APROXIMACIÓN AL CONTROL ESTADÍSTICO DE CALIDAD MEDIANTE VARIABLES MULTIDIMENSIONALES

F.C. LARIO ESTEBAN

El objeto de este trabajo es definir unas técnicas estadístico-matemáticas que permitan plantear y resolver el tema del Control de Calidad multidimensional con variables correlacionadas. Planteada y justificada la necesidad de utilización de variables multidimensionales para caracterizar individuos estadísticos complejos, se pasa revista a técnica de Identificación y Análisis de Datos, -Análisis Discriminante y k-vecino más próximo (k-NN ó k-Nearest Neighbor)-, (se caracteriza el Reconocimiento de Formas, -Pattern-Recognition-, mediante los Identificadores anteriores actuando sobre Muestras de Aprendizaje) y al estimar los Errores de Mala-Identificación, se puede plantear ya un plan de Control de Calidad. Se ha obtenido de este modo un Programa que establece un Plan de Control de Calidad en la Recepción, sobre variable aleatoria multidimensional normal, mediante una Muestra de Aprendizaje inicial supervisada.

0. EL CONTROL ESTADISTICO DE CALIDAD, REVISION DE LAS TECNICAS UNIDIMENSIONALES CLASICAS

Entre las variables controlables en un proceso productivo, y junto a características tales como la Capacidad de Producción y los Costes, hay otros aspectos que no por más conocidos se utilizan en toda su potencia, tal es el caso del Control de Calidad en los Procesos de Fabricación.

La evolución y creciente desarrollo de una tecnología cada vez más automatizada e interrelacionada exige un continuo perfeccionamiento y avance en las Técnicas de Control, con la consiguiente elaboración de instrumentos matemático-estadísticos que permitan la solución de Problemas industriales que por su propia realidad alcanzan un evidente grado de complejidad en su tratamiento y modelización.

La utilización en el campo unidimensional de las técnicas estadísticas (Estimación, Muestreo y Pruebas de Hipótesis) como instrumentos al servicio del Control de Calidad permitieron resolver los problemas que los Ensayos Destructivos representaban en la Inspección al 100%.

Aparece de esta manera el Control Estadístico de Calidad, en el que a partir de Muestras relativamente pequeñas pueden inferirse características de la Población, Lote recepcionado o Producto fabricado. Inferencias, que para su validez estadística, han de tener en cuenta las características no tan sólo técnicas sino propias de la teoría estadística.

Ante las primeras exigencias de Calidad, manifiestas en los Contratos de Compra-Venta, se plantea la solución a la Recepción de Lotes, y como primer paso se plantearon las Técnicas Estadísticas del Control de Calidad en la Recepción, mediante variables unidimensionales que representaban Atributos o Variables Cualitativas. Si como es lógico existen, pues se han desarrollado, otras técnicas estadísticas que pretenden resolver otros temas concretos, -Control de Calidad en la recepción por medidas, e incluso se ha visto la necesidad de perfeccionar y desarrollar las técnicas iniciales obteniéndose los Planes de Muestreo Múltiples y Secuenciales-, aquí se insistirá en el C.E. de Calidad en la Recepción por atributos en su aspecto simple, ya que se tratará de aplicar estos conceptos y técnicas al caso de Variables Multidimensionales.

El C.E.C. en la Recepción mediante Atributos, tiene como objetivo definir un Plan de Control, que teniendo en cuenta los objetivos del Plan (Punto de la Curva Característica), optimice el problema realizando inferencias estadísticas a partir de los resultados obte-

- F.C. Lario Esteban. ETSII. Valencia. Cátedra de Organización de la Producción. Camino de la Vega s/n. Valencia
- Article rebut el Març del 1980.

nidos en Muestras aleatorias aleatorias extraídas del Lote.

Mediante los conceptos estadísticos del Control de Recepción por Atributos,

- Se calcula el número de individuos que constituirán la Muestra utilizada en el Plan de Muestreo.

- Se calcula la proporción de piezas defectuosas máxima, que puede encontrarse en la Muestra para que se cumplan las Condiciones de Eficacia del Plan.

A partir de esta proporción y el número de individuos de la Muestra se calcula el número máximo de individuos o piezas defectuosas que pueden encontrarse en la Muestra.

- Conocido el tamaño de la Muestra, se extraen los n-individuos del Lote, tomando las precauciones necesarias para que pueda considerarse extracción aleatoria.

- Se clasifican los individuos de la Muestra en buenos o defectuosos, mediante la aplicación de un ente físico o matemático que lo permita (Calibre Pasa-No Pasa, en el caso de piezas mecánicas, que se clasifican en buenas o defectuosas según el Calibre máximo que admiten).

- Determinado el número de Individuos de la Muestra considerados defectuosos, se compara esta cifra con el número máximo de individuos defectuosos que podrían encontrarse en el Lote.

- Según que los encontrados sean mayor-igual, o menor que el número máximo, podrá considerarse que el Lote no tiene la Calidad exigible por las Características de Eficacia, y por lo tanto será rechazado, o por el contrario nada autoriza a pensar que la Calidad del Lote (estimada por lo ocurrido con la Muestra) sea inferior a las condiciones expuestas en las Características de Eficacia.

Se tendrá pues resuelto el Problema del C.E.C. en la Recepción, utilizando atributos, y dejando de lado la problemática de la extracción aleatoria de los individuos de la Muestra, la realización de los Ensayos destructivos o no, -necesarios para la clasificación en buenos o defectuosos de los individuos- e incluso la problemática laboral y humana que la implantación del Control impone.

