $\underline{x}$ a la población P_i si

$$(\underline{x} - 1/2\ \underline{\mu}_i)' \Sigma^{-1} \underline{\mu}_i = \max_j (\underline{x} - 1/2\ \underline{\mu}_j)' \Sigma^{-1} \underline{\mu}_j$$

para efectos prácticos μ_i es substituida por $\bar{x}_i$ y Σ por S. Este método es conocido como Función Discriminante Múltiple.

En cualquiera de las dos últimas soluciones las fronteras de las regiones de asignación son hiperplanos. En el caso del método que utiliza los vectores propios correspondientes a las correlaciones canónicas la dimensión de tales hiperplanos es min (k - 1, p) - 1 = f - 1 y en el caso de la Función Discriminante Múltiple la dimensión es p - 1.

Matusita (1971) discute el uso de la afinidad* entre distribuciones como una medida de información para la discriminación. Matusita (1973) empleando la afinidad presenta un método para determinar una proyección para efectuar la discriminación, el cual minimiza la afinidad entre las distribuciones involucradas. La función discriminante lineal presentada anteriormente para discriminar entre dos poblaciones con la misma matriz de varianza también se puede obtener por este método, así como una función discriminante para el caso de matrices de covarianza diferentes.

Matusita (1973) establece algunas propiedades de la afinidad como una medida de información para la discriminación, donde la afi

---

* en el sentido de Matusita,

nidad entre dos poblaciones se define como:

$$\rho_1 (F_1, F_2) = \int_R \sqrt{p_1(x) p_2(x)}\, dm$$

F_1 , F_2 dos distribuciones en el espacio R en el cual una medida m (Lebesgue, o de conteo, o mixta) esta definido y P_1 (x) , P_2 (x) son las funciones de densidad de F_1 y F_2 con respecto a m.

Así definida ρ_2 (F_1 , F_2) está relacionada a la distancia entre F_1 y F_2

$$d_2 (F_1, F_2) = \left\{ \int_R \left(\sqrt{p_1(x)} - \sqrt{p_2(x)} \right)^2 dx \right\}^{1/2}$$

mediante la expresión

$$d_2^2 (F_1 , F_2) = 2 (1 - \rho_2 (F_1 , F_2))$$

En el caso de varias distribuciones se define la afinidad como sigue. Sean F_1, ..., F_r distribuciones en R con funciones de densidad con respecto a m P_1 (x), ..., P_r (x) respectivamente.

$$\rho_r (F_1, \ldots, F_r) = \int_R (P_1 (x) \ldots P_r(x))^{1/r}\, dm$$

Este valor, Matusita afirma, que puede demostrarse que es independiente de la elección de la medida subyacente m. Además posee

la propiedad de ser simétrico en $F_1, \ldots, F_r$ y otras más (ver Matusita (1973)). Tiene características similares a medidas de información, tales como la de Kullback - Liebler para discriminación y es mejor que esta en cuanto a simetría en distribuciones y relación directa con la probabilidad de error en discriminación.

En cuanto a la búsqueda de una proyección, E, sobre un subespacio s - dimensional se busca que minimice

$$\rho_r (E(F_1), \ldots, E(F_r))$$

obteniendo el siguiente resultado (Matusita (1973)):

Sean B una matriz s x p de rango s y E una proyección de R_p sobre un subespacio s - dimensional entonces para resolver el problema de encontrar la proyección buscada no siempre hay necesidad de confinarse al conjunto de proyecciones E, se puede buscar una matriz B la cual minimice la afinidad $\rho_r (B(F_1), \ldots, B(F_r))$.

Considerando específicamente poblaciones normales p - dimensionales se tiene que si $F_1, \ldots, F_r$ son distribuciones normales $F_1 = N(\underline{\mu}_1, A_1^{-1}) \ldots, F_r = N(\underline{\mu}_r, A_r^{-1})$ donde $\underline{\mu}_j$ son vectores p - dimensionales y $A_1, \ldots, A_r$ matrices definidas positivas, simétricas, p x p. Entonces

$$\rho_r(\bar{T}_1,\ldots,\bar{T}_r) = \frac{\prod_{i=1}^{r}|A_i|^{\frac{1}{2r}}}{\left|\frac{1}{r}\sum_i A_i\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2r}\left[\sum_{i=1}^{r}\mu_i' A_i \mu_i - \left(\sum_{i=1}^{r} A_i\mu_i\right)'\left(\sum_i A_i\right)^{-1}\left(\sum_i A_i\mu_i\right)\right]\right\} *$$

(II.35)

que es invariante bajo transformaciones no singulares

$$T\underline{x} + \underline{b}.$$

En el caso de que las matrices de covarianza coincidan i.e. $A_1 = \ldots = A_r = A$ se tiene

$$\rho_r(\bar{T}_1,\ldots,\bar{T}_r) = \exp\left\{-\frac{1}{2r}\left[\sum_{i=1}^{r}\mu_i' A\mu_i - \left(\sum_{i=1}^{r} A\mu_i\right)'(rA)^{-1}\left(\sum_{i=1}^{r} A\mu_i\right)\right]\right\}$$

$$= \exp\left\{-\frac{1}{2r}\left[\sum_{i=1}^{r}\mu_i' A\mu_i - \left(\sum_{i=1}^{r}\mu_i\right)'\frac{1}{r}A^{-1}\left(\sum_{i=1}^{r}\mu_i\right)\right]\right\}$$

Como puede notarse, las soluciones mediante el enfoque clási co están muy restringidas en este caso.

---

* Inexplicablemente Matusita (1973) presenta un resultado que al parecer no es equivalente y no concuerda con resultados suyos anteriores ver Matusita (1967).

II.2 Soluciones bayesianas.

Geisser (1964) considera el problema de discriminación entre k poblaciones basando la decisión en la probabilidad de que la nueva observación $\underline{x}$ pertenezca a una de k poblaciones normales. Presenta una alternativa a los procedimientos clásicos ya que estos tienen la desventaja de que en el caso de desconocer los parámetros de las densidades normales involucradas, los problemas distribucionales que se enfrentan son muy complicados, además de que por su mismo desarrollo "son métodos de clasificación en masa, los cuales son útiles para clasificar gran número de observaciones sujetas a errores fijos de frecuencias de clasificación equivocada", pero que no permiten enunciado alguno acerca de la probabilidad de que una observación particular pertenezca a una u otra de las poblaciones.

Propone entonces asignar una densidad a priori (impropia) la

cual refleja la ignorancia que se tenga en el caso particular acerca de los parámetros para la población P_1 , entonces utilizando la verosimilitud de que $\underline{x}$ pertenezca a P_i y vía el teorema de Bayes obtener una solución para la probabilidad a posteriori de que la observación $\underline{x}$ pertenezca a P_i suponiendo conocidas las probabilidades de provenir de la población P_i. En lo que sigue $\underline{x}$ denotará una "nueva" observación por discriminar con probabilidad q_i de pertenecer a la población P_i , la distribución en P_i una normal con media $\underline{\mu}_i$ y matriz de covarianza Σ_i , $\bar{\underline{x}}_i = \hat{\underline{\mu}}_i$ y $S_i = \hat{\Sigma}_i$. Además se omitirán factores de proporcionalidad que no afecten el resultado final.

Caso I. Σ_i conocida , $\underline{\mu}_i$ conocida $\forall_i$

$$P(P_i \mid \underline{x}, q) \propto q_i \, |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2} t_r \Sigma_i^{-1} (\underline{x}-\underline{\mu}_i)(\underline{x}-\underline{\mu}_i)'\right\}$$

Caso 2. Σ_i conocida , $\underline{\mu}_i$ desconocida (medias diferentes) $\forall_i$ se tiene que

$$f(\bar{\underline{x}}_i \mid \underline{\mu}_i, \Sigma_i, P_i) \propto |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2} N_i t_r \Sigma_i^{-1} (\bar{\underline{x}}_i-\underline{\mu}_i)(\bar{\underline{x}}_i-\underline{\mu}_i)'\right\}$$

bajo la suposición de que la densidad impropia a priori de $\underline{\mu}_i$ es

$$g(\underline{\mu}_i)\, d\underline{\mu}_i \propto d\underline{\mu}_i \qquad \text{(no informativa)}^*$$

entonces

* para una discusión acerca de las densidades a priori no informativas de $\underline{\mu}_i$ y Σ_i ver Box y Tiao (1973) o mejor aún Geisser y Cornfield (1963)

$$P(P_i \mid \underline{x}, q) \propto \left(\frac{N_i}{N_i+1}\right)^{\frac{1}{2}p} |\Sigma_i|^{-\frac{1}{2}p} \exp\left\{-\frac{N_i}{2(N_i+1)} t_r \Sigma_i^{-1}(\bar{\underline{x}}_i - \underline{x})(\bar{\underline{x}}_i - \underline{x})'\right\}$$

Caso 3. Σ_i conocida , $\underline{\mu}_i = \underline{\mu}$ $\forall_i$ media común, desconocida

de nuevo

$$g(\underline{\mu}) d\underline{\mu} \propto d\underline{\mu} \quad .$$

$$P(P_i \mid \underline{x}, q) \propto q_i |\Sigma_i + \Lambda|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2} t_r (\Sigma_i + \Lambda)^{-1}(\bar{\underline{x}}_\omega - \underline{x})(\bar{\underline{x}}_\omega - \underline{x})'\right\}$$

donde

$$\Lambda^{-1} = \sum_{j=1}^{k} N_j \Sigma_j^{-1}$$

$$\bar{\underline{x}}_\omega = \Lambda \sum_{j=1}^{k} N_j \Sigma_j^{-1} \bar{\underline{x}}_j$$

Caso 4. Σ_i desconocida $\underline{\mu}_i$ conocida $\forall_i$

en este caso

$$N_i s_{ut_i} = \sum_{d=1}^{N_i} (x_{du_i} - \mu_{u_i})(x_{dt_i} - \mu_{t_i})$$

es el estimador de la componente ut – ésima de la matriz S_i que estima a Σ_i , además utilizando la densidad a priori, impropia y no informativa

$$g(\Sigma_i^{-1}) d\Sigma_i^{-1} \propto |\Sigma_i|^{\frac{1}{2}\sigma} d\Sigma_i^{-1} \qquad \sigma \leq N_i \quad *$$

se obtiene finalmente

---

* para una discusión sobre el valor v consultar Geisser y Cornfield (1963).

$$P(p_i \mid \underline{x}, q) \propto q_i \frac{\Gamma\left\{\frac{1}{2}(N_i+p-\sigma+2)\right\} \mid N_i S_i \mid^{\frac{1}{2}(N_i+p-\sigma+1)}}{\Gamma\left\{\frac{1}{2}(N_i-\sigma+2)\right\} \mid N_i S_i + \underline{x}\,\underline{x}' \mid^{\frac{1}{2}(N_i+p+2-\sigma)}}$$

Caso 5. Σ_i desconocida , $\underline{\mu}_i$ desconocida $\forall_i$

Este es el caso que ocurre más frecuentemente, aquí

$$(N_i-1)S_{uti} = \sum_{\alpha} (x_{\alpha ui} - \bar{x}_{ui})(x_{\alpha ti} - \bar{x}_{ti})$$

que será la componente ut – ésima de la matriz S_i que estima a la matriz Σ_i. La densidad conjunta a priori impropia y no informativa para $\underline{\mu}_i$ y Σ_i^{-1} es

$$g(\underline{\mu}_i, \Sigma_i^{-1})\, d\underline{\mu}_i\, d\Sigma_i^{-1} \propto |\Sigma_i|^{\frac{1}{2}\sigma}\, d\underline{\mu}_i\, d\Sigma_i^{-1} \qquad \sigma \leq N_i$$

finalmente se tiene

$$P(p_i \mid \underline{x}, q) \propto q_i \left\{\frac{N_i}{(N_i+1)\pi}\right\}^{\frac{1}{2}p} \frac{\Gamma\left\{\frac{1}{2}(N_i+p-\sigma+1)\right\}}{\Gamma\left\{\frac{1}{2}(N_i+1-\sigma)\right\} \mid (N_i-1)S_i \mid^{\frac{1}{2}}}$$

$$\times \left[1 + \frac{N_i (\bar{x}_i - \underline{\alpha})' S_i^{-1} (\bar{x}_i - \underline{\alpha})}{(N_i+1)(N_i-1)}\right]^{-\frac{1}{2}(N_i+p-\sigma+1)}$$

Caso 6. Σ_i desconocido , $\forall_i$ $\underline{\mu}_i = \underline{\mu}$ desconocida (pero vectores de medias iguales). Suponiendo la misma densidad a priori para $\underline{\mu}$ y la misma densidad a priori de Σ_i^{-1} como en los casos 4 y 5 se obtiene la densidad a priori conjunta

$$g(\underset{\sim}{\mu}, \Sigma_1^{-1}, \ldots, \Sigma_k^{-1}) \propto \prod_{i=1}^{k} |\Sigma_j|^{\frac{1}{2}\sigma}$$

y finalmente

$$P(P_i | \underset{\sim}{x}, q) \propto q_i \int \left\{ \left| \sum_{d=1}^{N_i} (\underset{\sim}{x}_{di} - \underset{\sim}{\mu})(\underset{\sim}{x}_{di} - \underset{\sim}{\mu})' + (\underset{\sim}{x} - \underset{\sim}{\mu})(\underset{\sim}{x} - \underset{\sim}{\mu})' \right|^{-\frac{1}{2}(N_i - p + 2 - \sigma)} \right.$$

$$\left. \times \prod_{j \neq i} \left| \sum_{d=1}^{N_j} (\underset{\sim}{x}_{dj} - \underset{\sim}{\mu})(\underset{\sim}{x}_{dj} - \underset{\sim}{\mu})' \right|^{-\frac{1}{2}(N_j + 1 + p - \sigma)} \right\} d\underset{\sim}{\mu}$$

Geisser apunto que debido a la dificultad de integrar sobre $\underset{\sim}{\mu}$ debe dejar el resultado en la forma anterior.

Caso 7. $\Sigma_i = \Sigma$ desconocida pero común, $\underset{\sim}{\mu}_i$ conocida $\forall_i$

De nuevo se utiliza el estimador de Σ, S como en el caso 4 y además se define como densidad a priori para la Σ común

$$g(\Sigma^{-1}) d\Sigma^{-1} \propto |\Sigma|^{\frac{1}{2}\sigma} d\Sigma^{-1} \qquad \sigma \leq N \qquad N = \sum_{j=1}^{k} N_j$$

y $S = \frac{1}{N} \sum_{j=1}^{k} N_j S_j$ obteniendo

$$P(P_i | \underset{\sim}{x}, q) \propto q_i \left\{ 1 + N^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_i)' S^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_i) \right\}^{-\frac{1}{2}(N + p + 2 - \sigma)}$$

Caso 8. $\Sigma_i = \Sigma$ desconocida pero común, $\underset{\sim}{\mu}_i$ desconocida $\forall_i$ (mismas suposiciones que en MANOVA)

Definiendo $S = \frac{1}{N-k} \sum_{i=1}^{k} (M_i - 1) S_i$ donde $N = \sum_{i=1}^{k} N_i$ y S_i definida

como en el caso 5, la densidad a priori para Σ^{-1} es la misma que en el caso 7 sólo que con la restricción $\nu \leq N - k$ y para $\frac{\mu_i}{\Gamma_i}$ la densidad a priori como en el caso 5.

Se tiene

$$P(P_i \mid \underline{x}, q) \propto \left(\frac{N_i}{N_i+1}\right)^{\frac{1}{2}p} \frac{\Gamma\left\{\frac{1}{2}(N-k+p-\nu+2)\right\}}{\Gamma\left\{\frac{1}{2}(N-k-\nu+2)\right\}}$$

$$\times \left[1 + \frac{N_i(\bar{\underline{x}}_i - \underline{x})' S^{-1} (\bar{\underline{x}}_i - \underline{x})}{(N_i+1)(N_i-k)}\right]^{-\frac{1}{2}(N-k+p-\nu+2)}$$

Como puede observarse, excepto para el caso 6, se tienen formas relativamente simples para expresar la solución con respecto a las probabilidades a posteriori.

La regla de decisión, como es común en el análisis bayesiano, se ha de basar en las distribuciones a posteriori (que indican el grado de credibilidad después de obtenida la información). Es por esto que no se proporciona una regla específica sino que ésta debe ser determinada por el investigador, por ejemplo usar "momios" relativos, o asignar la nueva observación , $\underline{x}$, a la población con mayor valor de la probabilidad a posteriori dado $\underline{x}$.

En cuanto al cálculo de los momios relativos (relative odds) -

Press (1972) presenta el cálculo en base a la densidad predictiva definida como sigue:

$$h(\underline{x}_{N+1} \mid \underline{x}_1, \ldots, \underline{x}_N) = \int g(\underline{x}_{N+1} \mid \theta)\, q(\theta \mid \underline{x}_1, \ldots, \underline{x}_N)\, d\theta$$

donde $q(\theta \mid \underline{x}_1, \ldots, \underline{x}_N) \propto p(\theta)\, f(\underline{x}_1, \ldots, \underline{x}_N \mid \theta)$

$h(\cdot \mid \cdot)$ denota la densidad predictiva

$g(\cdot \mid \cdot)$ denota la densidad de una observación futura

$p(\theta)$ es la densidad a priori para θ

$f(\underline{x}_1, \ldots, \underline{x}_n \mid \theta)$ es la función de verosimilitud de las observaciones para el parámetro θ

y

$q(\theta \mid \underline{x}_1, \ldots, \underline{x}_n)$ es la densidad a posteriori del parámetro dados los datos

La densidad predictiva resulta en caso de desconocer los parámetros ($\underline{\mu}$'s y Σ's) una densidad t de Student multivariada:

$$p(\underline{x} \mid \text{datos}, P_j) = \frac{k_j}{\left[1 + \frac{N_j}{N_j^2 - 1} (\underline{x} - \bar{\underline{x}}_j)' S_j^{-1} (\underline{x} - \bar{\underline{x}}_j)\right]^{N_j/2}}$$

donde k_j es una constante que no depende de $\underline{x}$ dada por

$$k_j = \left[\frac{N_j}{(N_j+1)\pi} \right]^{p/2} \frac{\Gamma\left(\frac{N_j}{2}\right) p_j}{\Gamma\left(\frac{N_j - p}{2}\right) \left| (N_j-1) S_j \right|^{1/2}}$$

P_j la probabilidad a priori de provenir de la población P_j

N_j es el número de observaciones extraídas de P_j

Para que sea aplicable este enfoque no es necesario que el tamaño de las muestras de ensayo sea grande y además las matrices de covarianza no necesitan ser iguales (como en general se requiere en los enfoques "clásicos").

Queda por discutir el problema cuando se desconocen las probabilidades a priori de pertenecer a alguna de las poblaciones P_i. - Geisser apunta que hay dos posibles situaciones, la primera cuando se considera que las q_i son estimables a partir de los cocientes N_i/N para $i = 1, \ldots, k$, $N = \sum_{i=1}^{k} N_i$. Esta situación se originaría en el caso de haber muestreado unidades aleatoriamente de una mezcla totalmente aleatorizada de las k poblaciones de manera que los cocientes N_i/N reflejen la probabilidad de que una observación provenga de la población respectiva. En este caso se tiene una densidad multinomial para las N_j's recordando que $N_k = N - (N_1 + N_2 + \ldots - N_{k-1})$ y $q_k = 1 - (q_1 + \ldots + q_{k-1})$ entonces la verosimilitud dadas las N_j's es

$$L(q_1, \ldots, q_k) \propto \prod_{j=1}^{k} q_j^{N_j}$$

Si se supone que la densidad a priori de las q_i's es de la forma Dirichlet se tiene

$$g(q_1, \ldots, q_{k-1}) \propto \prod_{j=1}^{k} q_j^{\alpha_j}$$

se obtiene así la densidad a posteriori:

$$P(q_1, \ldots, q_{k-1} \mid N_1, \ldots, N_{k-1}) \propto \prod_{j=1}^{k} q_j^{N_j + \alpha_j}$$

Finalmente utilizando esta densidad y la correspondiente a $\underline{x}$ dado $q_1, \ldots, q_{k-1}$ se obtiene la densidad a posteriori de P_i dado $\underline{x}$ que resulta

$$P(P_i \mid \underline{x}) = \frac{(N_i + \alpha_i + 1) f(\underline{x} \mid P_i)}{\sum_{j}^{k} (N_j + \alpha_j + 1) f(\underline{x} \mid P_j)}$$

donde $f(\underline{x} \mid P_i)$ denota cualquiera de las densidades predictivas previamente obtenidas con los numeradores representando los "momios relativos" de las varias P_i's.

Ahora bien, en caso de que los cocientes N_i/N no permitan estimar de una manera adecuada las probabilidades q_i se supone que las N_i's se escogen de tal manera que no guarden relación con el incremento de información con respecto a la estimación de las q_i's, lo que es

equivalente a suponer que $N_i = 0 \quad \forall_i$ en la expresión anterior de don de resulta que:

$$P(P_i \mid \underline{x}) \propto (a_i + 1)\, f(\underline{x} \mid P_i)$$

donde las α_i's pueden considerarse que reflejan frecuencias previas o impresiones acerca de las frecuencias de las diferentes P_i's, en caso de no contar con datos previos ni otra clase de información a priori, Geisser señala que al elegir $\alpha_i = \alpha \quad i = 1, \ldots, k$ se llega al mismo resultado que se hubiese obtenido de considerar a las k poblaciones igualmente probables a priori. (i.e. $q_i = 1/k$).

En cuanto a la elección de v se propone como un valor conveniente en caso de desconocimiento a priori de los parámetros de P_i - al usar $v = p + 1$. Como puede observarse en el caso de que se conozca Σ se escoge $v = 0$.

Dunsmore (1966) enfoca el problema desde el punto de vista de teoría de decisiones bayesiana, considerándolo como una categoría de un problema más amplio que tiene la siguiente estructura: una clase $\mathcal{F} = \{ F_x : x \in \mathcal{X} \}$ de futuros experimentos donde cada F_x tiene el mismo espacio de resultados $\mathcal{Y}$, además el espacio de decisiones coincide con $\mathcal{X}$ o con una clase de subconjuntos de $\mathcal{X}$. Para un

valor dado y, sea y_o , se requiere un estimador del correspondiente valor de x o una región en la que se piensa que x se encuentra. Se tienen los resultados de un experimento informativo $\mathcal{E}$, y_1, y_2, ..., y_n que son los resultados de realizaciones de experimentos independientes F_{x_1}, ..., F_{x_n} donde x_1, ..., x_n son conocidos, pero pueden haberse originado aleatoriamente. Otras categorías de este problema son i) la calibración (ver Aranda y Valencia (1974)), ii) la regularización y iii) la optimización.

Así planteado el problema, teniendo como objetivo maximizar la utilidad esperada (o minimizar la pérdida esperada) con una función de utilidad definida convenientemente, Dunsmore genera un modelo para problemas en los cuales F_x se ha desarrollado para un valor desconocido de x y lo aplica al problema de discriminación. En estos problemas los valores de x son cualitativos, es más,x actúa como una variable indicadora.

Dunsmore (1966) encuentra primero la solución para el caso en que las parejas (x_1 , y_1), ..., (x_n , y_n) del experimento, provienen de una densidad bivariada, además se restringe a y unidimensional, donde $p(y \mid x , \underline{\theta})$ es $N(\mu_x , \sigma^2)$, $\underline{\theta}$ descubre las características de incertidumbre en el modelo i, $p(x = i \mid \psi) = \phi_i$ $\quad i = 1, \ldots, k$, ψ

denota la incertidumbre en la distribución de x $\phi_k = 1 - \phi_1 - \cdots - \phi_{k-1}$.

Si se conoce σ^2 la varianza de la distribución condicional de y dado x se tienen dos parámetros $\underline{\Theta} = (\mu_1, \ldots, \mu_k)$ y $\Psi = (\phi_1, \ldots \phi_{k-1})$ definiendo distribuciones conjugadas a priori para $\underline{\Theta}$ y Ψ se obtiene finalmente que la utilidad esperada es de la forma

$$U(\underline{z}, y_0, R) \propto \sum_{i=1}^{k} V(R, x=i)(\beta_i + m_i)(\sigma^2 + B_{ii})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{(y_0 - A_i)^2}{(\sigma^2 + B_{ii})}\right\}$$

donde $\underline{z}$ denota el conjunto de parejas ordenadas $(x_1, y_1) \ldots, (x_n, y_n)$

$V(R, x = i)$. denota a la función de utilidad para la región R (la cual puede ser un punto) si el valor de la variable indicadora x es igual a i.

β_i $i = 1, \ldots, k$ es el parámetro de la distribución de Dirichlet que es la conjugada de Ψ.

m_i es el número de unidades en el experimento informativo que pertenecen a la población P_i.

B_{ii} es la componente ii de la matriz de covarianza B de la distribución a posteriori de $\underline{\Theta}$ dado $\underline{z}$.

A_i es el elemento i - ésimo del vector de medias de la distribución a posteriori de $\underline{\Theta}$ dado $\underline{z}$.

y_o es el valor por asignar a alguna de las k poblaciones.

Como podría esperarse el caso para σ^2 desconocida se complica aún más obteniéndose que

$$U(z, y_o, \theta) \propto \sum_{i=1}^{k} V(\theta, z=i)(p_i + m_i)\left\{C(1+B^{ii})\right\}^{-\frac{1}{2}}\left\{D + \frac{(y_o - A_i)^2}{C(1+B^{ii})}\right\}^{-\frac{1}{2}(D+1)}$$

con la misma notación de la fórmula anterior y donde D = d + n , d parámetro de la densidad conjugada normal gama* para $\underline{\Theta}$, n el total de observaciones en el experimento informativo.

$$C = \frac{1}{D}\left(\sum_i y_i^2 + \underline{a}'\underline{b}\,\underline{a} - A'BA + cd\right)$$

$\underline{a}$ vector de medias de la densidad conjugada normal gama citada anteriormente.

b matriz de covarianza de tal distribución ,

c parámetro de la misma distribución.

B^{ii} denota la componente ii de la inversa de la matriz B.

Para el caso en el cual los valores de $x_1, \ldots x_n$ en el experimento informativo (muestra de ensayo, como se le llamó en el capítulo

* ver De Groot (1970)

I) son especificados de antemano se tiene que la densidad a priori de x como $\pi(x = i) = \pi_i \quad i = 1, \ldots, k$ y la única alteración que sufren los dos resultados anteriores es que π_i substituye a $(\beta_i + m_i)$

Como puede observarse, a pesar de que es posible extender los resultados a y multidimensionales de una manera directa (según - afirma Dunsmore (1966)) la complejidad aumenta enormemente debido al incremento en el número de parámetros. Es la opinión de este autor que simplemente observando los resultados de páginas anteriores se nota la mayor simplicidad y tal vez efectividad de los resultados de Geisser (1964).

II.3 Soluciones específicas para casos que involucran variables diferentes de la normal.

Como es de suponer el problema de discriminación no solamente se presenta entre poblaciones normales, es más en muchas situaciones la discriminación se ha de basar en variables cualitativas por lo - que el estudio de casos con variables diferentes ha atraído el interés de muchos investigadores. Aquí se presenta un breve resumen de las soluciones presentadas agrupadas de acuerdo al tipo de distribución considerado.

III.3.1 Distribución multinomial.

Considerando una variable x multinomial con m categorías en cada una de las posibles poblaciones de origen.

Matusita (1956) propone una solución para discriminar entre dos poblaciones P_1 y P_2, contando con muestras de n_1 y n_2 unidades correspondientes respectivamente a P_1 y P_2. Se cuenta también con una muestra de n unidades que provienen de la población P la cual puede ser P_1 o P_2. La solución consiste en una regla de distancia mínima basada en las muestras antes citadas. La distancia se calcula entre las funciones de distribución empírica construídas para cada población, la función distancia se define como la raíz cuadrada de

$$\| F - G \|^2 = \sum_{i=1}^{m} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2$$

donde $(p_1, \ldots, p_n)$ y $(q_1, \ldots, q_n)$ son las probabilidades de cada categoría o celda correspondientes a las distribuciones F y G respectivamente, determinándose que la población de la que proviene la muestra no clasificada, es aquella para la cual la función distancia es menor. Matusita (1956) obtiene cotas mínimas para la probabilidad de clasificación correcta y un valor de tal probabilidad cuando los tamaños de las muestras son grandes, además discute el caso de que se tenga n = 1 i.e., un "individuo" a discriminar solamente.

Chernoff (1959) también enfoca el problema de dos poblaciones con la restricción de que una de ellas, P_1, tiene la misma probabilidad para cada una de sus categorías, en tanto que para la otra, P_2, simplemente se establece que se desconocen las probabilidades por categoría. Se tiene una muestra de n_2 unidades de la población P_2 y la solución se diseña para determinar si una muestra de tamaño n de una población P pertenece a P_1 ó P_2. Los resultados están dirigidos hacia aplicaciones cuando m, n_2 y n son grandes y el cociente de las probabilidades de error es muy pequeño o muy grande.

Wesler (1959) considera el problema de dos poblaciones multinomiales donde las probabilidades por categoría son cualquier permutación de las componentes de un vector de probabilidades $\underline{P}_{(i)}$ $(i = 1, 2)$. El objetivo es clasificar una muestra de n observaciones de la población P (que puede ser P_1 ó P_2) minimizando una probabilidad de error manteniendo controlado el otro en un valor máximo que se mantiene fijo. Wesler (1959) obtiene una solución aproximada para n grande y considera el caso m = 2 i.e., variables dicotómicas.

Cochran y Hopkins (1961), se interesaron en el problema debido a discusiones sostenidas acerca del posible uso de análisis discriminante en diagnóstico médico. El problema que se plantea es el del

manejo de datos cualitativos y donde cada medición toma sólo un mínimo finito (generalmente pequeño) de valores distintos. Los autores obtienen la forma de la regla de Bayes y consideran en particular la regla de máxima verosimilitud. Consideran el efecto de un tamaño de muestra finito sobre la estimación de la probabilidad de clasificación equivocada. Se proponen métodos para obtener estimadores insesgados de tal probabilidad y de la diferencia entre la PCE (probabilidad de clasificación equivocada) óptima teórica y la real. Se analiza también la relación que parece ser adecuada entre el tamaño de las muestras de ensayo y el número de estados o categorías.

Hills (1966) presenta desarrollos teóricos en los estudios de los errores de clasificación equivocada para la regla obtenida mediante el método de máxima verosimilitud para el caso de dos poblaciones. Para una población dicotómica, demuestra que la PCE para la regla de máxima verosimilitud, es mayor que la correspondiente para la misma regla bajo conocimiento completo de las distribuciones. Obtiene aproximaciones normales para el valor esperado del estimador de reasignación (Smith (1947)) de la PCE el cual subestima la PCE de la regla de máxima verosimilitud.

Bunke (1966) estudia asintóticamente una propiedad de la regla

minimax estimada mediante funciones de distribución empíricas para - distribuciones multinomiales.

Glick (1969) considera distribuciones discretas en general pero también trata especialmente la distribución multinomial. Este trabajo generaliza algunos de los resultados de Cochran y Hopkins (1961) y - Hills (1966), antes mencionados y proporciona pruebas rigurosas. Glick (1973) tratando en especial distribuciones multinomiales demuestra que la regla intuitiva basada en muestras para discriminación entre dos distribuciones multinomiales tiene una tasa de no-error real, la cual converge a la tasa óptima exponencialmente al aumentar el tamaño de muestra (n). Se prueba también que el sesgo optimista* de la tasa aparente de no-error (como un estimador del óptimo) es proporcional a $n^{-1/2}\, a^n$, donde $a < 1$ excepto para el caso especial considerado por Cochran y Hopkins (1961). Las convergencias anteriores se relacionan a los ejemplos numéricos de Hills (1966) y se comparan a estudios analíticos y - vía método Montecarlo de las correspondientes convergencias para discrimirantes lineales para normales multidimensionales basados en muestras.

Goldstein y Rabinowitz (1975), presentan un procedimiento basado en muestras para seleccionar un subconjunto óptimo de variables pa

---

* optimista en el sentido de que subestima

ra el problema de clasificación (discriminación) en dos grupos o poblaciones. Basándose en los trabajos de Glick (1972, 1973) examinan dos variaciones de un procedimiento el cual busca diferencias en los resultados discriminantes como un método para evaluar la importancia de varios subconjuntos de variables. Dos ejemplos sirven para ilustrar los métodos.

II.3.2. Distribuciones Bernoulli multivariadas.

La variable aleatoria $\underline{x}$ es un vector p x 1 y cada componente de $\underline{x}$ toma valores 0 ó 1.

Bahadur (1961) tratando el caso de dos poblaciones da algunas aproximaciones a la regla mediante logaritmo del cociente de verosimilitudes. También se obtienen algunas aproximaciones a la medida de información simétrica de Kullback - Leibler, J. Estas aproximaciones también son útiles cuando J es pequeño, p es grande y la interdependencia entre las componentes del vector $\underline{x}$ no es apreciable.

Solomon (1960, 1961) con respecto a dos poblaciones también presenta un estudio numérico de la efectividad medida en términos de la PCE y comparaciones relativas entre: reglas basadas en la suma de las componentes de $\underline{x}$, función discriminante lineal de Fisher, estadística de cociente de verosimilitudes y algunas funciones truncadas obtenidas de la representación en series para funciones de probabilidad de Bahadur (1961).

Hills (1967) para el caso de dos poblaciones, considera algunas alternativas simples a la FDL* de Fisher basadas en muestras de

* Función discriminante lineal

cada una de las poblaciones bajo consideración n_1 observaciones de P_1 y n_2 de P_2. Trata el problema de estimar el logaritmo del cociente de verosimilitudes en un punto dado $\underline{x} = \underline{x}_o$, sugiriendo diferentes estimadores útiles en diferentes situaciones: si el número m de categorías es pequeño el cociente de verosimilitudes en $\underline{x}_o$ puede estimarse mediante el cociente de frecuencias de sujetos de las muestras de P_1 y P_2 que cayeron en la misma celda que x_o. Si m es grande en relación con n_1 y n_2 , se propone un estimador por "vecino próximo" de orden 1:

$$\left(\frac{r_1 + r_1'}{n_1} \right) \Big/ \left(\frac{r_2 + r_2'}{n_2} \right)$$

donde r_j' es el número de vecinos próximos en una muestra de n_i de P_i cuyos valores $\underline{x}$ difieren de $\underline{x}_o$ en solamente un componente. También propone estimadores por "vecino próximo" de orden > 1 y estudia numéricamente las distribuciones de estos estimadores. Presenta extensiones a sus métodos para tratar con datos discretos no solamente del tipo (0, 1) y para más de dos poblaciones. Se sugiere un método paso a paso para seleccionar componentes mediante el uso de la medida de información J de Kullback y Leibler.

Martin y Bradley (1972) considerando una función de probabilidad de la variable x como

$$P_i(x) = f(x)\left[1 + h_s(a_i, x)\right]$$

en la población P_i donde h_s es más función lineal de los polinomios - ortogonales sobre el espacio muestral de x. Este trabajo trata el problema de estimación de a_i y f sujeto a algunas restricciones.

II.3.3 Distribuciones contínuas paramétricas diferentes de la normal.

Cooper (1962, 1963), demuestra que el procedimiento óptimo de discriminación para distribuciones normales multidimensionales en el cual se asigna la nueva observación en base a la comparación de formas cuadráticas puede extenderse a una clase de distribuciones mucho más amplia. Estas clases son las distribuciones del tipo de Pearson II y VII (ver Kendall y Stuart (1958, 1961)).

Battacharya y Das Gupta (1964), consideran el caso de dos poblaciones. La distribución de la variable aleatoria se toma como un miembro de la familia exponencial con un sólo parámetro. Se obtiene una clase de reglas de Bayes admisibles.

Nuevamente Cooper (1965), trata el problema considerando en esta ocasión que la función densidad de probabilidad de $\underline{x}$ en P_i es

$$f_i(\underline{x}) = A_i \; |\Sigma_i|^{-1/2} \; h_i\left[(Q_i(\underline{x}))^{-1/2} \right]$$

donde Q_i es una forma cuadrática definida positiva y $h_i(u)$ es una función monótona decreciente. Se estudia la estadística del cociente de verosimilitudes.

Day y Kerridge (1967), consideran la f.d.p. para $\underline{x}$ en P_i co

mo

$$f_i(\underline{x}) = d_i \exp\left[-1/2\,(\underline{x} - \underline{\mu}_i)'\,\Sigma^{-1}\,(\underline{x} - \underline{\mu}_i)\right] h(\underline{x})$$

estudiando dos casos: (1) $h(\underline{x}) \equiv 1$, (2) $\Sigma = I$ y $h(\underline{x}) = 1$ si todas las componentes de $\underline{x}$ son 0 ó 1 y $h(\underline{x}) = 0$ en otro caso. La probabilidad a posteriori de la hipótesis $H_i : P = P_i$, dado la observación $\underline{X} = \underline{x}$, se expresa como $\dfrac{\exp(\underline{x}'\,\underline{b} + c)}{1 + \exp(\underline{x}'\,\underline{b} + c)}$

Este trabajo trata principalmente con los estimadores máximo verosímiles de $\underline{b}$ y c. La idea de decisión "dudosa" se incorpora en el problema de discriminación.

Finalmente Anderson (1972), trata el caso de 2 ó más poblaciones, cuando todo o la mayoría de las observaciones son cualitativas. El método de discriminación logística sugerido por Cox (1966) y Day y Kerridge (1967), se extiende a la situación en que se tienen muestras separadas de cada población usando los resultados de y Silvey (1958) sobre estimación máximo verosímil con restricciones. Tal método de discriminación se extiende a más de dos poblaciones, se investigan las propiedades del método vía simulación y se aplica a un caso real.

II.3.4. Otros casos.

Kendall (1966), sugiere algunas reglas basadas en la categorización de los datos.

Marshall y Olkin (1968), en relación a la situación especial - formulada en el primer capítulo de este trabajo consideran una variable aleatoria binomial x, con una probabilidad de éxito π la cual se distribuye a su vez como uniforme en el intervalo (0, 1). Se obtiene la forma de la regla de Bayes.

Chang y Afifi (1974), presentan un procedimiento Bayesiano para clasificar una observación consistente de una observación dicotómica x y un vector b - dimensional con componentes contínuas y, tal procedimiento se deriva para un modelo biserial suponiendo que la distribución de y dado x es normal. El procedimiento se reduce a dos funciones discriminantes lineales, una para cada valor de x. Se presenta un ejemplo utilizando datos en pacientes en estado crítico. Se discuten además las extensiones, considerando variables politómicas o varias variables dicotómicas.

Shumway y Unger (1974) aplican ciertas aproximaciones espec

trales al problema de discriminar entre dos procesos normales me - diante filtrado lineal. Los valores límites de i) la tasa de informa ción de discriminación de Kullback - Liebler, ii) tasa de divergencia J y iii) probabilidad de detección, se expresan en términos de densi dades espectrales de las dos poblaciones y la transformada de Fourier - Stieltjes de la diferencia media entre ellas. Filtros discriminantes lineales que maximizan (i), (ii) y (iii) se aproximan por los mismos - métodos y se aplican a registros sísmicos de temblores seleccionados y explosiones nucleares.

Krzanowski (1975), deriva la regla de clasificación o discrimi nación mediante cociente de verosimilitudes a partir del modelo de lo calización (tratado en Afifi y Elashoff (1969), Chang y Afifi (1974) y en Olkin y Tate (1961)). Esta regla es aplicable cuando los datos contie nen variables binarias y contínuas. Se propone un método para esti - mar tal regla en situaciones prácticas y evaluar su funcionamiento. Las pérdidas que se incurren por este procedimiento de estimación se investigan y se estudia el uso de la función discriminante lineal para el caso de parámetros poblacionales conocidos. Se compara el funcio namiento del método propuesto con otras reglas de clasificación (discri minación) aplicándolo a algunos conjuntos de datos.

II.4 Soluciones no paramétricas o de distribución libre.

Como se vió en el capítulo I las soluciones en este terreno - se pueden agrupar en tres grandes categorías: reglas de inserción, uso de estadísticas involucradas en el desarrollo de algunas pruebas para problemas con 2 muestras o con k-muestras y algunos métodos que típicamente se han utilizado en problemas de discriminación como - por ejemplo distancia mínima, afinidad entre distribuciones, etc.

Entre las reglas de inserción con estimadores de densidad existen varios trabajos interesantes entre los que destacan las de Glick - (1969, 1972), en donde se estudian algunas propiedades de tales procedimientos.

Fix y Hodges (1951), consideran también estas reglas de entre las cuales las de vecino más próximo se presentan brevemente aquí:

Sea $\mathcal{X}^n = \{\underline{x}_1, \ldots, \underline{x}_n\}$ un conjunto de n observaciones - identificadas y sea $\underline{x}'_n \in \mathcal{X}^n$ la observación más cercana a $\underline{x}$, la observación a discriminar. La regla de vecino más próximo para clasificar a $\underline{x}$ es asignarlo a la población a la que pertenece $\underline{x}'_n$. Esta - regla es un procedimiento sub-óptimo, i.e., su uso generalmente conduce a cometer una tasa de error más grande que el mínimo posible, la correspondiente a regla de Bayes. Sin embargo puede demostrarse, ver Duda y Hart (1973), asintóticamente la tasa de error es a lo más el doble que la tasa de Bayes.

Una extensión directa de la regla del vecino más próximo es la de k vecinos más próximos*. Esta regla asigna $\underline{x}$ a la población a la que pertenecen más observaciones entre los k vecinos más próximos. En el capítulo III se estudiarán más a fondo las características en cuanto a tasas de error de las dos reglas anteriores.

Esto es tratando una sola muestra de una mezcla de las poblaciones, sin embargo, originalmente Fix y Hodges consideran $\{\underline{x}_{ij}$; $j = 1, \ldots n_i\}$ una muestra aleatoria de la i-ésima población, $i = 1, \ldots, k$. Se considera una función distancia de entre $\underline{x}_{ij}$ y $\underline{x}$ la observación a discriminar. Entonces se ordenan los valores de $d(\underline{x}_{ij}, \underline{x})$ $j = 1, \ldots, n_i$, $i = 1, \ldots k$, la regla R - NN (R vecinos más pró

* estas reglas están intimamente relacionadas con los con los procedimientos de estimación de densidades (ver Duda y Hart (1973)).

ximos) asigna x a la población P_i

Si $\frac{R_i}{n_i} = \max_j \frac{R_j}{n_j}$ donde R_i es el número de observaciones de P_i en las R observaciones más cercanas a $\underline{x}$, los casos de empate se resuelven de alguna manera preestablecida.

Matusita (1956), propone una regla de distancia mínima basada en la distancia de Matusita entre funciones de distribución empíricas. Das Gupta (1964), considera la regla de distancia mínima (con una distancia arbitraria) para el problema de k poblaciones y muestra la consistencia de tales reglas bajo condiciones apropiadas.

Glick (1969), desarrolla sistemáticamente las reglas "mejores en su clase" bajo la suposición de que la observación a ser clasificada proviene de una mezcla de m poblaciones. Considera la colección, $\mathcal{S}$, de particiones en m regiones ordenadas del espacio muestral, tales particiones se utilizan para asignar al individuo, $\underline{x}$, a alguna de las poblaciones. Define como regla mejor en su clase aquella (si existe) para la cual el estimador de Smith de la probabilidad de clasificación correcta alcanza su supremo en $\mathcal{S}$. Obtiene resultados acerca de las características del estimador citado en $\mathcal{S}$ y en una colección $\mathcal{S}'$ relativamente más restrictiva.

Otros autores (Aoyama (1950), Stoller (1954), Hudimoto (1956, 1957)) tratan casos más simples que el tratado por Glick y presentan resultados un tanto más débiles.

Un enfoque diferente al problema es el de diseñar reglas basadas en regiones de tolerancia. El germen de tal idea ya se encuentra en el trabajo de Fix y Hodges (1951), aunque es en el trabajo de Anderson (1966) donde se sugiere por vez primera. En ese trabajo Anderson sugiere definir bloques al asignar rangos a las observaciones vectoriales y posteriormente utilizar métodos univariados. Anderson también propone utilizar la muestra ensayo combinada para definir "bloques" y una observación $\underline{x}$ se considera como perteneciente a P_i si el bloque al cual $\underline{x}$ pertenece está definido por una mayoría de observaciones de P_i.

Quesenberry y Gessaman (1968), presentan un método para efectuar discriminación entre poblaciones en un espacio euclidiano con funciones de distribución contínuas. El espacio de decisiones (de cardinalidad $2^k - 1$ si son k las posibles poblaciones) que los autores llaman "parciales" y las probabilidades de error son variables aleatorias con distribución beta (ver Wilks (1962)). El procedimiento es el siguiente: suponiendo que se tiene una muestra $\{x_{j1}, \ldots, x_{jn_j}\}$ de la

población P_j $j = 1, \ldots, k$; sean $(\alpha_1, \ldots, \alpha_k)$ con $\alpha_i \in (0, 1)$ $i = 1, \ldots, k$ constantes y a_j denota la parte entera de $\alpha_j (n_j + 1)$. Usando la teoría de coberturas (Wilks (1962)) y la j-ésima muestra se construye una región de tolerancia no paramétrica A_j que contiene a_j bloques sobre el espacio euclidiano de medida $\mathcal{X}$ para la distribución. Cada conjunto A_j formado de la j-ésima muestra y su complemento $\bar{A}_j$ constituye una partición en dos conjuntos del espacio muestral $\mathcal{X}$. Una partición "producto" se forma de estas particiones como sigue:

Defínanse los conjuntos $S_{i_1}, \ldots, S_{i_s}$ y S_o como

$$S_{i_1, \ldots i_s} = \bar{A}_{i_1} \cdots \bar{A}_{i_s} A_{i_{s+1}} \cdots A_{i_k} \qquad s = 1, \ldots, k-1$$

$$S_o = (A_1 \cdots A_k) \cup (\bar{A}_1 \cdots \bar{A}_k)$$

donde $\{i_1, \ldots, i_s\}$ es un subconjunto propio de s elementos de $\{1, \ldots, k\}$ diferente del vacío, i.e., $s \geq 1$.

En base a lo anterior el procedimiento de discriminación d^* se define por

$$d^*(\underline{x}) = \begin{cases} \delta_{i_1, \ldots, i_s} & \text{si } \underline{x} \in S_{i_1, \ldots, i_s} \\ \delta_o & \text{si } \underline{x} \in S_o \end{cases}$$

donde

$\delta_{i1}, \ldots, i_s$ significa decidir que $P \in \{P_{i1}, \ldots, P_{is}\}$

δ_o significa reservarse juicio (no asignar a alguna población)

P es la población de procedencia de $\underline{x}$

Dado que se comete un error si la observación $\underline{x}$ proviene de la población P_j pero $\underline{x}$ cae en el conjunto $B_j = A_j \cap (\bigcup_{j=1}^{k-1} \bar{A}_{i1} \ldots \bar{A}_{is} A_{i_{s+1}} \ldots A_{ik})$ donde para la primera unión sólo se pueden escoger índices tomados de $\{1, \ldots, j-1, j+1, \ldots, k\}$ y $\{i_{s+1}, \ldots, i_{k-1}\}$ es en cada caso el conjunto restante. Como $B_j \subset A_j$ esto implica que $P(B_j) \leq P(A_j) \quad j = 1, \ldots k$ donde $P(A_j)$ es una variable Beta con parámetros $(a_j, n_j - a_j + 1)$ con media

$$\frac{a_j}{n_{j+1}} = \alpha_j + O\left(\frac{1}{n_j}\right) \qquad O\left(\frac{1}{n_j}\right) \geq 0$$

y varianza $a_j (n_j - a_j + 1) / (n_j + 1)^2 (n_j + 2)$. Si $\alpha_j (n_j + 1)$ es entero la media es α_j y la varianza $\alpha_j (1 - \alpha_j)/(n_j + 2)$

Aplicando la desigualdad de Chebyshev se tiene que

$$P_j(A_j) \xrightarrow{P} \alpha_j \qquad \text{si} \qquad n_j \to \infty \qquad \forall j = 1, \ldots k$$

Puede observarse que el procedimiento así definido tiene la propiedad de que cuando $\underline{x}$ tiene distribución F_j (distribución de la po

blación P_j) la probabilidad de error puede acotarse por una variable aleatoria P_j (A_j) con distribución beta y parámetros $(a_j , n_j - a_j + 1)$. La elección de las A_j no es única lo que hace que el procedimiento tenga gran flexibilidad, la cual puede aprovecharse para obtener ciertas propiedades deseables.

Cuando se consideran procedimientos basados en distribuciones, un límite natural lo proporcionan los procedimientos basados en distribuciones completamente conocidas (ver definición procedimientos consistentes en Quesenberry y Gessaman (1968)).

En la práctica puede suceder que se cuente con tan poca información acerca de las distribuciones que sea imposible sugerir una familia paramétrica probable, con la cual "calibrar" el procedimiento, i.e., evaluar comportamiento del método no paramétrico frente al óptimo que usa la información acerca de las distribuciones (así lo hacen los autores en dos ejemplos). También en muchos casos no se conocen procedimientos no paramétricos óptimos y aún conociéndolos es probable que procedimientos no paramétricos consistentes con respecto a los paramétricos óptimos estén asociados con particiones muy complicadas del espacio de observaciones, las cuales dificulta su uso en la práctica.

La selección de A_j , j = 1, ... k , determina el procedi - miento. Parece razonable seleccionar A_j de tal manera que la densi dad de la distribución para P_j produzca un valor pequeño de probabili dad de estar en A_j, tal vez ésto no conduzca a procedimientos ópti - mos pero para muestras grandes puede originar procedimientos buenos.

Los procedimientos propuestos tienen la propiedad de que pro porcionan un control de las probabilidades de clasificación equivocada para toda población con función de distribución contínua y para todo ta maño de muestra. Este control de probabilidades de error se debe al hecho de considerar decisiones de reserva y decisiones parciales, i.e., considerar a $\underline{x}$ como proveniente de un grupo de poblaciones. En caso de que las poblaciones estén muy "juntas" o que las regiones de tole - rancia se escojan de manera inadecuada la probabilidad del juicio de re serva (no asignación) puede ser alta.

En cuanto a la comparación que hacen los autores de sus pro cedimientos con los de Fix y Hodges (1951, 1952), estos últimos tie - nen la ventaja de consistencia (en el sentido de Quesenberry y Gessa man (1968) que a su vez es una ligera generalización del dado en Fix y Hodges (1951)), ante clases más grandes de distribuciones. Sin em bargo, los procedimientos de Quesenberry y Gessaman son más fáciles de usar en la práctica ya que una vez que se tienen las muestras de

ensayo (que ellos llaman de calibración) se determinan las regiones de clasificación y solamente ha de determinarse en que región cae la observación a discriminar, lo que para distribuciones uni o bidimensionales puede ser llevado a cabo fácilmente aún por personal no especializado.

Pelto (1969), trata el caso de dos poblaciones de las cuales supone tiene muestras de tamaño n_1 y n_2 respectivamente. Propone un procedimiento basado en la inserción de una hiperesfera de radio r en el espacio de caracteres, esencialmente su idea es estimar las densidades de probabilidad mediante el conteo de "puntos" identificados - que se encuentran dentro de la citada hiperesfera, el radio r de ésta se fija de manera que se minimice la pérdida esperada estimada de la regla de decisión. Este procedimiento puede operarse de dos formas, una es en base a muestras de origen conocido y proceder a clasificar (discriminar) una serie de observaciones cuyo origen es desconocido, i.e., utilizar una muestra supervisada para efectuar la discriminación; en la otra forma se supone que existe un intervalo de tiempo entre cada arribo de observaciones con origen desconocido de tal manera que se pueda verificar la población de la cual proviene la observación $\underline{x}$, después de asignarla y así esta observación aumenta la información y por lo tanto se tiene una muestra post-supervisada (ver

capítulo I de este trabajo).

Anderson y Benning (1970), evitan en cierta manera la dificultad que significa la información insuficiente acerca de las distribuciones involucradas en el problema de discriminación, mediante el uso de técnicas de agrupamiento (clustering) para obtener información sobre los cocientes de verosimilitud. Patrick y Fisher (1970) usan las regiones de tolerancia para estimar las funciones densidad de probabilidad y poder usarlas en reglas de inserción. En su trabajo sobre una comparación de algunos procedimientos de discriminación con variables multidimensionales, Gessaman y Gessaman (1972), también sugieren algunos procedimientos basados en bloques equivalentes estadísticamente, los autores llevan un estudio de ellos mediante métodos Monte Carlo.

Kendall y Stuart (1966), sugieren un método de discriminación basado en la suposición de que una vez que ha sido empleada una variable en el análisis, todo el poder discriminatorio o información proporcionada por esa variable se ha agotado. El procedimiento que proponen para discriminar entre dos poblaciones es un método basado en el examen secuencial de las distribuciones de frecuencia marginal de los individuos no clasificados, dividiendo el espacio p-dimensional de observaciones en regiones mutuamente exclusivas.

Posteriormente Richards (1974) demuestra que la suposición de Kendall y Stuart (1966) no es necesariamente cierta, ya que aun - que es cierto que no vale la pena re-examinar una variable ya incluí da en la regla de discriminación cuando se usa ya sea la función li - neal o la cuadrática usuales, el criterio no se aplica a un método co mo el sugerido por Kendall y Stuart. Richards (1974), sugiere un re finamiento que consiste en examinar casi todas las variables, incluyen do aquellas que previamente se han empleado en cada paso, la única - variable la cual no debe ser examinada en un paso dado es la varia - ble usada en el paso inmediatamente anterior. Además propone una - extensión, aplicable a los casos de duda, la cual consiste en examinar las distribuciones de frecuencia conjunta de los casos en los cuales per sista la indecisión, después de haber utilizado el análisis de Kendall y Stuart. El objetivo del examen es detectar separaciones que sean con sistentes con la regla de discriminación basada en el análisis de distri bución de frecuencias marginales. Tal extensión no necesita restringir se a las distribuciones de frecuencia conjunta de dos variables sino que se pueden utilizar más variables.

Das Gupta (1962, 1964), sugiere el uso de las estadísticas em pleadas en procedimientos no paramétricos de pruebas de rangos. Con sidera en la discriminación de dos poblaciones el uso de la estadística

de Wilcoxon, siendo la regla de decisión considerar que las observaciones a discriminar provienen de P_i, $i = 1, 2$, si el valor absoluto de W_i, estadística de Wilcoxon basada en muestras de P_i, $i = 1, 2$ y la muestra a clasificar, es el menor de los dos valores calculados. Hudimoto (1964), utilizando también la estadística de Wilcoxon, modifica la regla de decisión considerando W_i en vez de $|W_i|$ cuando $F_1(x) \geq F_2(x)$ $\forall$, proporciona una cota para la probabilidad de clasificación correcta y posteriormente en otro trabajo Hudimoto (1965), estudia la regla en el caso de ocurrir empates. Kinderman (1972), discípulo de Das Gupta, propone una clase de pruebas basadas en estadísticas de rangos lineales. Calcula la eficiencia relativa asintótica (en el sentido de Pitman) de la regla de clasificación específica para el caso de dos poblaciones. Govindorajulu y Gupta (1972), consideran estadísticas similares para el caso de k poblaciones donde los tamaños de muestra de ensayo de cada población pueden ser diferentes y obtienen una regla que controla el promedio (con respecto a probabilidades a priori de procedencia conocidas), de las probabilidades de clasificación correcta.

En cuanto a un enfoque de Bayes empírico, Johns (1961) considera el problema de dos poblaciones y la variable indicadora I se considera aleatoria y tal que las dos categorías son definidas por una

partición del espacio de resultados de I . El autor propone una regla basada en una muestra de ensayo de tamaño N y demuestra que el riesgo de Bayes de tal regla tiende al riesgo mínimo de Bayes, i.e., bajo conocimiento total de la distribución de (X, I) (ver capítulo I de este trabajo), Johns (1961) trata los siguientes casos: 1) muestra de ensayo supervisada y X del tipo discreto , 2) muestra de ensayo supervisada y X del tipo contínuo y 3) muestra de ensayo postsupervisada y X del tipo discreto.

Gessaman y Gessaman (1972), presentan un método de discriminación, sin embargo, la parte esencial de su trabajo es la comparación de métodos de discriminación por lo que es referido al capítulo III de este trabajo.

Habbema et al. (1974), tratan el caso de discriminación entre dos poblaciones, para la solución del problema plantean un modelo de teoría de decisiones el cual admite decisiones no conclusivas, i.e., casos de duda. Bajo el supuesto de que se desconocen completamente las distribuciones en cada una de las dos poblaciones estiman las densidades, utilizando el estimador propuesto por Sebestyen (1962)y Specht (1967) y llevan a cabo la asignación de un individuo mediante el cociente de verosimilitud estimado, aplicando a éste la regla óptima diseña

da para minimizar la pérdida esperada correspondiente a la acción de asignar el individuo a la población P_1 ó P_2 ó no asignarlo.

El objetivo de presentar un desarrollo detallado para algunos de los métodos es hacer explícitas las suposiciones que en muchas ocasiones pasan inadvertidas. Además de esto, se considera importante la presentación de soluciones poco conocidas desarrolladas mediante el enfoque bayesiano o bajo el de procedimientos no paramétricos, pues proporcionan una alternativa que en algunos casos puede ser más adecuada que los métodos tradicionales de los cuales se abusa un poco.

BIBLIOGRAFIA

CAPITULO II.-

II.1.1.

Fisher	(1936)
Wald	(1944)
Anderson	(1951) , (1958)
Sitgreaves	(1952)
Kabe	(1963)
Cramér	(1946)
Press	(1972)
Kshirsagar	(1966)

II.1.2.

Anderson y Bahadur	(1962)

II.1.3.

Von Mises	(1945)
Anderson	(1958)
Blackith y Reyment	(1971)
Seal	(1964)
Matusita	(1971) , (1973) , (1967)

II.2

Geisser	(1964)
Press	(1972)
Dunsmore	(1966)

II.3

Matusita	(1956)
Chernoff	(1959)
Wesler	(1959)
Cochran y Hopkins	(1961)
Hills	(1966)
Bunke	(1966)
Glick	(1969) , (1973)
Goldstein y Rabinowitz	(1975)
Bahadur	(1961)
Solomon	(1961)
Hills	(1967)
Martin y Bradley	(1972)
Cooper	(1962) , (1963) , (1965)
Battacharya y Das Gupta	(1964)
Day y Kerridge	(1967)
Anderson	(1972)

Kendall (1966)

Marshall y Olkin (1968)

Chang y Afifi (1974)

Shumway y Unger (1974)

Krzanoski (1975)

II.4

Glick (1972) , (1969)

Fix y Hodges (1951) , (1952)

Duda y Hart (1973)

Matusita (1956)

Das Gupta (1964)

Aoyama (1950)

Stoller (1954)

Hudimoto (1956) , (1957)

Anderson (1966)

Quesenberry y Gessaman (1968)

Pelto (1969)

Kendall y Stuart (1966)

Richards (1974)

Anderson y Benning (1970)

Patrick y Fisher (1970)

Gessaman y Gessaman (1972)

Das Gupta (1962) , (1964)

Hudimoto (1964) , (1965)

Kinderman (1972)

Govindarajulu y Gupta (1972)

Johns (1961)

Habberna et al. (1974)

III. ESTUDIO DE LAS CARACTERISTICAS DE LAS SOLUCIONES.-

III.1 Estimación de las probabilidades de clasificación correcta e incorrecta.

Este problema ha atraído últimamente la atención de un gran número de investigadores, aunque siempre está presente en relación a reglas de decisión basadas en muestras i.e., cuando se desconocen total o parcialmente los parámetros y aún la distribución de las poblaciones bajo las diversas alternativas. El problema surge debido a que las probabilidades de clasificación (correcta o equivocada) dependen de la distribución exacta de la variable aleatoria considerada, i.e., dependen de la forma de la función y de los parámetros que se desconocen y por otra razón hay que estimar las probabilidades de clasificación.

Debe señalarse que existen varias medidas de probabilidades

de clasificación equivocada las cuales se consideran generalmente, que son:

1) La probabilidad de error (clasificación equivocada) o probabilidad de Bayes que resulta cuando se tiene conocimiento completo de las densidades, para las poblaciones involucradas, con las cuales construir la regla de decisión óptima o de Bayes. Se la denotará por P_e^B.

2) La probabilidad de error para el caso anterior pero en vez de utilizar la regla óptima se utiliza alguna otra. Se la denotará P_e^o.

3) La probabilidad de error en futura utilización de la regla diseñada, también conocida como probabilidad condicional, se le denotará como $\hat{P}_e$.

4) El valor esperado de la anterior $E\{\hat{P}_e\}$, tanto $\hat{P}_e$ como $E\{\hat{P}_e\}$ aproximan la llamada probabilidad "real" de error P_e (que es aquella en la que no se hace distinción entre P_e^B y P_e^o).

5) La probabilidad de error "aparente" P_e^* que se puede obtener estimando las distribuciones poblacionales o sus parámetros y substituyendo estos valores estimados en la expresión de la probabilidad de error. También se puede obtener al usar el clasificador

o función discriminante en un número infinito de observaciones que provengan de distribuciones cuyos parámetros sean los estimados en vez de los reales, generalmente este valor esta sesgado optimístamente.

6) La probabilidad de que una observación perteneciente a la clase i sea clasificada como de la j se denota en general como $P_{e_{j|i}}$. $i, j = 1, \ldots, k \quad i \neq j$.

7) Para la probabilidad de error condicionada en una clase se tiene que $P_{e|c_i} = \sum_{\substack{j=1 \\ j \neq i}}^{k} P_{e_{j|i}}$ que es la probabilidad de que una observación de la clase i se clasifique mal.

8) La probabilidad de error dada una observación particular $P_{e|x}$, que se relaciona con P_e mediante la siguiente identidad $P_e = \int P(x) P_{e|x} \, d X$, donde $P(x)$ es la distribución incondicional o una mezcla de las distribuciones involucradas.

Se han presentado trabajos que tratan sobre las relaciones entre algunas de las probabilidades anteriores, por ejemplo Sorum (1972), concluye que los problemas de estimación de $E\{P_e\}$ y P_e pueden considerarse prácticamente uno solo y $\hat{P}_e$ ha de considerarse separadamente, pues este último, a diferencia de los dos anteriores, es también -

una función de la muestra. Posteriormente en 1968 (ver Sorum (1968)) se avoca a estimar P_e^B y $E(P_e)$.

Mc Lachlan (1974) utiliza la expansión del error cuadrático - medio asintótico (ECMA), encuentra relaciones interesantes entre los estimadores de P_e^B, P_e y $E(P_e)$ que además mejoran los de Sorum.

Cabe comentar que si las muestras son "suficientemente grandes" se pueden dividir en dos subconjuntos y utilizar uno de ellos en el diseño del procedimiento de discriminación, las observaciones restantes se utilizan para estimar las probabilidades de error. Sin embargo, la situación que parece ocurrir más a menudo es la de tener - muestras relativamente pequeñas por lo que la aplicación de la regla- citada no da buenos resultados, en ciertos casos se utiliza simplemente para tener una idea aproximada de las probabilidades de clasificación equivocada, pero la regla de decisión se construye posteriormente usando todas las observaciones, por lo que en este caso, se tiene - la desventaja de que los resultados son un tanto engañosos.

Entre los métodos que se han propuesto para estimar las probabilidades de error en el caso de dos poblaciones se tienen los siguientes:

Los estimadores de Fisher para la probabilidad de errores de clasificación de la regla minimax que consiste en substituir en la expresión que proporciona ese valor, i.e. $\Phi(-\frac{\Delta}{2})$, en lugar de Δ el estimador muestral D*. Este método depende de la suposición de normalidad, además es sesgado pues como D sobreestima Δ (ver Kshirsagar (1972)) entonces, el estimador obtenido subestima la probabilidad de error de clasificación. Para muestras pequeñas generalmente produce resultados engañosos.

Estimador por conteo o de reasignación para errores de clasificación equivocada, sugerido por Smith (1947), para el caso de discriminar entre dos poblaciones, que consiste en tomar una muestra de la población P_1 y ver con que frecuencia se asignan a la población P_2 en base a la regla de decisión diseñada de antemano. El mismo Smith (1947) señala que si se usa la misma muestra que se utilizó en la construcción de la regla de decisión para posteriormente estimar las probabilidades de error (lo cual sugiere él mismo), el error estándar de los estimadores no será el valor que correspondería a una media de variables bernoulli. Este método puede usarse para cualquiera que sea la distribución de las observaciones. Tiene la desventaja de que subestima las probabilidades de clasificación equivocada (por su misma construcción) y debe usarse con mucho cuidado, en particular cuando se

* donde $D^2 = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' S^{-1} (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$ la distancia muestral de Mahalanobis.

tengan muestras pequeñas.

John (1961), deriva la distribución de los errores de clasificación equivocada para la regla de decisión minimax y además obtiene - la media de estos bajo la suposición de que la matriz común de- covarianza se conoce, obtiene resultados similares cuando la - constante contra la cual se compara para llevar a cabo la deci - sión es diferente de cero. Presenta aproximaciones para el caso en el cual se desconoce la matriz de covarianza y los tamaños de muestra de cada población son grandes. Posteriormente el mismo autor en John (1963), estudia las probabilidades de clasificación - equivocada correspondientes a las estadísticas de clasificación R - propuesta por Rao (1954) y una de las propuestas por Anderson - (1951, 1958); en (1964) John trata el mismo problema que en su trabajo de (1961), pero bajo la suposición de que la media de la población de la cual procede la observación por discriminar es diferente de $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ las medias correspondientes a las poblaciones propuestas como alternativas de procedencia.

Dunn y Varady (1966), usando métodos de simulación Monte Carlo, investigan la relación entre la probabilidad de clasificación correcta usando la función discriminante lineal muestral y el estimador de tal probabilidad el cual puede obtenerse estimando la distancia de

Mahalanobis entre las dos poblaciones. Presentan gráficas para obtener intervalos de confianza conservadores para las probabilidades de clasificación correcta dada la probabilidad estimada. Utilizando los mismo métodos obtienen que, para un número dado de variables usadas para calcular la función discriminante, el estimador de la probabilidad es más satisfactorio (en términos de longitud del intervalo de confianza construído), a medida de que el tamaño de muestra se incrementa. Además para un tamaño de muestra fijo el estimador de la probabilidad de clasificación correcta se vuelve menos satisfactorio al incrementarse de dos a diez el número de variables. Se presentan tablas que muestran ciertos percentiles en las distribuciones de estas relaciones.

Hills (1966) obtiene la distribución de $\hat{P}_{e_{2|1}}$ para la regla minimax con parámetros estimados, compara las esperanzas de $\hat{P}_{e_{2|1}}$ y $P_{e_{2|1}}$ también para la regla minimax y las del estimador de Smith mediante expresiones exactas y cálculos numéricos. También menciona los resultados en ese entonces inéditos de Lachenbruch (1965). Obtiene que la $E\{P_e \text{ minimax, parámetros conocidos}\} < E\{P_e \text{ minimax parámetros estimados}\}$ Entre lo más interesante del artículo esta la discusión que de él hacen estadísticos de renombre como P. Armitage, W. G. Cochran, D. V. Lindley y otros más que presentan -

atrayentes sugerencias.

Lachenbruch (1967) propone otro método para estimar los errores de clasificación equivocada. El método puede ser usado para cualquiera que sea la distribución de las observaciones y produce estimadores casi insesgados para las probabilidades $P_{2|1}$ y $P_{1|2}$. Lachenbruch señala que experimentos muestrales indican que puede usarse para obtener intervalos de confianza aproximados para las probabilidades de clasificación equivocada. El método puede ser usado también en problemas de discriminación no paramétrica.

Lachenbruch y Mickey (1968) evalúan mediante simulación varios métodos para estimar tasas de error en análisis discriminante. Estos autores obtienen que los dos métodos más comúnmente usados son significativamente inferiores que otros métodos nuevos que ellos proponen.

Entre los métodos que ellos evalúan se encuentran el de Fisher y el de Smith antes mencionados, también consideran otro método inspirado en el que se usa cuando n_1 y n_2 no son grandes con respecto a p, sigue el razonamiento del de Fisher pero el estimador de Δ^2 que se usa es el insesgado.

$$D^{*2} = \left\{ D^2 - (m-2)\, m\, p / n_1\, n_2\, (m-p-3) \right\} \times (m-p-3)/(m-2)$$

donde $m = n_1 + n_2$, que sin embargo, tiene la desventaja de que D^{*2} frecuentemente es negativo. Lachenbruch y Mickey proponen en cambio

$$DS = (m-p-3)\, D^2 / (m-2)$$

que corrige parcialmente el sesgo en D^2 y es consistente para Δ^2 y es no negativo para el rango de valores de interés. El método basado en DS parece tener algunas ventajas sobre el de Fisher.

Los autores presentan también un método(*) basado en la distribución asintótica para las probabilidades de clasificación equivocada para la estadística $W(\underline{x})$ de Anderson que es la siguiente:

$$P_{2|1} = P\left[W(\underline{x}) < 0 \;\middle|\; \underline{x} \in P_1 \right] = \Phi\left(-\frac{1}{2}\Delta\right) + \frac{a_1}{n_1} + \frac{a_2}{n_2} + \frac{a_3}{m-2}$$

$$\frac{b_{11}}{n_1^2} + \frac{b_{12}}{n_2^2} + \frac{b_{12}}{n_1 n_2} + \frac{b_{13}}{n_1(m-2)} + \frac{b_{23}}{n_2(m-2)} + \frac{b_{33}}{(m-2)^2} + O_s$$

(III.1)

un resultado análogo se tiene para $P_{1|2}$. Okamoto ha tabulado los valores de las a's y de las b's para algunos casos especiales, así se pueden estimar $P_{2|1}$ y $P_{1|2}$ substituyendo en (III.1) Δ por D. Una modificación consiste en usar el estimador insesgado de Δ^2 dado anteriormente, tal modificación es llamada OS.

(*) Con base en el trabajo de Okamoto (1963)

Los mismos autores proponen un método que como el método de Smith hace uso de todas las observaciones pero que no tiene las desventajas de un sesgo serio. En este método $W(\underline{x})$ no se calcula de todas las $n_1 + n_2$ observaciones, sino que se omite una de ellas ya sea que pertenezca a P_1 o a P_2 y $W(\underline{x})$ se obtiene de $n_1 + n_2 - 1$ observaciones. Tal método presenta la ventaja de que tiene sesgo pequeño pero para distribuciones discretas tiene una varianza mayor que el método de resubstitución de Smith, como demuestra Glick, este problema también ha sido estudiado por Lunts y Brailovskiy (1962). En Kshirsagar (1972) se presenta el desarrollo computacional, aún cuando los autores demuestran que sólo se necesita un reajuste matricial (ver Bartlett (1952)) para llevar a cabo los cálculos, también para el caso de discriminación mediante métodos de distribución libre, continúa existiendo el problema de cálculo, sin embargo, hay que hacer notar que este trabajo aumenta si los tamaños de muestra son grandes.

En el mismo artículo proponen el método $\bar{U}$ que consiste en calcular las medias $\bar{D}_l$, $l = 1, 2$ de las estadísticas de clasificación W_j^* calculadas en base a $n_1 + n_2 - 1$ observaciones para $j = 1, \ldots, n_1, n_1 + 1, \ldots n_2$ dentro de cada muestra y dividirlas entre la varianza muestral S_{D_l} de las n_l cantidades $W_j^*(\underline{x}_j)$ $l = 1, 2$ y en base a esto calcular

$$P_{2|1} = \Phi(-\bar{D}_{1}/S_{D_1}) \qquad y \qquad P1|2 = \Phi(-\bar{D}_{2}/S_{D_2})$$

Como puede observarse este método también involucra gran - trabajo computacional. Lachenbruch y Mickey (1968) efectúan una evaluación comparativa de los métodos citados anteriormente basada en experimentos de simulación Monte Carlo. Ellos llegan a la conclusión de que tanto el método de Smith como el de D son relativamente malos. Los mejores métodos son los denotados como OS, U y $\bar{U}$ aunque el método O tuvo un buen desempeño general. Ahora bien, si puede suponerse normalidad aproximada los métodos OS y $\bar{U}$ son buenos, el U no saca ventaja de la suposición de normalidad. Para muestras de tamaño pequeño los métodos U y $\bar{U}$ se deben usar y no el OS.

Cochran (1968), en un comentario al artículo reseñado anteriormente señala los puntos buenos de tal artículo aunque erróneamente se refiere al método U como una aplicación de la técnica de "jackknife" - (de la cual objeta el nombre por el desconocimiento del significado original de tal palabra). Cochran lleva a cabo algunas simulaciones de - las cuales presenta resultados de las comparaciones que él efectúa, observándose que en general los métodos OS y DS tienden a resultar ligeramente mejores de lo que sugieren Lachenbruch y Mickey. Propone además que sería útil mayor investigación con respecto a los méto

dos más prometedores bajo poblaciones no normales.

Lissack y Fu (1972) proponen un método que llaman método F, encuentran que presenta menor sesgo y varianza que el método U, este resultado lo obtuvieron para experimentos con datos provenientes de poblaciones normales.

Toussaint (1969, 1970 junto con Donaldson) propone el método π, el cual consiste en separar pequeños subconjuntos de observaciones muestrales de ensayo en vez de una sola unidad y probar en tales observaciones el clasificador diseñado en base a los restantes, para de esta manera estimar el error a cometer con el funcionamiento futuro. En el caso de que los subconjuntos separados cuenten con un solo elemento este método coincide con el U antes citado.

Posteriormente Toussaint (1974), presenta un estimador para la probabilidad de error en funcionamiento futuro de una regla de clasificación, el cual es una combinación convexa del valor del estimador mediante el método π y el valor obtenido mediante el método R de Smith. Al menos empíricamente puede demostrarse que este estimador es esencialmente insesgado, tiene como ventaja el que parece proporcionar buenos resultados aún con muestras de ensayo de tamaño pe

queño (ver Foley (1972)) y requiere un trabajo computacional menor que para el método U.

McLachlan (1974 Biometrics) sugiere otra técnica para estimar las tasas de error en discriminación, tal método que se designará como método M se basa exclusivamente en las muestras de ensayo y además de tener un sesgo pequeño con respecto al de otros estimadores (su sesgo es el término de tercer orden en la expansión de la esperanza de $P_{2|1}$ dada por Okamoto (1962)), requiere menor esfuerzo computacional que el método U. En comparación con el método OS se encuentra mediante simulación Monte Carlo y bajo el criterio de Error Cuadrático Medio Asintótico que el estimador compite aceptablemente. El estimador proporcionado por este método puede expresarse explícitamente como una función de D^2, p, n_1 y n_2. Las desventajas que presenta se deben principalmente a que su construcción se apoya en el hecho de que las poblaciones son normales. Su uso esta restringido, pues no es recomendable para casos en los que se tiene un valor pequeño de D^2 (sea $D^2 < 1$) cuando n_1 y n_2 no son relativamente grandes con respecto a p. El estimador de $P_{1|2}$ se obtiene fácilmente pues basta intercambiar los papeles de n_1 y n_2 en la expresión para $P_{2|1}$. Posteriormente McLachlan (1974) basado en el estimador anterior propone uno para la probabilidad óptima de error.

III.2 Selección de variables y problemas relacionados.

Cochran (1964) toca este tema al interesarse en el problema de predecir la probabilidad de clasificación equivocada (al considerar dos poblaciones), a partir de las probabilidades dadas por las variables usadas individualmente. El resultado principal que obtiene es que "es generalmente seguro excluir de un discriminante, antes de calcularlo, un grupo de variables cuyas potencias discriminatorias individuales sean bajas, excepto por alguna que tenga correlaciones negativas con la mayoría de las buenas discriminadoras". Otro resultado es que el funcionamiento de la función discriminante puede ser predicho satisfactoriamente a partir del conocimiento de las potencias discriminatorias individuales y el coeficiente de correlación promedio, dado que bajo la suposición de que todo par de variables tiene una correlación igual al promedio, produce buenos resultados en la predicción del funcionamiento de la función discriminante lineal de Fisher.

Weiner y Dunn (1966), estudian el comportamiento de tres reglas de selección de variables frente a la selección aleatoria. Se supone que el número de variables se ha reducido tanto como se ha podido utilizando el conocimiento del problema, sin embargo, se requiere disminuir aún más el citado número, para lo cual se requiere un procedimiento de selección de variables. Enfocando la situación de dos poblaciones estudian tres métodos: uno basado en seleccionar las variables que muestran las más grandes diferencias "studentizadas" * entre las observaciones de las muestras de ensayo, en el segundo la selección se lleva a cabo considerando los coeficientes normalizados más grandes en la función discriminante y el tercero selecciona las variables mediante un programa de selección de variables por pasos originalmente diseñado para tratar el problema en casos de regresión múltiple. Los tres métodos se comparan con la selección aleatoria de variables para establecer si son relevantes y notar diferencias entre ellos. El procedimiento de discriminación (clasificación) que los autores usan es el basado en la estadística W de Anderson.

Las comparaciones se realizan aplicando el método de discriminación basado en diez, cinco y dos variables seleccionadas de veinticinco de ellas, observando el número de observaciones mal clasificadas en las propias muestras ensayo y después aplicando las funciones

---

* Diferencias entre medias de una variable en cada población suponiendo la varianza desconocida.

discriminantes a otros individuos. El número de individuos por clasificar cambia según el estudio (los autores consideran cinco estudios diferentes) que se trate. Se presentan los resultados de tales comparaciones para las muestras ensayo y para las observaciones "nuevas" se debe considerar que los resultados para las muestras ensayo han de tomarse con reservas pues es más notable el efecto que sufre la potencia discriminatoria de W a medida que se utilizan menos variables, ya que se reasignan las observaciones que se utilizaron para disenar la función discriminante respectiva. En cambio para las "nuevas" observaciones los resultados para 10 y 5 variables muestran que el mejor método de selección (a grosso modo) es el basado en la selección por pasos (paso a paso), sin embargo, no difiere significativamente en el sentido estadístico a un nivel del 0.1 de ninguno de los otros tres (aún del aleatorio), para dos variables, el que numéricamente se comporta mejor es el basado en las diferencias de t pero no es significativamente diferente del método paso a paso y el de coeficientes de la función discriminante. El método de selección al azar consistentemente fue el peor como era de esperarse. Cabe comentar que tal vez se debió tratar con un nivel de significancia mayor que al 10% ya que sería deseable conocer diferencias que existan entre los métodos, no obstante que sean pequeñas. Queda aún mucho trabajo por hacer al respecto.

Elashoft et al. (1967) enfocan el problema de selección de variables en el caso en que estas son dicotómicas, encuentran que los resultados de Cochran con respecto al mismo problema no son aplicables en este caso. Los autores examinan en detalle el caso de seleccionar dos variables de entre p de ellas, obteniendo como resultado el que la correlación positiva puede incrementar la discriminación mientras la negativa la decrementa.

Francis (1967) trata el caso de seleccionar v variables de entre p en el caso de dos poblaciones normales, supone que se tienen dos muestras de ensayo. Utiliza un método paso a paso para seleccionar las v variables en base a las muestras ensayo, sin embargo, señala que las variables así relacionadas bien pueden ser las de mayor poder discriminante para las muestras ensayo pero en el caso de discriminar nuevas observaciones pueden presentar el efecto de regresión, i.e., su funcionamiento regresa hacia el promedio. Escoge la probabilidad de clasificación equivocada como medida de la habilidad para discriminar, esta es entonces una función de la distancia poblacional de Mahalanobis Δ^2 para las v variables seleccionadas, que a su vez es una función de las diferencias entre las medias poblacionales, δ_i, de las variables seleccionadas. De la distribución conjunta a posteriori para las v diferencias mencionadas se puede encontrar el valor esperado a posterio

ri de Δ^2, entonces se tiene un estimador de Δ^2 (y de la probabilidad de clasificación equivocada) el cual esta libre del efecto de regresión. El autor muestra como encontrar la distribución conjunta a posteriori de las δ_i's (en particular de las v variables seleccionadas). Esto no es posible en general sino sólo en caso de que la matriz de covarianza de las δ_i's pueda aproximarse por una matriz de cierta forma.

Hills (1967) considera también el problema de selección de variables para el caso en que estas son discretas, presenta tres procedimientos, dos paso a paso y uno mixto que incluye tanto la técnica paso a paso, como la de vecino más próximo (ver comentario sección II. 4, de este trabajo). Propone dos medidas de discriminación para poder comparar las reglas basadas en los subconjuntos de variables formados por los procedimientos de selección, tales medidas satisfacen las cuatro propiedades que debe poseer una medida de discriminación. Como se comentó anteriormente, en el trabajo se presentan extensiones para variables con más de dos estados. Se refiere también a otros procedimientos de selección de variables que son el de Belson (1959) y el de Mac Naughton - Smith (1963), los cuales son del tipo paso a paso no restringidos (aunque el de Belson fue sugerido originalmente para producir muestras apareadas). Se refiere al modelo logístico sugerido por

Cox (1966) y sugiere que los procedimientos paso a paso presentados en Hills (1967) pueden ser útiles para seleccionar variables a incluir en el modelo logístico y para sugerir posibles términos de interacción.

Hortan et al. (1968) analizando datos agrupados, sugieren la necesidad de un método que combine las ventajas del análisis multivariado con la economía en relación al número de variables requerido. Para alcanzar este objetivo utilizan dos programas de computadora, el primero de ellos se ocupa de la selección de un subconjunto de las variables originales tal que la varianza entre grupos restante no es significativa, el segundo programa realiza un análisis canónico (que como ya se vió en el capítulo II, esta íntimamente relacionados con la discriminación entre varias poblaciones). Tal análisis permite identificar variables que mediante las técnicas univariadas no presentaban diferencias significativas entre poblaciones pero que consideradas en conjunto contribuyen significativamente a la determinación de diferencias entre poblaciones.

En el caso de discriminación entre dos poblaciones basada en muestras de ensayo de tamaños n_1 y n_2 respectivamente. Boullion et al. (1975) estudian las relaciones entre la estimación de la probabilidad de clasificación equivocada y la selección de variables mediante -

métodos de simulación Monte Carlo, su trabajo tiene dos objetivos: el primero es estimar la probabilidad esperada de clasificación equivocada la cual ocurriría en muestras repetidas de ensayo de tamaño $n_1 = n_2 = n$ y segundo, estudiar el problema de seleccionar el "mejor" subconjunto de variables para discriminación. Ya que la probabilidad de clasificación equivocada es asintóticamente dependiente de Δ^2 (la distancia de Mahalanobis) y no de p el número de variables, entonces con tal - que la separación entre poblaciones permanezca constante (medida en términos de Δ^2), es factible que un número menor que p variables sea usado con fines discriminatorios sin disminución en la exactitud del procedimiento. En el caso de poblaciones normales con matrices de covarianza Σ_1 y Σ_2 diferentes la medida de divergencia sugerida por Kullback (1952, 1968 p 190) queda como:

$$J = \frac{1}{2} tr\,(\Sigma_1 - \Sigma_2)(\Sigma_2^{-1} - \Sigma_1^{-1}) + \frac{1}{2} tr\,(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$$

que para el caso en que $\Sigma_1 = \Sigma_2$ se reduce a la Δ^2 de Mahalanobis. En situaciones reales cuando $\Sigma_1 \neq \Sigma_2$ es más fácil usar Δ^2 en vez de J, utilizando como matriz de covarianza común $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$, esta substitución se usa en el trabajo de Boullion et al. (1975) y estos autores concluyen que es razonable.

Boullion et al. efectuaron las simulaciones para diferentes va

lores de Δ^2, p y n comparándolos contra aquellos en que se conocen $\underline{\mu}_1$, $\underline{\mu}_2$ y Σ; para analizar la factibilidad de seleccionar un número menor de variables en base a las cuales discriminar y las consecuencias que esto tiene en cuanto a la separación de las poblaciones. También se estudia la equivalencia entre Δ^2 y J como criterios para la selección de variables. Las conclusiones que se obtienen para el caso que estudian detenidamente es que independientemente de la medida de separación seleccionada se obtiene muy poco incremento en la máxima separación al considerar más de cierto número de variables. Además se observa que cualquiera de las dos medidas de separación es un criterio adecuado para la selección del subconjunto de variables deseado.

Con respecto a la probabilidad de clasificación equivocada (usando la estadística W de Anderson) se obtiene que $P_{2|1}$ es una función decreciente de Δ^2 (fijos n y p) y de n (fijos Δ^2 y p). Para n fija los resultados indican que $P_{2|1}$ es una función creciente de p.

La conclusión final es que para tamaños de muestra pequeños puede escogerse un subconjunto de variables de tal manera que produzca mejores resultados que el total de ellas y que la selección de variables se puede lograr mediante la consideración de la separación entre poblaciones.

III.3 Comparación de algunos métodos de discriminación.

Este tema ha interesado hasta últimas fechas a los estadísti - cos. La necesidad de estudiarlo nace principalmente del uso extendi do que se ha hecho y se hace de la función discriminante lineal de Fi sher (FDL), por lo que las comparaciones que se han realizado com - prenden a la FDL y a otros métodos de discriminación.

Gilbert (1968) en un trabajo que forma parte de su tesis docto ral, presenta los resultados de la comparación de cinco métodos para discriminar entre dos poblaciones en base a variables vectoriales, $\underline{x}$, tales que la distribución marginal de cada componente es Bernoulli, - considera una variable Bernoulli adicional, y, cuyos valores caracteri zan a las poblaciones i.e., si es cero la observación pertenece a una población y si es uno a otra. El problema de discriminación se redu ce a predecir el valor de y a partir del vector $\underline{x}$. Los resultados que obtiene basados en los criterios del coeficiente de correlación y la pro

babilidad de clasificación equivocada son que para una dimensión del vector de observaciones igual a 6 y varios tamaños de muestras de ensayo, existe poca diferencia entre los procedimientos lineales (entre ellos la FDL) usados pero todos son ligeramente superiores al sugerido por Cochran y Hopkins (1961). El resultado más importante es - que, los experimentos Monte Carlo llevados a cabo y la evaluación de la FDL considerando conocidos los parámetros, sugieren que la pérdida o costo debido a usar la FDL en vez de otro procedimiento diseñado para discriminación con el tipo de variables considerado es muy pequeño para ser de importancia. Además el comportamiento de la FDL se mantiene bastante estable cuando el número de componentes del vector de observaciones aumenta, y es muy probable que produzca una regla de discriminación superior a casi todos los métodos considerados. Revo (1970) estudia el funcionamiento de varias reglas para tratar variables discretas ordenadas y encuentra que la FDL se comporta bastante bien en relación a tales reglas consideradas.

Posteriormente Gilbert (1962) analiza el efecto de heterocedasticidad sobre la FDL comparada con la función cuadrática óptima[1], suponiendo conocidos los parámetros para el caso de discriminación entre dos poblaciones, se estudia el coeficiente de correlación entre ambas - reglas y las probabilidades de clasificación equivocada respectiva.

1. Obtenida al utilizar el cociente de verosimilitud para diseñar la regla de discriminación

Sólo se considera el caso en que $\Sigma_2 = d\Sigma_1$, d una constante (que corresponde a la situación en la que las correlaciones son iguales y las únicas distintas son las varianzas y covarianzas, las cuales difieren únicamente en una constante de proporcionalidad).

Los resultados se pueden resumir en que la función discriminante cuadrática y la FDL muestran una concordancia adecuada solamente para un rango moderado de valores de d y con cierta separación (medida en términos de la Δ^2 de Mahalanobis), tal concordancia disminuye a medida de que p (dimensión del vector de observaciones) aumenta. Además, se observa que la FDL puede ser adecuada para la clasificación pero no para estimar la probabilidad a posteriori de cierta población dada una observación específica.

Gessaman y Gessaman (1972) presentan un nuevo método de discriminación basado en el uso de bloques estadísticamente equivalentes (ver Wilks (1962)), tal método cae dentro de los de distribución libre. Los autores llevan a cabo comparaciones con respecto a otros métodos de discriminación sin decisiones de "reserva" * y con métodos que si permiten decisiones de reserva, tales comparaciones se basan en los resultados de experimentos de simulación Monte Carlo realizados para 3 problemas de dos poblaciones alternativas normales bivaria

* i.e., decisiones no conclusivas.

das, además se toman tres tamaños diferentes de muestras de ensayo: 729, 200 y 64, los procedimientos se calibran con muestras de 250 - observaciones de cada una de las poblaciones consideradas. Es la opinión de este autor que las conclusiones extraídas de tales resultados - son un tanto engañosas por varias razones:

1°. El tipo específico de distribuciones, ya que exclusivamente se trata con normales bivariadas que son muy "adecuadas" para utilizar el método propuesto por Gessaman y Gessaman, pero merecería pensar acerca de la relación entre número de variables y potencia discriminante de los métodos (que más adelante se trata).

2°. Los tamaños de las muestras ensayo, ya que contar con muestras de 729 y 200 observaciones para cada población se antoja un tanto ideal, en muchos casos aún 64 individuos parece difícil de conseguir sobre todo en problemas taxonómicos y arqueológicos.

3°. No se compara contra otros métodos que podrían haberse comportado bastante bien sobre todo en los casos de diferentes medias y varianzas y en los de diferencia nula en la medias (ver Anderson y Bahadur (1962), Geisser (1973)).

Press (1972) discute el uso de las matrices de confusión pre

sentadas en el trabajo de Massy (1965), para probar la potencia discriminatoria de un procedimiento. Esta potencia discriminatoria se prueba comparando mediante una prueba ji-cuadrada, los resultados obtenidos al aplicar un cierto método con los que hipotéticamente se obtendrían al efectuar la discriminación aleatoriamente.

Al aplicar esta prueba, se debe tener cuidado al interpretar el resultado puesto que solamente es válido si se aplica a nuevas observaciones, pero si se usan las observaciones de las muestras de ensayo utilizadas para diseñar la función discriminante se incurre en un sesgo optimista, puesto que tal función es la mejor para discriminar entre las observaciones de ensayo. Para una discusión más amplia de este asunto ver el trabajo de Frank, Massy y Morrison (1965).

Lachenbruch et al. (1973a) analizan la robustez de la FDL y la función discriminante cuadrática (FDC) ante ciertos tipos de no-normalidad, los mismos autores comentan que nada se ha hecho al respecto hasta donde ellos conocen. Las distribuciones de origen consideradas presentan un alto grado de no-normalidad y son las sugeridas por Johnson (1949). El criterio de funcionamiento elegido son las probabilidades de clasificación equivocada.

En base a experimentos de simulación para diferentes números

de variables, distancia entre las poblaciones y tamaños de las muestras de ensayo, los resultados principales son los siguientes: la FDL se ve muy afectada por la no normalidad de las poblaciones, además este efecto depende del tipo específico de no normalidad que se trate, para las distribuciones estudiadas los menores efectos se observan para las distribuciones del tipo logístico (las cuales están acotadas por arriba y por abajo), la FDC tiene un funcionamiento pobre y en algunos casos peor que la FDL, dependiendo del tipo de no normalidad, ésto no es muy raro ya que a pesar de estar diseñadas ambas funciones para tratar con variables normales, la FDL tiene la ventaja de que un modo alternativo de llegar a ella es el determinar la combinación lineal de las observaciones que maximice la variación entre grupos con respecto a la variación dentro de grupos, por lo que su aplicación aún en situaciones con poblaciones diferentes de la normal tiene una justificación intuitiva. De ser posible, deben de transformarse los datos para aproximarse a la normalidad y así usar la FDL.

Lachenbruch (1973b), efectúa una comparación entre dos métodos para discriminar entre k poblaciones, uno de ellos basado en la idea de Fisher, el cual se denominará aquí el de vectores propios o variables canónicas (VC), el otro es conocido como función discriminante múltiple (FDM). Se trata el caso de poblaciones normales homocedás

ticas.

El objetivo del trabajo de Lachenbruch es responder a las siguientes preguntas:

i) ¿Bajo qué condiciones el método VC se comporta mejor que el FDM?

ii) ¿Cuál es el efecto del tamaño de las muestras de ensayo en el comportamiento de los métodos analizados?

iii) ¿Cómo afecta el número de poblaciones el comportamiento de los métodos?

iv) ¿Cómo los afecta la dimensión de las observaciones?

La comparación se efectúa en base a experimentos de simulación, los resultados principales son los siguientes:

a) El método VC se comporta mucho mejor para la estructura colineal (C)* de medias poblacionales que para la estructura simplex (S) de tales medias.

b) El incremento del tamaño de muestra conduce a un decremento en la probabilidad aparente de clasificación correcta (ver sección

* En el trabajo solamente se analizan dos estructuras de las medias poblacionales: colineal con distancias iguales y simplex k dimensional, para diferentes distancias.

III.1), la explicación más plausible es debido a que se usa el método de resubstitución para estimar tal probabilidad, el sesgo "optimista" de esta se reduce al aumentar el número de variables.

c) Un incremento en el número de poblaciones conduce a decrementar la probabilidad media de clasificación correcta pues aumentan las oportunidades de efectuar asignaciones erróneas.

d) La probabilidad de clasificación correcta se incrementa al aumentar el número de variables por observación. Esto es sumamente interesante ya que no se proporciona más información en las variables adicionales, lo que parece suceder (según el autor), es que el conjunto de datos puede ser ajustado más exactamente cuando se usan más variables.

De lo anterior se desprende que si las medias de las variables son colineales (o casi) el método VC se comporta aproximadamente como el FMD. Por lo tanto, si el número de variables es grande (dimensión de las observaciones) el método VC es una alternativa computacionalmente adecuada. En cuanto a los efectos del tamaño de muestra el mismo Lachenbruch menciona que son un tanto inciertos aunque tal parece que ambos métodos se ven afectados de manera semejante.

El número de grupos está inversamente relacionado al funcionamiento de ambos métodos afectándolos de igual manera aproximadamente.

Debido a que se han analizado poco los métodos para más de 2 poblaciones, sería deseable que se llevasen a cabo más trabajos semejantes al de Lachenbruch sobre todo para analizar robustez con respecto a no normalidad, violación de la hipótesis de homocedasticidad y efectos de diferencias en las proporciones de las muestras de ensayo.

Cabe hacer el comentario de que los métodos FMD y VC son equivalentes, en el caso de que el método VC no se use desechando aquellas variables canónicas cuyas correlaciones canónicas sean no significativas, esto se debe a que tanto la función discriminante de Fisher como la D^2 de Mahalanobis son invariantes bajo trasnformaciones no singulares de $\underline{x}$, vector de observaciones (Kshirsagar y Arseven (1975)).

Fisher y Van Ness (1973) proponen comparar varios algoritmos de discriminación mediante la proposición de propiedades deseables requeridas para un caso específico y así en base a ellas poder encontrar el procedimiento admisible con respecto a tales propiedades.

Comparan varios procedimientos, cinco de ellos utilizados para efectuar agrupamiento (clustering) que son vecino más próximo, vecino más alejado, centroides, eslabonamiento promedio y mínimos cuadrados con otros usados exclusivamente en discriminación FDL (dos modalidades) FDC (dos modalidades) y una regla basada en probabilidades a posteriori vía el teorema de Bayes.

Es interesante seguir su discusión aunque de los criterios propuestos para verificar la admisibilidad, algunos tienen poca importancia en el problema de discriminación y parecen más interesantes para el problema de agrupamiento (cabe hacer notar que los autores presentaron anteriormente un trabajo semejante sobre técnicas de agrupamiento).

En base a los criterios establecidos por ellos, los autores encuentran que los procedimientos de vecino más próximo, vecino más lejano y el que se basa en el teorema de Bayes son los "mejores". Se debe tomar este estudio simplemente como una proposición y determinar el "mejor" procedimiento para el caso específico en base a las propiedades deseables adecuadas en la situación que se trate.

Marks y Dunn (1974) comparan mediante experimentos de si-

mulación Monte Carlo, el comportamiento de tres funciones discriminantes para el caso de dos poblaciones con matrices de covarianza diferentes. Las funciones comparadas son la cuadrática (FDC), la mejor lineal de Anderson y Bahadur* y la FDL. La comparación se lleva a cabo usando muestras y en forma asintótica en base a la probabilidad general de clasificación equivocada.

El problema se estudia en su forma canónica (sugerida por - Dunn y Holloway (1966) con respecto al análisis de la T_0^2 de Hotteling) i.e., $\underline{\mu}_1 = \underline{0}$, $\underline{\mu}_2 = \underline{\nu}$, $\Sigma_1 = I$, $\Sigma_2 = \Lambda$ una matriz diagonal con valores en ésta $\lambda_1, \ldots, \lambda_p$. Se tratan diversas combinaciones para los valores de $(\lambda_1, \ldots, \lambda_p) = \underline{\lambda}$, componentes de $\underline{\nu}$, N_i $i = 1, 2$ (los tamaños de las muestras ensayo). Se presentan cuatro combinaciones de $\underline{\lambda}$, $\underline{\nu}$, diferentes distancias entre las poblaciones y diferente número de componentes del vector de observaciones. Los resultados principales pueden resumirse en que para muestras grandes de distribuciones normales la FDC es mucho mejor que la FDL para valores grandes de λ (valor utilizado para formar diferentes matrices Λ), para los valores pequeños es ligeramente mejor. Para muestras pequeñas la FDC funciona peor que la FDL para valores pequeños de λ y tal comportamiento se hace más crítico si el número de parámetros aumenta. Para valores pequeños de λ la mejor función lineal se com

* presentada inicialmente por Clunies - Ross y Riffenburgh (1960a , 1960b)

porta ligeramente mejor que la FDL y para λ más grande se comporta mucho mejor, sin embargo, para tales valores la FDC es generalmente la mejor. Welch y Wimpress (1961) comparan la FDC y la mejor lineal en un problema más específico con un conjunto de datos.

Goldstein (1975) presenta una comparación entre algunos procedimientos de clasificación basados en estimadores de densidad, el objetivo de tal estudio es analizar la "capacidad" para clasificar correctamente observaciones no clasificadas como una mejor alternativa para el investigador, al que sólo le preocupa clasificar sus observaciones en vez del comportamiento asintótico de un procedimiento particular.

El estudio se basa en los resultados de experimentos de simulación Monte Carlo, las poblaciones consideradas son normales bivariadas y diferentes tamaños de muestra para estimar las densidades, los métodos utilizados son el de Loftsgaarden y Quesenberry (LQ) y el de Rosenblatt, Parzen y Cacoullos (RPC), el método paramétrico frente al cual se comparan es el de inserción de estimadores para la regla de discriminación entre varias poblaciones. Los resultados obtenidos son que los métodos no paramétricos se comportan muy bien frente al procedimiento paramétrico, aunque el mérito de los métodos LQ y RPC, dependen en gran manera de la elección de las constantes involu

cradas en su definición. Se concluye que el método LQ parece ser más adecuado o natural para el problema de discriminación.

Se debe comentar que se hacen necesarios estudios de este tipo para analizar la robustez de los métodos, así también probarlos - con observaciones de una dimensión mayor. Un problema que se avisora es que los métodos no paramétricos requieren en ocasiones un - mayor conocimiento de las técnicas utilizadas (a diferencia de la situación en otros casos), por lo que se puede pensar que esto es una desventaja de tales procedimientos.

III.4 Tópicos especiales.

En esta sección se presentan los resultados de algunos trabajos relacionados con situaciones especiales del problema de discriminación.

III.4.1 Los efectos de clasificación equivocada de las muestras de ensayo.

De nuevo es Lachenbruch (1966), el que se preocupa por el problema de discriminación, en este caso estudia la situación en la - cual algunas observaciones en las muestras de ensayo están mal clasificadas, los resultados que obtiene son que para el caso de muestras grandes es que la distribución de la función de discriminación es una normal univariada con las medias respectivas a cada población más - cercanas de lo que teóricamente se supone, los coeficientes de discriminación basados en las muestras mal clasificadas difieren de los verdaderos por una constante de proporcionalidad que depende de las proporciones de observaciones mal clasificadas para cada población, los tamaños respectivos de muestras de ensayo y la distancia de Mahalanobis entre las poblaciones. Además la función discriminante difiere de la verdadera en un término constante lo que significa que se usará una constante equivocada al efectuar la discriminación.

Vale la pena hacer notar que las dos probabilidades de clasificación equivocada, para el caso de costos de clasificación errónea y probabilidades a priori iguales, no se ven alteradas si la proporción - de individuos mal clasificados es igual para ambas muestras de ensayo.

Para el caso de muestras pequeñas, debido a la dificultad de analizar la distribución de la función discriminante, el estudio se basa en analizar experimentos de simulación en los cuales solamente se clasifica equivocadamente una proporción de una de las poblaciones, los resultados indican que lo observado en el caso de muestras de ensayo grandes es una buena aproximación al comportamiento en muestras pequeñas.

McLachlan (1972), presenta un enfoque alternativo al de Lachenbruch (1966) ya que el efecto de clasificación equivocada lo expresa en la forma de expansión asintótica. Los resultados para pequeñas muestras difieren de los del citado trabajo de Lachenbruch pero para muestras grandes concuerdan en general. El autor considera que se cuenta con una muestra de ensayo de cada población en las cuales hay algunos individuos mal clasificados, los cuales se supone que son una muestra aleatoria de su población de origen (tal vez esta es una suposición que no se cumple en la práctica y que puede llevar a resultados un tanto engañosos). Bajo la consideración de que $\alpha_1 = 0$ (α_i $i = 1, 2$, denota la proporción de observaciones de P_i clasificados como de P_{3-i}) encuentra las esperanzas de $P_{2|1}$ y $P_{1|2}$, sus respectivas varianzas y analiza el efecto que tiene, tanto en las esperanzas como en las varianzas, el valor de α_2. Los resultados pueden resumirse en que Var $(P_{2|1})$ es una función creciente de α_2 en tanto que -

Var $(P_{1|2})$ es decreciente, suponiendo que los tamaños de las muestras de ensayo son iguales, i.e., $n_1 = n_2 = n$, $E(P_{2|1})$ (correcta hasta el orden n^{-1}) es una función creciente de d_2 lo cual concuerda con los resultados de Lachenbruch, para $E(P_{1|2})$ no se tiene un resultado general, sin embargo, para cierta relación (McLachlan (1974) pág. 410), entre la dimensión de los vectores de observaciones y el tamaño común de las muestras de ensayo*, se tiene que $E(P_{1|2})$ es una función decreciente de d_2.

Finalmente, Lachenbruch (1974), estudia el problema considerando dos modelos de clasificaciones equivocadas iniciales no aleatorias. Basándose en los resultados de experimentos Monte Carlo evalúa el comportamiento de la función discriminante lineal en este caso. Los modelos de clasificación equivocada son el llamado de separación completa definido como sigue: asignar la observación de ensayo $\underline{x}$ a la población para la cual $(\underline{x} - \underline{\mu}_i)' (\underline{x} - \underline{\mu}_i)$ $i = 1, 2$ tenga el valor menor; el segundo modelo es una generalización del anterior pero en este caso la observación $\underline{x}$ proveniente de P_i, se clasifica erróneamente si $(\underline{x} - \underline{\mu}_i)' (\underline{x} - \underline{\mu}_i)$ es mayor que un valor constante V_i, en su estudio Lachenbruch usa para tales valores percentiles escogidos de una χ^2_p donde p es la dimensión de las observaciones $\underline{x}$. Los resultados que obtiene son para diferentes valores de p, n (tamaño común de

* esta relación varía también en relación a la distancia entre las poblaciones

las muestras de ensayo), V_1 y V_2. Además considera tres valores de la distancia entre las medias de las poblaciones. Se observa que las probabilidades verdaderas de clasificación equivocada solamente son afectadas ligeramente por los tipos de clasificación equivocada considerados, es más, el efecto general parecen ser que se separan mejor las poblaciones, la razón de esto puede ser que otros métodos eliminan los efectos que cualquier observación "extraña" (outlier) tenga sobre la función discriminante.

En el estudio también se analiza el comportamiento de dos estimadores de errores de clasificación, que se presentaron en la sección III.1 : el método de resubstitución y el D (de Fisher). Puede concluirse que tanto el método R (resubstitución) como el D, en caso que se sospeche la existencia de cualquier clasificación equivocada inicial, no deben usarse ya que los valores proporcionados por los métodos citados son afectados considerablemente.

III.4.2 El tratamiento de valores faltantes.

Chan y Dunn (1972), presentan un trabajo en el cual utilizan la probabilidad de clasificación correcta bajo los métodos de manejo de valores faltantes más comúnmente utilizados para realizar una comparación entre éstos. Bajo la suposición de que se tienen dos muestras de ensayo independientes de dos poblaciones normales multivariadas homocedásticas, la regla para efectuar la discriminación es la FDL, (por cierto los autores llaman incorrectamente a la regla de discriminación que usan función discriminante lineal de Fisher, en realidad se refieren a la regla W de Anderson). La probabilidad incondicional de clasificación correcta, ver Dunn y Varady (1966), se usa como criterio para comparar los métodos en base a métodos Monte Carlo.

Los métodos comparados son los siguientes:

Método A. En este método no se usan las observaciones a las cuales les falta algún valor.

Método B. En vez de desechar las observaciones con valores faltantes, se utiliza toda la información posible.

Método BA (modificación del B). Debido a que en algunas ocasiones la matriz de covarianza muestral no es definida positiva al usar el método B, se verifica esto mediante el valor estimado de la distancia de Mahalanobis y si éste es negativo se utiliza el método A.

Método C. Método de substitución de la media. Se calculan primero las medias muestrales para las variables en base a todos los valores que se tengan, posteriormente se substituyen por los valores faltantes respectivos.

Método D. Método de Regresión. En este caso se obtienen ecuaciones de regresión entre las variables en base a las observaciones completas y entonces se substituye por cada valor faltante el estimado de la ecuación apropiada.

Método DS (modificación del método D). Semejante al D, pero en este caso se va haciendo uso de manera secuencial de la información en las observaciones con un valor faltante, posteriormente con dos y así en adelante.

Método E. Método de Componentes Principales. La primer componente principal de cada una de las dos matrices de observaciones

muestrales se usa para estimar los valores faltantes.

Método O. En este caso no hay valores faltantes en las observaciones (se usa como control).

La comparación se lleva a cabo bajo el supuesto de que las poblaciones son normales multivariadas homocedásticas con parámetros $\underset{\sim}{\mu}_1 = \underset{\sim}{0}$, $\underset{\sim}{\mu}_2 = \underset{\sim}{\mu}$ y $\Sigma = \mathcal{Q}$, $\mathcal{Q}$ matriz de correlación con estructura de igual correlación i.e., $\rho_{ii} = 1$ y $\rho_{ij} = \rho$ $i \neq j$. Además se utilizan 250 matrices de correlación escogidas aleatoriamente. Los parámetros básicos son p_1, Δ, m (proporción de valores faltantes, escogidos de manera aleatoria) n_1 y n_2. El criterio de comparación es la probabilidad general de clasificación correcta.

Los resultados pueden resumirse como sigue: para $p = 2$ el método D parece ser el mejor para valores pequeños de $|\mathcal{Q}|$, el método E para valores moderados y el C para valores grandes. Aunque el método A no es tan bueno como los métodos D, E y C, las diferencias son muy pequeñas; sin embargo, el método A va comportándose cada vez peor en tanto p aumenta (excepto para valores extremadamente pequeños de $|\mathcal{Q}|$). El método E es preferible para valores pequeños de $|\mathcal{Q}|$ (excepto para valores extremadamente pequeños de $|\mathcal{Q}|$) y el

C para valores grandes.

Todos los métodos de manejo de valores faltantes mejoran (en relación al método O) cuando Δ se incrementa. Continúan las relaciones entre ellos pero las diferencias disminuyen. El rango de valores de $|\rho|$ en el cual el método C es preferible al E se incrementa al cambiar Δ de 2 a 4.

Todos los métodos mejoran al aumentar los tamaños de las muestras de ensayo, especialmente el método A. Una cosa similar sucede al disminuir el valor de m.

En términos de la esperanza de la probabilidad de clasificación correcta parece que los tamaños de muestra diferentes tienen poco efecto sobre las diferencias entre los métodos de manejo de valores faltantes.

Para valores negativos de ρ con $p > 2$ las matrices de equicorrelación produjeron resultados diferentes que las matrices con ρ positiva. A excepción del A, todos los métodos funcionaron más satisfactoriamente para valores positivos de ρ.

Se utilizaron diferentes formas de μ_j desde $(d, 0, \ldots, 0)'$ has

ta (d, d, ..., d)'. Los métodos C y D mejoraron gradualmente en relación a estos cambios en $\frac{\mu}{\sigma}$.

Finalmente los autores analizan la variabilidad del funcionamiento de los métodos en base al error estandar del valor estimado de la esperanza de clasificación correcta y en los estimadores del $5^{o}(P_{05})$, $50^{m_o}(P_{50})$ y $95^{o}(P_{95})$ percentiles de la distribución de la PCC. En general los métodos con valores bajos de E (PCC) muestran valores de error estándar grandes y valores bajos de P_{05}. Por otro lado los valores de P_{95} fueron muy similares para todos los métodos.

La variabilidad en los métodos se incrementó al aumentar el número de variables. Los valores de muestra diferentes causaron una mayor variabilidad de los estimadores de E (PCC).

Posteriormente los mismos Chan y Dunn (1974), comparan los métodos C, D y E en vase al comportamiento asintótico de la probabilidad de clasificación correcta bajo cada método. Las comparaciones las efectúan mediante experimentos de simulación Monte Carlo con el modelo de equicorrelación* antes referido. Los resultados son los siguientes: aunque hay ligeras variaciones en el comportamiento asintótico entre los diferentes métodos, las diferencias de E (PCC) no son -

* aunque un poco reducido en cuanto a variedad de situaciones

apreciables o substanciales. El que parece ser el método más débil es el E (método de componentes principales). Además, se nota una tendencia de E (PCC) a incrementarse en relación directa con k cuando ρ es positiva y decrementarse con respecto a k para ρ negativa.

Debido a que los métodos A y B siempre alcanzan la máxima-probabilidad asintóticamente y dado que para los métodos C, D y E los valores asintóticos no están muy lejanos del máximo, el sesgo asintótico parece tener poca importancia práctica. De lo anterior se desprende que el estudio del tratamiento de valores faltantes en análisis discriminante debe de circunscribirse a muestras de tamaño pequeño y mediano.

III.4.3 Un procedimiento iterativo de reclasificación.

McLachlan (1975), considera el problema de dos poblaciones normales con diferente vector de media pero matriz de covarianza común, iguales probabilidades a priori e iguales costos de clasificación equivocada. Supone que se tienen muestras de ensayo de tamaños n_1 y n_2 y además, que hay M individuos sin clasificar.

El procedimiento iterativo consiste en calcular la estadística W de Anderson en base a las $n_1 + n_2$ observaciones de ensayo, usar la función encontrada para discriminar las M observaciones nuevas obteniendo m_{11} y m_{21} (donde $m_{11} + m_{21} = M$) que son observaciones asignadas a P_1 y a P_2 respectivamente y entonces construir

$$W_1 = \left\{ \underline{x} - 1/2 \, (\underline{x}_{11} + \underline{x}_{21})' \quad S_1^{-1} \, (\underline{x}_{11} - \underline{x}_{21}) \right\}$$

donde $\underline{x}_{i1}$ es la media de $n_i + m_{i1}$ observaciones ($i = 1, 2$) y S_1 es la matriz de covarianza muestral usual basada en $n_1 + m_{11}$, $n_2 + m_{21}$ observaciones considerando como si las $n_i + m_{i1}$ observaciones pertenecieran a P_i ($i = 1, 2$).

La estadística W_1 se aplica nuevamente a las M observaciones no asignadas obteniendo m_{12} y m_{22} de manera similar a la ante

rior se calcula W_2 y de manera iterativa se obtienen nuevas estadísticas W_k.

Al analizar el comportamiento asintótico de tal procedimiento, se obtiene que el riesgo (o pérdida esperada) promedio, e_k, tiende al riesgo del procedimiento óptimo, r (cuando se conocen todos los parámetros), cuando el número de iteraciones tiende a infinito y además se considera M muy grande en relación a $n_1 + n_2$. Además para una Δ distancia de Mahalanobis grande, la diferencia $e_k - r$ es muy pequeña después de sólo unas cuantas iteraciones.

El procedimiento sugerido se evalúa para M finita, mediante un experimento Monte Carlo. Los resultados obtenidos sugieren que una regla con un riesgo promedio menor que para la original se obtiene generalmente después de una iteración.

La idea de usar las $n_1 + n_2 + M$ observaciones con propósitos de mejorar los estimadores y diseñar una regla de discriminación mejor, no es nueva. Hartley y Rao (1968), discuten la obtención de estimadores máximo verosímiles de los parámetros considerando primero estimadores máximo verosímiles condicionados en todas las posibles subdivisiones de los M observaciones no asignadas entre las dos pobla

ciones. Sin embargo, el procedimiento sugerido por McLachlan (1975), parece más manejable computacionalmente.

III.4.4 Enfoque secuencial al problema de discriminación.

Samuel (1963a), considera el siguiente problema: se debe clasificar un grupo de n individuos que se sabe pertenecen a la población P_0 ó a la P_1. Los individuos del grupo llegan secuencialmente para inspección y clasificación, la clasificación del i-ésimo individuo se efectúa inmediatamente después de que ha sido inspeccionado.

Existe una estructura de costos de clasificación equivocada para el problema que en general se consideran diferentes para cada acción errónea. El interés se centra entonces en la pérdida, o costo promedio, esperada debida a la clasificación equivocada.

Samuel propone un procedimiento secuencial en el cual se utilizan las observaciones ya clasificadas para la clasificación de la observación actual, este procedimiento forma una clase completa mínima, (ver primer capítulo de este trabajo), de reglas de decisión. Además las reglas propuestas son "fuertemente secuenciales" en el sentido de que si se establece una sucesión de decisión en base a tales reglas para toda n, $n = 1, 2, \ldots$ la regla sugerida forma la parte inicial de la sucesión, i.e., no se necesita conocer de antemano el número de individuos pertenecientes a la población i, $i = 0, 1$ y que se deben clasificar.

Entre las interpretaciones posibles para la regla propuesta se tiene que es una regla que usa en cada etapa todas las observaciones previas para llevar a cabo la asignación de un nuevo individuo. Además la distribución a priori, para la cual la regla es la solución de Bayes, se va actualizando a cada etapa, tomando en cuenta todos los individuos ya clasificados (una especie de uso de muestras post-supervisadas).

La misma autora propone en Samuel (1963b), soluciones para el caso en que el grupo por clasificar (discriminar) no provenga totalmente de la misma población.

Posteriormente Srivastava (1973), considera el problema en el cual las poblaciones son normales multivariadas. Este autor considera tres poblaciones P_1 , P_2 y P_o , la cual coincide con P_i para i = 1 ó 2, en realidad el problema que resuelve corresponde al de identificar a la población P_o. Bajo la suposición de que P_1 es $N(\underset{\sim}{\mu}_1 , \Sigma)$ y P_2 es $(\underset{\sim}{\mu}_2 , \Sigma)$, trabaja inicialmente el caso en el cual conoce $\underset{\sim}{\delta} = \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ sin pérdida de generalidad sólo se necesita muestrear de una de las 2 poblaciones, el resultado que obtiene es que se debe tomar un número n de observaciones tanto de P_1 ó P_2 y de P_o , la variable de alto (parada) se define por

N = menor n ($\geq n_0$) tal que $\quad n \geq \dfrac{g a^2}{\underline{\delta}' S_m^{-1} \underline{\delta}}$

donde a es tal que la distribución de una normal estándar en a es $1 - d$, d el valor control para la probabilidad de clasificación equivocada del procedimiento

$$S_m = \frac{1}{m} \sum_{i=0}^{1} \sum_{j=1}^{n} (\underline{X}_{ij} - \underline{\bar{X}}_{in})(\underline{X}_{ij} - \underline{\bar{X}}_{in})' \qquad m = 2(n-1)$$

y

$$n \underline{\bar{X}}_{in} = \sum_{j=1}^{n} \underline{X}_{ij} \qquad i = 0, 1$$

(en este caso x muestrea sólo de P_1 y P_0)

no es tal que $\quad 2n_0 \geq p + 2$

Cuando se termina de muestrear se clasifica P_0 como P_1 ó P_2 de acuerdo a si

$$(\underline{\bar{X}}_{0n} - \underline{\bar{X}}_{1n} + 1/2\, \underline{\delta})' S_m^{-1} \underline{\delta} \gtrless 0$$

respecto a la media

Srivastava (1973), demuestra que tomando un número finito k, de observaciones adicionales a la n fijada por la regla de parada se puede lograr un control d sobre la probabilidad de clasificación equivocada, k puede depender de μ_1 pero no de μ_1 , μ_2 y Σ .

En el caso de δ desconocida se muestrea de todas las pobla

ciones y entonces la regla de muestreo queda como

$$N = \text{menor entero } n \ (\geq n_0) \text{ tal que } \quad n \geq \frac{6a^2}{\underline{\delta}_n' \omega_L^{-1} \underline{\delta}_n}$$

donde

$$\omega_L = \frac{1}{L} \sum_{i=0}^{2} \sum_{j=1}^{n} (\underline{X}_{ij} - \bar{\underline{X}}_{in})(\underline{X}_{ij} - \bar{\underline{X}}_{in})' \qquad L = 3(n-1)$$

$$\bar{\underline{X}}_{in} = \frac{1}{n} \sum_{j=1}^{n} \underline{X}_{ij} \qquad i = 0, 1, 2$$

$$\underline{\delta}_n = \bar{\underline{X}}_{1n} - \bar{\underline{X}}_{2n}$$

La regla de discriminación queda como: clasificar como P_1 ó P_2 de acuerdo a si

$$\left[\bar{\underline{X}}_{0n} - \tfrac{1}{2} (\bar{\underline{X}}_{1n} + \bar{\underline{X}}_{2n}) \right]' \omega_L^{-1} \underline{\delta}_n \gtrless 0$$

Como un comentario se tiene el que no es necesario tener muestras de igual tamaño de las tres poblaciones, se puede muestrear de acuerdo a alguna proporción especificada.

Se demuestra en el artículo que el costo de desconocer $\sum$ la matriz de covarianza es el de necesitar las k observaciones adicionales a las que marca la regla de parada.

BIBLIOGRAFIA

CAPITULO III.-

III.1

Hingleyman (1962)

Kanal y Chandrasekaran (1971)

Armitage (1966)

Chow (1962)

Duda y Fossum (1962)

Duda y Hart (1973)

Smith (1947)

Lunts y Brailovskdy (1967)

Lissack y Fu (1972)

Toussaint (1969) , (1970 junto con Donaldson), (1972)

(1974)

Sorum (1968) , (1972)

McLachlan (1974)

Kshirsagar (1972)

John (1961) , (1963) , (1964)

Rao (1954)

Anderson (1951 , 1958)

Dunn y Varady	(1966)
Hills	(1966)
Lachenbruch	(1965) , (1967) , (1968 junto con Mickey)
Cochran	(1968)
Foley	(1972)

III.2

Cochran	(1964)
Allais	(1964) , (1966)
Chu y Chueh	(1967)
Estes	(1965)
Cornfield	(1967)
Gaffey	(1951)
Weiner y Dunn	(1966)
Elashoff et al.	(1967)
Francis	(1967)
Hills	(1967)
Belson	(1959)
Mc Naughton - Smith	(1963)
Cox	(1966)
Horton et al.	(1968)
Boullion et al.	(1975)
Kullback	(1952, 1968)
Hughes	(1968) , (1969)

Duda y Hart (1973)

III.3

Gilbert (1968) , (1969)

Cochran y Hopkins (1961)

Gessaman y Gessaman (1972)

Press (1972)

Massy (1965)

Frank, Massy y Morrison (1965)

Lachenbruch et al. (1973)

Lachenbruch (1973a) , (1973b)

Kshirsagar y Arseven (1975)

Fisher y Van Ness (1973)

Marks y Dunn (1974)

Goldstein (1975)

III.4

Lachenbruch (1966) , (1974)

McLachlan (1972) , (1974) , (1975)

Chan y Dunn (1972) , (1974)

Hartley y Rao (1968)

Samuel (1963a) , (1963b)

Srivastava (1973)

IV.- Algunos ejemplos de aplicaciones.

En este capítulo se presentan algunos ejemplos de aplicaciones de los procedimientos de discriminación. Se describen brevemente las características de las variables involucradas y se comentan los resultados del uso de la técnica. Los ejemplos se presentan en orden cronológico.

Kossack (1945), presenta una aplicación en la cual el objetivo es clasificar un individudo entrenado en el grupo con trabajo satisfactorio o en el de trabajo insatisfactorio en el primer período del curso de matemáticas (los individuos estaban cubriendo los prerequisitos de ingeniería en la Universidad de Oregon), las variables usadas fueron los resultados en tres diferentes pruebas, dos de matemáticas y una general. Bajo la suposición de distribución normal de los vectores de observaciones lleva a cabo la discriminación usando la estadística de

Wald (llama la atención el hecho de que presenta desglosados todos los valores numéricos utilizados), con la cual obtiene una probabilidad de clasificación equivocada igual al 20% aproximadamente, la cual compara con la que realmente se cometió resultando que el valor estimado aproximó bastante el valor real. Es interesante seguir los cálculos de Kossack aunque en la actualidad ya resulte de poco valor práctico su estudio.

Tintner (1946), discute en su trabajo una aplicación efectuada por Durand (1941) a datos financieros para discriminar entre préstamos buenos y malos. Las variables utilizadas fueron el enganche, el precio del préstamo, el ingreso mensual (todos en dólares) y la longitud (en meses) del período del contrato. Se obtuvo la función discriminante lineal de Fisher, la cual parece apropiada aunque es obvio que la longitud del período del contrato no se distribuye normalmente, sin embargo, por las características de la FDL ésta funcionó bien en tal situación.

Tintner en el mismo artículo trata el caso de aplicar el procedimiento de Fisher para discriminar entre los precios de bienes de producción y los bienes de consumo, las variables consideradas fueron: la mediana de la longitud del ciclo en meses (el ciclo medido de mínimo a mínimo), la mediana del porcentaje de la duración de precios as

cendentes relativa a la duración total del ciclo, la mediana de la amplitud cíclica y la tasa media de cambio en el ciclo. Tintner discute los resultados obtenidos y comenta que las variables utilizadas probablemente no se distribuyan normalmente con lo que una combinación no lineal de las variables fuese más adecuada en ese caso.

Rao y Slater (1949), aplican los métodos para distinguir individuos neuróticos de normales, los resultados se basan en tres clases de mediciones sicológicas. Los individuos se clasifican en 5 grupos de neuróticos y un grupo control de individuos normales. El estudio no sólo abarca lo concerniente a la discriminación sino que también analiza los datos mediante correlación canónica de una manera bastante amplia, para llevar a cabo la discriminación un método sugerido por Rao (1948) con el objeto de minimizar el número de individuos mal clasificados suponiendo conocidas las poblaciones (vía probabilidades a priori), este método prácticamente coincide con el expuesto por Anderson (1951). Los resultados son inadecuados para un uso práctico puesto que el propósito del artículo es presentar algunas técnicas del análisis multivariado, las cuales tienen una aplicación potencial en sicología.

Banks (1950) utilizó la FDL para discriminar entre compradores y no compradores (potenciales) de un producto. Se tratan específi

camente dos casos: un blanqueador y una marca de café. Las variables consideradas para el primero fueron apariencia del empaque, habilidad para limpiar, abrasividad en el uso, irritación para las manos, olor y precio, en tanto que para el café, las variables fueron a su vez apariencia del empaque, sabor, capacidad para vender más tazas por libra y precio. Obviamente las variables que se trataron fueron cualitativas y se les asignaron ciertos valores (rangos). Se utilizan métodos para variables cuantitativas lo cual es evidente que causa alguna distorsión en los resultados. Los coeficientes de la función discriminante se interpretan en forma intuitiva, lo que no es muy válido debido a que la escala elegida para asignar valores a los resultados fue arbitraria. Con los procedimientos con los que se cuenta actualmente para tratar variables cualitativas se estima que se hubiesen obtenido resultados más confiables.

Anderson (1951), analizó el problema considerado por Rao (1948), en el que se deseaba discriminar entre tres castas comunes en la India (Brahmin, Artisan y Korwa) en base a las mediciones de estatura, altura estando sentado, profundidad nasal y altura de la nariz. Compara sus resultados (mediante la solución minimax (ver capítulo I), con los de Rao (1948) corregidos por el propio Anderson y encuentra que el procedimiento minimax produce una probabilidad máxima de cla

sificación equivocada menor que el método de Rao.

Evans (1959), aplicó la técnica para discriminar entre dos poblaciones una de poseedores de autos Ford y la otra de Chevrolet, el objetivo del estudio era probar la facultad de los métodos psicológicos y objetivos para discriminar entre los poseedores de un automóvil de una clase de las dos citadas marcas. Las variables consideradas fueron mediciones de actitudes psicológicas: logro, deferencia, exhibiciones, autonomía, afiliación, intracepción, dominancia, humillación, cambio, agresión y heterosexualidad. El método de discriminación usado fue la FDL de Fisher, observándose que el porcentaje de individuos mal clasificados fue del 37% global (siendo mayor tal porcentaje para los poseedores de Ford), los mismos datos fueron presentados a un grupo de psicólogos, los cuales obtuvieron una clasificación más pobre que la efectuada mediante la FDL. Se discuten los resultados obtenidos llegando a la conclusión de que las necesidades de personalidad, medidas en el estudio son de poco valor para discriminar en este caso. Se eligieron otras variables para comparar siendo estas: "edad" del automóvil que se poseía en ese momento, uso del automóvil por más o menos de 10 000 millas al año, comprador que trató con más de un vendedor antes de la compra, fumador o no fumador, poseedores de casa contra arrendatarios, 3 ó más hijos que viven en casa o no,

preferencia religiosa, asistencia a la iglesia más de una vez o menos de ello al mes, partido político de preferencia, edad del dueño, si el dueño había permanecido en su trabajo más de 5 años o menos de ese lapso, ingreso anual familiar. Los resultados obtenidos fueron que la FDL basada en las anteriores variables (demográficas) no tenía suficiente poder predictor para ser de uso práctico.

Se intentó un tercer método mezclando las variables con mayor significancia en las funciones discriminantes anteriores (medido esto mediante los valores absolutos de los coeficientes asociados a las variables, a mayor valor absoluto del coeficiente más significativa la variable), este intento tampoco tuvo el éxito deseado, aunque se mejora en cuanto a poder discriminante de una manera ligera. Se compararon estadísticamente los resultados de las tres funciones discriminantes intentadas.

Cabe comentar que una posible causa de que las técnicas utilizadas no fuesen de gran ayuda práctica, puede ser el hecho de utilizar una función discriminante diseñada para variables esencialmente contínuas. Tal vez con métodos específicamente diseñados para tratar con variables de tipo discreto los resultados hubiesen sido mejores.

Cochran y Bliss (1961), presentan un ejemplo ampliamente desarrollado con la intención de familiarizar a los investigadores en los métodos de discriminación utilizando covariables en la cual los métodos para llevar a cabo un análisis de covarianza se usan junto con los de discriminación. Las variables utilizadas fueron lecturas de azúcar en la sangre de las unidades muestrales (conejos) a diferentes períodos después de aplicarles una dosis de insulina. El objetivo del estudio era encontrar una combinación lineal entre tres lecturas (para simplificar los cálculos sólo utilizan tres mediciones) de azúcar en la sangre que mediese mejor el efecto de la insulina (ésta se aplicó en 4 dosis diferentes). Este estudio es muy interesante por las características del problema que se estudia, aunque no es un ejemplo de discriminación puro.

Lubischew (1962), utiliza la función discriminante lineal para discriminar entre dos especies del género Haltica*, las cuales se pueden identificar con certeza únicamente mediante los órganos copulativos del macho de la especie. El objetivo de Lubischew era encontrar un medio de efectuar la identificación en base exclusivamente en mediciones externas y de una manera confiable. Las variables utilizadas fueron 21 entre las cuales se escogieron 4 en base al coeficiente de discriminación sugerido por Lubischew (1959), que tiene la desventaja de

* de ordevacea y carduorum

considerar individualmente a la variable en estudio. En base a tal criterio se eligieron las variables: distancia de la garganta o ranura transversa desde la orilla posterior del protorax, la longitud de los élitros, la longitud de la segunda articulación de las antenas. Los resultados obtenidos fueron satisfactorios.

Bartlett y Please (1963), estudiaron el caso en el cual las poblaciones no difieren en cuanto a la media. El ejemplo que tratan es el de discriminar entre gemelos* monocegóticos y dicigóticos. Las observaciones fueron las diferencias entre las observaciones en cada par de individuos siendo específicamente: altura, peso, longitud de la cabeza, anchura de la cabeza, perímetro craneal, distancia interpupilar, presión sanguínea - sistólica y diastólica -, intervalo del pulso, intervalos de la respiratorio, presión manual - mano derecha e izquierda-, agudeza visual - ojo derecho y ojo izquierdo -. Se escogieron 30 pares de gemelos hombres y 30 de mujeres donde 15 pares de cada sexo fueron juzgados como monocigóticos y 15 como dicigóticos. Sólo 10 de las mediciones anteriores fueron usadas efectivamente pues en las variables 8^a, 13^a y 14^a existían muchos valores faltantes y para la décima sus correlaciones con las otras mediciones variaba grandemente entre los cuatro grupos de gemelos.

---

* del mismo sexo

Los sexos se consideraron separadamente y se supuso un patrón de equicorrelación entre las variables (que no parecía irrazonable observando las matrices de correlación muestreal). La discriminación fue mejor en el caso de mujeres que en el de hombres. Se trató de utilizar una función discriminante única para ambos sexos lo cual mostró no ser satisfactorio.

Geisser y Desu (1968), en relación al mismo problema estudiado por Bartlett y Please presentan una solución bayesiana al problema. Los datos que utilizan y las variables consideradas fueron las mismas que en el estudio de Bartlett y Please lo cual permite la comparación de los métodos utilizados. En base a los resultados no hay diferencia aparente en la potencia discriminatoria de los dos procedimientos comparados aún cuando, como es obvio, la interpretación que se da a los resultados obtenidos según cada método es diferente.

Massy (1965), presenta una aplicación en el área de mercadeo (marketing) y publicidad en el que el análisis se uso para discriminar entre radioescuchas de 5 estaciones de FM en Boston. El estudio es interesante pues presenta el ataque al problema de tener un gran número de variables (47), lo cual se resolvió utilizando en la primera etapa análisis factorial para así determinar un número reducido de va

riables (12) en base a las cuales efectuar la discriminación mediante 5 funciones discriminantes lineales (tantas como poblaciones). Una novedad es el uso de las llamadas matrices de confusión para analizar el poder discriminante de las funciones y establecer posibles relaciones entre las poblaciones de interés. Además Massy en base a los coeficientes asociados a las variables para cada una de las poblaciones propone un esbozo de las características socioeconómicas de los oyentes para cada estación de radio FM. Este estudio es muy atractivo pues muestra un enfoque fresco y una aplicación novel del análisis discriminante. La hipótesis de normalidad que se hace acerca de las poblaciones tal vez valdría reconsiderarse así como la suposición de igualdad de probabilidades a priori de proveniencia de alguna de las poblaciones. Como ya se mencionó anteriormente los coeficientes de las variables en las funciones discriminantes (en este caso múltiples) se interpretan intuitivamente.

Addelman y Morris (1968), presentan un estudio para determinar las funciones lineales de características de funcionamiento nacional que discriminen mejor entre grupos de naciones en vías de desarrollo. El método que utilizan es el basado en las variables canónicas (denotado anteriormente como de vectores característicos). Todas las variables utilizadas están relacionadas con funcionamiento económico. Lo

más sobresaliente es la utilización iterativa llevada a cabo con lo cual reubican ciertas naciones y clasifican naciones no asignadas inicialmente. Obtienen además probabilidades individuales de pertenecer a un grupo dado para cada nación lo que en este caso es muy importante, pues proporciona un medio para evaluar el desarrollo potencial de países previamente no clasificados, así como para reevaluar la estimación original (i.e., los datos que sirven como muestras de ensayo) de las perspectivas de desarrollo de una nación.

Higgins (1970), aplicó el método de FDL para discriminar entre empresas de defensa y no de defensa en relación al empleo en cada una de ellas. Las variables utilizadas fueron de caracter económico (longitud promedio de oscilación en meses, promedio de razones de crecimiento, amplitud promedio de las oscilaciones, promedio de tasa mensual de cambio y error estándar muestral de la variable anterior) los resultados obtenidos son bastante satisfactorios, pues permiten discriminar entre las dos poblaciones consideradas en cada caso (se analizó el empleo total y el empleo en producción).

Anderson (1972), analiza la aplicación de la técnica de discriminación al problema de diagnóstico médico. Debido a que hay un riesgo de que personas que sufren artritis reumatoide contraigan kerato-conjuntivitis sicca y aunque este mal puede ser diagnosticado de una

manera confiable por un oftalmólogo, se presenta el problema de que estos servicios no están disponibles para todos los pacientes con artritis reumatoide. Sería deseable contar con un procedimiento simple - que permitiese al grupo de médicos de un centro reumático, quienes no sean oftalmólogos, decidir cuáles pacientes enviar a la sección (hospital, centro) oftálmico. Anderson aplica discriminadores logísticos (ver sección II.3 de este trabajo) para resolver el problema. El sistema de diagnóstico se basó en 10 síntomas del tipo ausencia o presencia (i.e., variables dicotómicas). Las muestras de ensayo contaban con 40 pacientes que padecían el mal oftálmico y 37 que no lo padecían. Incluyéndose decisiones no conclusivas, los resultados demuestran ser satisfactorios. Mayores detalles se pueden encontrar en el trabajo de - Anderson, Whaley, Williamson y Buchanan (1972).

Los ejemplos presentados son sólo una muestra de los posibles campos de aplicación de los métodos de discriminación. Otros ejem - plos pueden encontrarse en los textos de Bock (1973), Seal (1964), Cooley y Lohnes (1962 y 1966), un compendio hasta el año de 1950 se encuentra en el trabajo (catalogado como excelente) pero bastante inaccesible de Hodges (1950).

B I B L I O G R A F I A

CAPITULO IV.-

Kossack (1945)

Tintner (1946)

Durand (1941)

Rao y Slater (1949)

Banks (1950)

Anderson (1951)

Evans (1959)

Cochran y Bliss (1961)

Lubischew (1959) , (1962)

Bartlett y Please (1968)

Massy (1965)

Addelman y Morris (1968)

Higgins (1970)

Anderson (1972) , Anderson et al. (1972)

Bock (1973)

Seal (1964)

Cooley y Lohnes (1962 , 1966)

Hodges (1950)

OBSERVACIONES

Es evidente de los resultados presentados, el que el tema de Discriminación Estadística está lejano de ser agotado.

En cuanto a los posibles caminos de investigación se hallan:

i) la comparación metódica de métodos sugeridos hasta la fecha.

ii) el estudio a fondo de las técnicas para estimar las probabilidades de clasificación, correcta o incorrecta, y en su caso mejorarlas.

iii) continuar el estudio del problema para k-poblaciones sobre todo mediante el enfoque clásico

En relación a las aplicaciones, estas son muy amplias, como puede observarse de los ejemplos presentados en el capítulo IV y en el I.

Apéndice. Programas de computadora.

Se presentan en este capítulo los datos referentes a varios programas para computadora con los cuales se puede efectuar clasificación.

1. Nombre: Análisis Regional y Clasificatorio mediante iteraciones discriminantes.

Número de código del programa y lugar en el que puede ser localizado:

UCSM 503

Centro de Cómputo, Universidad de Chicago, Chicago, Illinois.

Máquina para la cual el programa fue escrito:

IBM 7094

Lenguaje: FORTRAN II

Limitaciones: hasta 150 observaciones

hasta 20 poblaciones

hasta 10 dimensiones por observaciones

los datos deben estar ortonormalizados.

Forma de Salida: i) Carga de las dimensiones en cada función de dis

criminación.

ii) Resultados de cada observación con respecto a cada función discriminante.

iii) Suma de las raíces propias en cada etapa.

Descripción general y notas: Este programa identifica y evalúa clasificaciones. La iteración discriminante aplica un procedimiento discriminante a todos los objetos de una clasificación con los cuales el procedimiento esta funcionalmente relacionado.

2. Nombre: Análisis Discriminante para varios grupos.

Número de código del programa y lugar donde puede ser localizado:

BMD 05 M

Recurso de Cómputo de Ciencias de la Salud, UCLA,

Los Angeles, California.

Máquina para la cual fue escrito el programa:

IBM 7094

Lenguaje: FORTRAN IV

Limitaciones: Hasta 5 grupos (g)

g a 25 dimensiones

hasta 175 observaciones por grupo

Forma de Salida:

i) Matriz de productos cruzados

ii) Matriz de covarianza y su inversa

iii) D^2 de Mahalanobis

iv) Coeficientes y constantes

v) Resultados para cada observación.

Descripción general y notas: Este programa calcula un conjunto de funciones lineales así que un individuo puede ser clasificado en uno de varios grupos. El procedimiento de asignación de grupo está derivado de un modelo de una distribución Normal multivariada de observaciones, dentro de grupos tales que la matriz de covarianza es la misma para todos los grupos. El individuo es clasificado en el grupo para el cual la densidad de probabilidad estimada es mayor.

3. Nombre: Análisis Discriminante para Dos Grupos.

Número de código del programa y lugar en el que puede ser localizado:

BMD 04 M

Recurso de Cómputo Ciencias de la Salud, UCLA,

Los Angeles, California.

Máquina para la cual fue escrito el programa:

IBM 7094

Lenguaje: FORTRAN IV

Limitaciones: hasta 25 (p) dimensiones

de p a 300 observaciones por cada grupo

Forma de Salida: Matriz de suma de cuadrados y productos cruzados de desviaciones y su inversa.

Descripción general y notas: Este programa proporciona la función discriminante entre dos grupos para la discriminación entre los índices medios de cada grupo. La diferencia entre las medias dividida por la desviación estándar de los índices es maximizada.

4. Nombre del programa: Mesa 97 (NYBMUL)

Número de código del programa y lugar en el que puede ser localizado:

UCSL 705

Centro de Cómputo de la Universidad de Chicago,

Chicago, Illinois.

Máquina para la cual fue escrito el programa:

IBM 7094

Lenguaje: FORTRAN IV

Limitaciones: Hasta cerca de 100 variables y cerca de 4,000 celdas, si se necesita almacenamiento adicional, partes del programa pueden ser removidas.

Forma de Salida:

i) Varianza de la variable canónica

ii) Criterio de Roy

iii) Coeficientes de la función discriminante

iv) Criterio de traza de Hotteling

v) Prueba ji-cuadrada de Bartlett para la significación de variables canónicas sucesivas.

vi) Forma canónica de las estimadas por mínimos cuadrados.

Descripción general y notas: Este programa efectúa un análisis de función discriminante para cada hipótesis entre celda. Correlación canónica, análisis de varianza multidimensional y regresión son opciones precedentes del programa.

5. Nombre del programa: Analizador Estadístico Multidimensial.

Número de código del programa y lugar en el que puede ser localizado:

Sin número de código

Centro de Cómputo de la Universidad de Harvard,

Cambridge, Massachusets.

Máquina para la cual fue escrito el programa

IBM 7090

Lenguaje: FORTRAN II

Limitaciones: hasta 80 dimensiones

hasta 50 grupos

(D^2 de Mahalanobis hasta para 30 dimensiones y 20 grupos).

Forma de Salida: Coeficientes y constantes

Resultados para cada observación

D^2 de Mahalanobis

Descripción general y notas: Este sistema puede también manejar análisis factorial, análisis de covarianza y problemas de discriminación.

6. Nombre: Programas de Estadística Multidimensional.

Número de código del programa y lugar en el que puede ser localizado:

MSP

Laboratorio de Biometría, Universidad de Miami,

Coral Gables, Florida.

Máquina para la cual fue escrito el programa:

IBM 7090

Lenguaje: FORTAN IV

Limitaciones: 2 poblaciones

hasta 100 dimensiones

hasta 99,999 observaciones por población

sin datos extraviados

Forma de Salida: i) Función discriminante

ii) Resultados para cada observación

Descripción general y notas: Esta rutina proporciona análisis discriminante para dos grupos. Rutinas adicionales del sistema incluyen análisis de varianza multidimensional, correlaciones canónicas y análisis factorial.

7. Nombre: SSLPAC - Programa de Ciencias Sociales, Paquete Estadístico de Biblioteca.

Número de código del programa y lugar en el que puede ser localizado:

UCSL 509

Centro de Cómputo Universidad de Chicago,

Chicago, Illinois.

Máquina para la cual fue escrito el programa.

IBM 7094

Lenguaje: FORTRAN II

Limitaciones: hasta 80 dimensiones (d)

número de discriminantes $\leq$ d ó $\leq$

número de poblaciones menos 1

Forma de Salida:

i) Raíces características

ii) Coeficiente eta

iii) Porcentaje de varianza extraída representado por cada discriminante

iv) Cociente de verosimilitudes.

Descripción general y notas: Este programa también efectúa análisis factorial y correlaciones canónicas.

8. Nombre: Análisis Discriminante Paso por Paso.

Número de código del programa y lugar en el que puede ser localizado:

BMD 07 M

Recurso de Cómputo de Ciencias de la Salud, UCLA,

Los Angeles, California.

Máquina para la cual fue escrito el programa:

Lenguaje: FORTRAN IV

Limitaciones: hasta 80 dimensiones

2 a 80 grupos

Forma de Salida:

i) Correlación entre grupos y matrices de covarianza

ii) Estadísticas U y F

iii) Matriz de Clasificación .

iv) Raíces características

v) Correlaciones canónicas y coeficientes para variables canónicas.

vi) Probabilidad a posteriori de estar en cada grupo para cada observación

vii) D^2 de Mahalanobis

Descripción general y notas: Este programa lleva a cabo análisis discriminante múltiple paso por paso. En cada paso la variable con el valor F más grande entra en el conjunto de variables discriminadoras. Una variable es suprimida si su valor de F se vuelve muy bajo. Este programa también calcula correlaciones canónicas y coeficientes para variables canónicas.

9. Nombre: Clasif

Lugar donde puede ser localizado el programa:

Multivariate Data Analysis

William W. Cooley y Paul R. Lohnes, Wiley 1966.

Máquina para la cual el programa fue escrito:

IBM 1620

Lenguaje: FORTRAN IV

Limitaciones: hasta 99 variables

hasta 99 grupos

hasta 99,999 individuos

Forma de Salida: El programa cuenta con dos opciones, bajo una de ellas suprime la escritura y perforación de probabilidades para los individuos y da cuenta de los aciertos y equivocaciones y bajo la otra imprime probabilidades individuales y suprime la cuenta. Esto se debe a que está diseñado ba jo la suposición de que si se desean las probabi lidades un criterio de membrecía real no está disponible y viceversa.

Descripción general y notas: Este programa calcula probabilidades de clasificación para individuos, asigna cada individuo a un grupo en base a su más alta probabilidad de pertenencia y lleva cuenta de aciertos y equivocaciones (si la pertenencia real de los individuos a un grupo se conoce). El método que usa es el propuesto por Geisser (1964), para estimadas de centroides de grupo separados y una matriz de varianza (dispersión) común para los grupos.

Otro programa aparece en Blackith y Reyment (1971), junto con una discusión del problema enfocada a biólogos.

B I B L I O G R A F I A G E N E R A L .

Se presenta aquí la bibliografía para este trabajo. Inicialmente se planeó una bibliografía exhaustiva del tema pero finalmente se optó por presentar la que aquí aparece. Será motivo de otro trabajo de este autor la presentación de una bibliografía mas extensa que pueda merecer el adjetivo de exhaustiva.

No debe sorprender el que se encuentren algunos artículos íntimamente ligados con reconocimiento de patrones, ya que los problemas que se enfrentran, en cuanto a estimación de probabilidades de clasificación, selección de variables y tratamiento de valores faltantes, son comunes tanto a la técnica de Discriminación como a la de reconocimiento de patrones. Por lo tanto, puede aprenderse mucho de los avances que se han logrado en la solución de los problemas antes citados y al conjuntar tales avances allanar el camino para mejores soluciones.

Adelman, I. y Morris, C.T.(1968), Performance criteria for evaluating economic development potential: an operational approach. Quaterly Journal of Economics V82, 268-280.

Allais, D.C.(1964), Selection of measurements for prediction. Stanford Electron. Res. Lab., Calif., Rep. SEL-64-115, TR6103-9.

_____(1966), The problem of too many measurements in pattern recognition . en IEEE Int. Conv. Rec. V14 pt. 2, 124-130

Anderson, J.A.(1969), Discrimination between k populations with constraints on the probabilities of misclassification. JRSS, B V31, 123-139.

_____(1972), Separate sample logistic discrimination. Biometrika V59, 19-35.

_____(1974), Diagnosis by logistic discriminant function: further practical problems and results. App. Stat. V23, 397-404.

Anderson, T.W.(1951) Classification by multivariate analysis. Psychometrika V16, 31-50

_____(1958), An Introduction to Multivariate Statistical Analysis, John Wiley & Sons Inc. New York.

_____ y Bahadur, R.A.(1962), Classification into two - multivariate normal distributions with different covariance matrices. Ann. Math. Stat. V33, 420-431.

Armitage, P.V.(1966) Recent developments in medical statistics. Rev. Int. Stat. Inst. V34, 27-42.

Aoyama, H.(1950) A note on classification data. Ann, Inst. Stat. Math. V2, 17-20.

Ashton, E.H., Healy, M.R. y Lipton, S.(1957), The descriptive use of discriminant functions in physical anthropology. Proc. Roy. Soc. B V146, 552-572.

Banerjee, K.S. y Marcus, L.F.(1965) Bounds in a minimax classification procedure. Biometrika V52, 653-

Banks, S.(1950), The relationship between proference and purchase of brands. Journal of Marketing V15, 145-157.

Bartlett, M.S.(1951), An inverse adjustment arising in discriminant analysis. Ann. Math. Stat. V22, 107-111.

_____ y Please, N.W.(1963), Discrimination in the case of zero mean differences. Biometrika V50, 17-

Baten, W.D.(1944), The discriminant function applied to spore measurements. Michigan Acad. of Sci., Arts and Letters V29, 3-7.

Baten. W.D.(1945), The use of discriminant functions in comparing judges' scores concerning potatoes.
JASA V40, 223-227.

_____ y Dewitt, C.C.(1944), Use of discriminant function in the comparison of proximate coal analyses.
Ind. Eng. Chem. Anal. Ed, V16, 32-34.

_____ y Hatcher, H.M.(1944), Distinguishing method differences by use of discriminant functions.
Jour. of Exp. Ed.

Berkson, J.(1957), Cost-utility as a measure of the efficiency of a test.
J A S A V42, 246-255.

Blackith, R.E. y Reyment, R.A.(1971), Multivariate Morphometrics. Academic Press, London.

Bock, R.D.(1973), Multivariate Statistical Methods in Behavioral Research. McGraw-Hill, New York.

Bose, R.C. y Roy, S.N.(1938), The distribution of the studentised D^2-Statistic.
Sankhya V4, 19-38.

Boullion, T.L., Odell, P.L. y Duran, B.S.(1975), Estimating the probability of misclassification and variate selection.
Pattern Recognition V7, 139-145.

Bowker, A.H.(1960), A representation of Hotellings' T^2 and Anderson's classification statistics W in terms of simple statistics.
en Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. I. Olkin et al.(eds), 142-149.
Stanford University Press, Stan. Calif.

_____ y Sitgreaves, R.(1961), An asymptotic expansion for the distribution function of the classification statistic W
Tech. Rep. No. 53, App. Math. and Stat. Labs. Stanford University.

Bradley, R.A. y Martin, D.C.(1973), Inference for multivariate dichotomous populations.
Proc. BISI 39a. sesión, Vol XLV.

Brier, G.W., Schoot, R.G. y Simmons, V.L.(1940), The discriminant function applied to quality rating in sheep.
Proc. Amer. Soc. An. Prod. V1, 153-160.

Burnaby, T.P.(1966), Growth invariant discriminant functions and generalized distances.
Biometrics V22, 96-110.

Chaddha, R.L. y Marcus, L.F.(1968), An empirical comparison of distance statistics for populations with unequal covariance matrices.
Biometrics V24, 683-694.

Chang, P.C. y Afifi, A.A.(1974), Classification based on dichotomous and continous variables.
JASA V69, 336-339.

Chan, L.S. y Dunn, O.J.(1972), The treatment of missing values in discriminant Analysis-I: the smapling experiment.
JASA V67, 473-477.

Chernoff, H.(1956), A classification problem.
Tech. Rep. 33, App. Math. and Stat. Lab. Stanford University.

_____(1973), Bound on the efficiency of a classification - procedure.
Proc. BISI 39a. sesión.

_____(1973), Some measures for discriminating normal multi_ variate distributions with unequal covariance matrices.
en Multivariate Analysis III, 337-344. ed. Krishnaiah, P.R., Aca_ demic Press, New York.

Chow, C.K.(1962),A recognition method using neighbour dependence.
IRE Trans. Electron. Comput. Vol. EC-I1, 683-690.

Chu, J.T. y Chueh, . .(1967), Error probabilities in decision fun ctions for character recognition.
J. Ass. Comput. Mach. V14, 273-280.

Clunies-Ross, C.W. y Riffenburgh, R.H.(1960), Geometry and linear discrimination.
Biometrika V47, 185-189.

Cochran, W.G.(196), On the performance of linear discriminant - functions.
Proc. BISI 34a. sesión, 435-447.

_____(1964a), Comparison of two methods of handling cova_ riates in discriminatory analysis.
Ann. Inst, Stat. Math. V16, 43-53.

_____(1964b), On the performance of the linear discriminant function.
Technometrics V6, 179-190.

_____(1968), Commentary on 'Estimation of error rates in discriminant analysis'.
Technometrics V10, 204-205.

_____ y Bliss, C.I.(1948), Discriminant functions with co_ variance.
Ann. Math. Stat. V19, 151-176.

_____ y Hopkins, C.E.(1961), Some classification problems with multivariate qualitative data.
Biometrics V17, 10-32.

Cooley, W.W. y Lohnes, P.R.(1962), Multivariate Procedures for the Behavioral Sciences, John Wiley & Sons, New York.

_____ y _____(1971), Multivariate Data Analysis, John Wiley & Sons Inc., New York.

Cooper, P.W.(1963), Statistical Classification with quadratic forms.
Biometrika V50, 439-448.

Cornfield, J.(1967), Discriminant functions.
Rev, Int. Stat. Inst. V35, 142-153.

Cox, D.R.(1966), Some procedures connected with the logistic qua_
litative response curve.
Res. Papers Statis. Festchr. Neyman, 55-71.

Cox, G.M. y Martin, W.P.(1937), Use of discriminant function for
differentiating soils with different azotobacter populations.
Iowa St. Coll. Jour. Sci. V11, 323-332.

DasGupta, S.(1965), Optimum classification rules for classifica_
tion into two multivariate populations.
Ann. Math. Stat. V36, 1174-1184.

_____(1968), Some aspects of discrimination function coeffi_
cients.
Sankhya A V30, 389-400.

Day, B.B. y Sandomire M.M.(1942), Use of discriminant function for
more then two groups.
JASA V37, 461-472.

Dempster, A.P.(1960), Random allocation designs I: On general cla_
sses of estimation methods.
Ann. Math. Stat. V31, 885-905.

_____(1964), Tests for the equallity of two covariance-
matrices in relation to best linear discriminator analysis.
Ann. Math. Stat. V35, 190-199.

Dhrymes, P.J.(1970), Econometrics, Statistical Foundations and -
Applications. Harper & Row, New York.

Duda, R.O. y Fossum, H.(1966), Pattern classification by iterati_
vily determined linear and piecewise linear discriminant functions.
IEEE Trans. Electron. Comput. Vol. EC-15, 220-232.

_____ y Hart, P.E.(1973), Pattern Classification and Scene -
Analysis. John Wiley & Sons Inc., New York.

Dunn, O.J.(1971), Some expected values for probabilities of co_
rrect classification in discriminant analysis.
Technometrics V13, 345-353.

_____ y Varady, P.D.(1966), Probabilities of correct classifi_
cation in discriminant analysis.
Biometrics V22, 908-924.

Dunsmore, I.R.(1966), A bayesian approach to classification.
JRSS B V28, 568-577.

Durand, D.(1941), Risk elements in consumer installment Financing,8.
National Bureau of Economic Research, 125-142.

Elashoff, J.D., Elashoff, R.M. y Goldman, G.E.(1967), On the cho_
ice of variables in classification problems with dicotomous variables
Biometrika V54, 668-670.

Ellison, B. E.( ), A multivariate k-population problem.
Technical Report No. 703006, Lockheed Aircraft Corp. , Calif.

Ellison, B.E.(1962), A classification problem in which informa_
tion about alternative distributions is based on samples.
Ann. Math. Stat. V33, 213-223.

_____(1965), Multivariate-normal classification with cova_
riances known.
Ann. Math. Stat. V36, 1787,1793.

Estes, S.E.(1965), Measurement selection for linear discriminants
used in pattern classification.
Tesis doctoral(Ph.D), Stanford University, Calif.

Evans, F.B.(1959), Psychological and objective factors in the pre_
diction of brand choice, Ford versus Chevrolet.
Journal of Bussines V32, 340-369.

Fix, E. y Hodges, J.L.(1951), Discriminatory analysis, nonparame_
tric discrimination: consistency propierties.
USAF School of Aviation Medicine, Project No. 21-49-004, Rep. No. 4

_____y _____(1952), Discriminatory analysis, nonparame_
tric discrimination: small sample performance.
USAF School of Aviation Medicine, Project No. 21-49-004, Rep. No. 11

Fisher, L. y Van Ness, J.W.(1973), Admisible discriminant analysis.
JASA V68, 603-607.

Fisher, R.A.(1936), The use of multiple measurements in taxonomic
problems.
Ann. of Eugenics V7, 179-188.

Frank, R.E., Massy, W.F. y Morrison, D.G.(1965), Bias in multiple
discriminant analysis.
Journal of Marketing Research V2, 250-258.

Foley, D.H.(1972), Considerations of sample and feature size.
IEEE Trans, Inform. Theory Vol. IT-18, 618-626.

Garret, H.E.(1943), The discriminant function and its use in psy_
chology.
Psychometrika V8, 65-79.

Geisser, S.(1973), A note on linear discriminants.
Proc. BISI, 39a. sesión Vol. XLV.

_____(1964), Posterior odds for multivariate normal classi_
fications.
JRSS B V26, 69-76.

_____(1966), Predictive discrimination.
en Multivariate Analysis, ed. P.R. Krisnaiah, Academic Press, N.Y.

_____(1967), Estimation associated with linear discriminants.
Ann. Math. Stat. V38, 807-817.

_____(1973), Multiple birth discrimination.
en Multivariate Statistical Inference, ed. D.G. Kabe y R.P. Gupta
49-58, North Holland/American Elsevier.

_____y Desu, M.M.(1968), Predictive zero-mean uniform dis_
crimination.
Biometrika V55, 519-524.

Gessaman, M.P. y Gessaman, P.H.(1972), A comparison of some multivariate discrimination procedures.
JASA V67, 468-472.

Gilbert, E.(1969), The effect of unequal variance-covariance matrices on Fisher's linear discriminant functions.
Biometrics V25, 505-515.

Glick, N.(1972), Sample-based classification procedures derived from density estimators.
JASA V67, 116-121.

_____(1973), Sample-based multinomial classification.
Biometrics V29, 241-256.

Goldstein, M.(1975), Comparison of some density estimate classification procedures.
JASA V70, 666-669.

Habbema, J.D.F., Hermans, J. y Van der Burgt, A.T.(1974)

Biometrika V61, 313-323.

Hamilton, M.(1950), The personality of dyspeptics.
Brit. Jour. Med. Psych. V23, 182-198.

Han, C.P.(1968), A note on discrimination in the case of unequal covariance matrices.
Biometrika V55, 586-587.

Han, C.P.(1969), Distribution of discriminant function when covariance matrices are proportional.
Ann. Math. Stat. V40, 979-985.

Harter, L.H.(1951), On the distribution of Wald's classification statistic.
Ann. Math. Stat. V21, 58-67.

Hartley, H.O. y Rao, J.N.K.(19), A now approach to the classification problem.
Proc. BISI, V39, 317-320.

_____ y _____(1968), Classification and estimation in analysis of variance problems.
Rev. Inst. In. Stat. V36, 141-147.

Hermans, J., Habbema, J.D.F. y Van der Burgt, A.T.(19), Cases of doubt in allocation problems, k populations.
Proc. BISI 39a. sesión, Vol. XLV.

Higgins, G.F.(1970), Discriminant analysis of employment in defense and nondefense industries.
JASA V65, 613-623.

Hills, M.(1966), Allocation rules and their error rates.
JRSS B, V28, 1-31.

_____(1967), Discrimination and allocation with discrete data.
App. Stat. V16, 237-250.

Highleyman, W.H.(1962), Linear decision function, with application to pattern recognition.
Proc. IRE V50, 1501-1514.

Hodges, J.L.(1950), Discriminatory analysis(survey of discrimina_ tory analysis).
USAF School of Aviation Medicine, Project No. 21-49-004, Rep. No. 1.

Hoel, P. y Peterson, R.P.(1949), A solution to the problem of op_ timum classification.
Ann. Math. Stat. V20, 433-438.

Horton, I.F., Russell, J.S. y Moore, A.W.(1968), Multivariate-co variance and canonical analysis: a method for selecting the most effective discriminators in a multivariate situation.
Biometrics V24, 845-858.

Hudimoto, H.(1956), On the distribution-free classification of an individual into one of two groups.
Ann. Inst. Stat. Math. V8, 105-112.

_____(1957), A note on the probability of the correct cla_ ssification when the distributions are not specified.
Ann. Inst. Stat. Math. V9, 31-36.

Hughes, G.F.(1968),On the mean accuracy of statistical pattern re_ cognizers.
IEEE Trans. Inform. Theory Vol. IT-14, 55-63.

_____(1969), Number of pattern classifier design samples per class.
IEEE Trans. Inform. Theory, 615-618.

Jenden, D.J., Fairchild, M.D., Mickey M.R., Silverman, R.W.,Yale, C.(1972), A multivariate approach to the analysis of drug effects on the electroencephalogram.
Biometrics V28, 73-80.

John, S.(1959), The distribution of Wald's classification statis_ tic when the dispersion matrix is known.
Sankhyā V21, 371-376.

_____(1960), On some classification statistics.
Sankhya V22, 309-317.

_____(1961), Errors in discrimination.
Ann. Math. Stat. V32, 1125-1144.

Johnson, P.O.(1950), The quantification of qualitative data in dis_ criminant analysis.
JASA V45,

Kabe, D.G.(1963), Some results on the distribution of two random matrices used in classification procedures.
Ann. Math. V34, 181-185, enmendado V35(1964) 924.

Kanal, L. y Chandrasekaran, (1971), On the dimensionality and - smaple size in statistical pattern classification.
en Proc. Nat. Electron. Conf. V24, 2-7, aparece también en Pattern Recognition V3, 225-234.

Kendall, M.G. (19), A Course in Multivariate Analysis. Griffin's Statistical Monographs & Courses, London.

_____(1966), Discrimination and classification. en Multivariate Analysis, ed. P.R. Krisnaiah, Academic Press, N.Y.

_____y Stuart, A.(1966), The Advanced Theory of Statistics V2, Hafner Publish. Company, New York.

Kossack, C.F.(1945), On the mechanics of classification. Ann. Math. Stat. V16, 95-98.

_____(1949), Some techniques for simple classification. Proc. of the Berkeley Symposium in Stat. and Prob., 345-352, Univ. of California Press, Berkeley.

Kshirsagar, A.(1972), Multivariate Analysis. Statistics: textbooks and monographs, volume 2., Marcel Dekker Inc., N. Y.

_____ y Arseven, E.(1975), A note on the equivalency of discrimination procedures.
The American Statistician V29, 38-40.

Kullback, S.(1952), An application of information theory to multi_ variate analysis. I y II
Ann. Math. Stat. V23, 88-102., 122-146, correcciones V27(1956), 860

_____(1959), Information Theory and Statistics. John Wiley & Sons Inc., N.Y.

Lachenbruch, P.A.(1966). Discriminant analysis when the initial-samples are misclassified.
Technometrics V8, 657-662.

_____(1967), An almost unbiased method of obteining confidence intervals for the probability of misclassification in discriminant analysis.
Biometrics V23, 639-646.

_____(1968), On the expected probabilities of miscla_ ssification in discriminant analysis, necessary sample size and relation with the multiple correlation coefficient.
Biometrics V24, 823-834.

_____(1974), Discriminant analysis when the initial - samples are misclassified II: non-random misclassification models.
Technometrics V16, 419-424.

_____y Mickey, M.R.(1968), Estimation of error rates in discriminant analysis.
Technometrics V10, 1-11.

_____, Sneering, C. y Revo, L.T.(1973), Robustness of linear and quadratic discriminant function to certain types of nonnormality.
Communications in Statistics V1, No. 1, 39-56.

Lissack, T. y Fu, K.S.(1972), A separability measure for feature and selection and error estimation in pattern recognition.
Sch. Elec. Eng. Purdue Univ., Lafayette, Ind., Tech. Rep. TR-EE 72-15

Lubischew, A.A.(1962), On the use of discriminant functions in tax_ xonomy.
Biometrics V18, 455-477.

Lunts, A.L. y Brailovskiy (1967), Evaluation of attributes in statistical decision rules.
Eng. Cybern(Rusia), No. 3, 98-109.

MacNaughton-Smith, P.(1963), The classification of individuals by the possesion of attributes associated with a criterion.
Biometrics V19, 364-366.

McLachlan, G.J.(1972), Asymptotic results for discriminant analysis when the initial samples are misclassified.
Technometrics V14, 415-422.

McLachlan (1974a), The asymptotic distributions of the conditional error rate and risk in discriminant analysis.
Biometrika V61, 131-135.

(1974b), An asymptotic unbiased technique for estimating the error rates in discriminant analysis.
Biometrics V30, 239-249.

(1974c), Estimation of the errors of misclassification on the criterion of asymptotic mean square error.
Technometrics V16, 255-260.

(1974d), The relationship in terms of asymptotic mean square error between the separate problems of estimating - each of the three types of error rate of the linear discriminant function.
Technometrics V16, 569-575.

(1975), Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis.
JASA V70, 365-369.

Mahalanobis, P.C.(1927), Analysis of race mixture in Bengal.
J. Asiat. Soc. Beng. V23, 301-333.

(1930), On tests and measures of group divergence.
J. Asiat. Soc. Beng. V26, 541-588.

Marks, S. y Dunn, O.J.(1974), Discriminant functions when covariance matrices are unequal.
JASA V69, 555-559.

Martin, E.S.(1936), A study of egyptian series of mandibles, with special reference to mathematical methods of sexing.
Biometrika V28, 149-178.

Massy, W.F.(1965), On methods: discriminant analysis of audience characteristics.
Journal of Advertising Research V5, 39-48.

Matusita, K.(1967), Classification based on distance in multivariate gaussian cases.
Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, V1, University of California Press, 299-304.

(1956), Decision rule, based on the distance, for the classification problem.
Ann. Inst. Stat. Math. V8, 67-77.

Matusita, K.(1964), Distance and decision rule.
Ann. Inst. Stat, Math. V16, 305-315.

_____(1966), A distance and related statistics in multivariat riate analysis.
en Multivariate Analysis, ed. P.R. Krisnaiah, Academic Press, N.Y.

Memon, A.Z. y Okamoto, M.(1970), The classification Statistic W* in covariate discriminant analysis.
Ann. Math. Stat. V41, 1491-1499.

Morrison, D.F.(1967), Multivariate Statistical Methods. McGraw-Hill, New York.

Okamoto, M.(1961), Discrimination for variance matrices.
Osaka Math. Journal V13, 1-39.

_____(1968), An asymptotic expansion for the distribution of the linear discriminant function.
Ann. Math. Stat. V34, 1286- ; corrección en V39, 1358-1359.

Pearson, K.(1926), On the coefficient of racial likeness.
Biometrika V18, 105-117.

Pelto, C.R.(1969), Adaptive nonparametric classification.
Technometrics V11, 775-792.

Press, S.J.(1964), Some hypothesis testing problems involving mul_ tivariate normal distributions with unequal and intraclass struc_ tured covariance matrices.
Technical Report No. 12, Dept. of Statistics, Stanford University.

_____(1972), Applied Multivariate Analysis. Holt, Rinehart & Wilson Inc., Series in Quantitative Methods for Decision-Making.

Quesenberry, C.P. y Gessaman, M.P.(1968), Nonparametric discrimi_ nation using tolerance regions.
Ann. Math. Stat. V39, 664-673.

Rao, C.R.(1946), Tests with discriminant functions in multivaria_ te analysis.
Sankhya V7, 407-414.

_____(1947), The problem of classification and distance bet_ ween two populations.
Nature V159, 30-31.

_____(1948a), Test of significnace in multivariate analysis.
Biometrika V35, 58-79.

_____(1948b), The utilization of multiple measurement in pro_ blems of biological classification.
JRSS b V10, 159-193.

_____(1948c), A statistical criterion to determine the group to which an individual belongs.
Nature V160, 835-836.

_____(1949), On some problems arising out of discriminating with multiple characters.
Sankhya V9, 343-366.

Rao, C.R.(1950), Statistical inference applied to classificatory problems. Part I: Null hypothesis, discriminatory problems and distance power tests.
Sankhya V10, 229-256.

_____(1951a), Statistical inference applied to classificatory problems II: The problem of selecting individuals for various duties in a specified ratio.
Sankhya V11, 107-116.

_____(1951b), The applicability of large sample test for moving average and auto-regressive schemes to series of short length and experimental study. Part III: The discriminant function approach in the classification of time series.
Sankhya V11, 257-272.

_____(1952), Advanced Statistical Methods in Biometric Research. John Wiley & Sons Inc., N.Y., Heffner (1971), N.Y.

_____(1953), Discriminant functions for genetic differentiation and selection. Part IV de Statistical inference applied to classificatory problems.
Sankhya V12, 229-246.

_____(1954), A general theory of discrimination when the information about alternative population distributions is based on samples.
Ann. Math. Stat. V25. 651-670.

_____(1954b), On the use and interpretation of distance functions in statistics.
Proc. BISI, 34a. sesión parte II, 90-97.

_____(1962), Use of discriminant and allied functions in multivariate analysis.
Sankhya A V24, 149-154.

_____(1964), Sir Ronald Aylmer Fisher-The architect of multivariate analysis.
Biometrics V20, No. 2, volúmen en memoria de R.A. fisher.

_____(1965), Linear Statistical Inference and Its Applications. John Wiley & Sons Inc., New York, 2a. ed. 1973.

_____(1966a), Discriminant function between composite hypothesis and related problems.
Biometrika V53, 339-345.

_____(1966b), Discrimination among groups and assigning new individuals. The role of Methodology of Classification in Psychiatry and Psycophatology, 229-240.
U.S. Dept. of Health Education and Welfare, Public Health Services.

_____(1972), Recent trends of research in multivariate analysis. Biometrics V28, 3-22.

_____ y Slater, P.(1949), Multivariate analysis applied to differences between groups.
Bri. J. Psych.(Stat. Sec.) V2, 17-29.

Richards, L.E.(1974), Refinement and extension of distribution-free discriminate analysis.
App. Stat. V23, 174-176.

Romeder, J.M.(1973), Methodes et programmes d'analyse discriminante. Dunod, París.

Roy, S.N.(1939), p-Statistics or some generalisations in analysis of variance appropiate to multivariate problems.
Sankhya V4, 381-396.

Rulon, P.J., Tiedeman, D.V., Tatsuoka, M. y Langmuir, C.R.(1964), Multivariate Statistics for Personnel Classification. John Wiley & Sons Inc., New York,

Samuel, E.(1963), Note on a sequential classification problem.
Ann. Math. Stat. V34, 1095-1097.

Saxena, A.K.(1967), A note on classification.
Ann. Math. Stat. V38, 1592-1593.

Seal, H.L.(1964), Multivariate Statistical Analysis for Biologists.
Methuen and Co. Ltd., London.

Shumway, R.H. y Unger, A.N.(1974), Linear discriminant functions for stationary time series.
JASA V69, 948-956.

Sitgreaves, R.(1952), On the distribution of two random matrices used in classification procedures.
Ann. Math. Stat. V23, 263-270.

_____(1961), Some results on the distribution of W-classification statistics.
USAF Sch. of Aviation Medicine, aparece también en Stud. Item Analysis and Prediction, ed. Solomon, H.(1961), 241-251.

Smith, C.A.B.(1947), Some examples of discrimination.
Ann. of Eugenics V13, 272-

Smith, H.F.(1941), A discriminant function for plant selection.
Ann. of Eugenics V7, 240-250.

Solomon, H.(1961), Studies in Item Analysis and Prediction.
Stanford University Press, Stanford, Calif., capítulos 15-19.

Sorum, M.(1971), Estimating the conditional probability of misclassification.
Technometrics V13, 333-343.

_____(1972a), Three probabilities of misclassification.
Technometrics V14, 309-316.

_____(1972b), Estimating the expected and the optimal probabilities of misclassification.
Technometrics V14, 935-943.

_____(1973), Estimating the expected probability of misclassification for a rule based on linear discriminant function: univariate normal case.
Technometrics V15, 329-339.

Srivastava, M.S.(1967), Classification into multivariate normal populations when the populations means are linearly restricted.
Ann. Inst. Stat. Math. V19, 473-478.

_____(1973), A sequential approach to classification: cost of not knowing the covariance matrix.
J. Multivariate Analysis V3, 173-183.

Stevens, W.L.(1940), The standarization of rubber flexing tests.
India Rubber World, August 1.

Stoller, D.C.(1954), Univariate two population distribution-free discrimination.
JASA V49, 770-777.

Tatsuoka, M.M.(1970), Discriminant analysis. The Study of Group Differences. Selected topics in advanced statistics, an elemen_ tary approach, No. 6. Publicado por el Institute for Personality and Ability Testing, Champaign Ill.

_____y Tiedeman, D.V.(1954), Discriminant analysis.
Rev. of Educational Research V24, 402-420.

Tintner, G.(1946), Some applications of multivariate analysis to economic data.
JASA V41, 472-500.

Toussaint, G.T.(1969), Machine recognition of independent and con_ textually constrained contour-traced handprinted characters.
Tesis M.A.Sc. Univ. British Columbia, Vancouver, Canada.

_____(1972), Feature evaluation criteria and contextual decoding algorithms in statistical pattern recognition.
Tesis doctoral, Univ. British Columbia, Vancouver, Canada.

_____(1974), Bibliography on estimation of misclassificatio cation.
IEEE Trans. on Inf. Theory Vol. IT-20, 472-479.

_____y Donaldson, R.W.(1970), Algorithms for recognizing co contour-traced handprinted characters.
IEEE Trans. Comput. Vol. C-19, 541-546.

Travers, R.M.W.(1939), The use of discriminant function in the treament of psychological group differences.
Psychometrika V4, 25-32.

Von Mises, R.(1945), On the classification of observation data in_ to dinstinct groups.
Ann. Math. Stat. V16, 68-73.

Wald, A.(1944), On a statistical problem arising in the classificatio cation of an individual into one of two groups.
Ann. Math. Stat. V15, 145-163.

Wallace, N. y Travers, R.M.W.(1938), A psychometric sociological study of a group of speciality salesmen.
Ann. of Eugenics.

Weiner, J.M. y Dunn, O.J.(1966), Elimination of variates in li_ near discrimination problems.
Biometrics V22, 268-275.

Welch, B.L.(1939), Note on discriminant function.
Biometrika V31, 218-220.

Welch, P.A. y Wimpress, R.S.(1961), Two multivariate statistical computer programs and their application to the vowel recognition problem.
The. Journal of the Acoustical Society of America V33, 426-434.

Wesler, O.(1959), A classification problem involving multinomials.
Ann. Math. Stat. V30, 128-133.