1. PROBLEMATICA ACTUAL, NECESIDAD DE LA UTILIZACION DE LAS VARIABLES MULTIDIMENSIONALES

El creciente desarrollo de la tecnología, junto con la utilización cada vez más racional y eficaz de los recursos van imponiendo nuevas necesidades de Control en los sistemas. La complejidad en las características de los Productos, junto a la automatización de los Procesos que los utilizan como Materias Primas, exigen un máximo conocimiento de las citadas M.P., que en definitiva se manifiesta en la necesidad de utilizar variables multidimensionales para conocer la Calidad de Conformidad de un Producto.

Si bien inicialmente la Calidad de un Producto podía definirse mediante una variable unidimensional que representaba el estado de una sola característica objetiva, una tecnología cada vez más compleja automatizada e interrelacionada, demanda la definición de la Calidad mediante varias características, en definitiva mediante variables multidimensionales.

Si los distintos componentes de la variable Multidimensional pueden considerarse independientes, cabría la posibilidad de definir la Calidad como una Suma de la Calidad referida a cada uno de los parámetros, en definitiva mediante tanto Planes de Muestreo y Calidad como componentes de la variable. No obstante quedaría la dificultad de fijar claramente las Condiciones de Eficacia, pues ya no sería de un Plan sino de varios; igualmente podría ocurrir que un individuo considerado como aceptable para cada una de las características, fuese realmente defectuoso en tanto en cuanto se encontrase en una situación límite para todas las características.

También suelen presentarse casos en los que los componentes están interrelacionados entre sí, y éste es el caso en que aparece con mayor justificación la utilización en el Control Estadístico de Calidad de variables Multidimensionales.

Si de la idea de mecanizado clásico de una Pieza se pasa al Mecanizado mediante Máquinas transfer, ya no puede hablarse de la característica diámetro mecanizado de un eje sino que debe ampliarse a un concepto donde aparezcan los distintos diámetros de agujeros, las distancias entre centros, sus profundidades, etc., y será con referencia a estas variables que puede definirse la Calidad de una Pieza.

De igual manera en la recepción de un Producto Químico, con diferentes componentes y propiedades frente a determinadas circunstancias o reactivos, no podrá utilizarse para su caracterización una sola variable unidimensional, y además entre los distintos componentes de la varaible multidimensional aparecerán inter-relaciones más o menos significativas.

Por último, planteada la calidad de un Hilado esta utilizará características del tipo Título, Torsión, Regularidad, Resistencia a la tracción, etc., para su definición. Podría pensarse en la utilización de variables multidimensionales para su caracterización, pues también parece objetivamente posible la existencia de relaciones entre varias de esas características.

En definitiva la tendencia a considerar la Teoría de Sistemas y Procesos dentro de la actividad industrial, con la consiguiente modelización matemática del tema, implica la necesidad de un perfeccionamiento en las Técnicas de Control Estadístico que asuma la existencia de características multidimensionales correlacionadas en la Calidad.

2. ENFOQUE DEL CONTROL DE CALIDAD EN LA RECEPCION MEDIANTE VARIABLES MULTIDIMENSIONALES

Cuando se ha comentado el C.E.C. en la Recepción mediante variables unidimensionales, se han citado las distintas fases que lo componen. Conceptualmente hablando, la única fase donde incide que la variable sea Multidimensional o no, es en la de Clasificación de los Individuos.

Asignar un Eje mecanizado a la Clase Ejes Bien Mecanizados o a la Clase Ejes Mal Mecanizados, según pase o no por un Calibre, es un proceso clásico y sencillo, pues en definitiva trata con variables unidimensionales.

Por el contrario cuando el Individuo estadístico exige para su caracterización de Calidad una variable Multidimensional, la Clasificación ya no aparece como un proceso tan claro y sencillo. Se entra ya en un Problema de Identificación, donde podrán distinguirse dos etapas:

- Definición de las distintas Clases en que puede particionarse el conjunto de individuos, que en este caso solo serán dos, Clase de los Individuos o Piezas Buenas y Clase de los Individuos o Piezas Malas. Se conoce con el nombre de Problema de Clasificación.

- Determinación de la Clase a la cual pertenece el individuo a Identificar, o sea Asignación de un Individuo no-identificado a una de las dos Clases a las que puede pertenecer. Problema de Asignación o Identificación (específico).

Debe hacerse constar que el Problema de Identificación o Asignación implica una Toma de Decisión, pues en definitiva en función de una serie de características del Individuo este se asigna a una de las clases a las que puede pertenecer. Como en todo problema de Decisión, en la Identificación existe un Riesgo de Mala-Identificación, y éste que se presenta sobre cada individuo de la Muestra, deberá relacionarse con los riesgos del Plan de Control de Calidad, ya que en éste la decisión de aceptar o rechazar el Lote se calculará a partir de los Individuos de la Muestra identificados en cada Clase.

Ya se ha hecho constar que con carácter previo a la Identificación deberá resolverse la Clasificación. Si bien para este caso las Clases solo son dos, la de Individuos buenos y la de Individuos defectuosos, la definición explícita y matemática de las Clases suele hacerse necesaria para la Identificación. Normalmente para ello se utiliza una Muestra de Aprendizaje, que en el caso que incluya la información sobre la pertenencia de cada uno de los individuos a una clase se llamará Supervisada y con Profesor, y de no ser así recibe el nombre de no-supervisada y sin Profesor. Para el tratamiento del C.E.C. será su-

ficiente considerarla supervisada con Profesor, pues puede considerarse la existencia de un archivo (Banco de Datos) donde se almacena toda la información sobre Individuos aceptados y sus características.

La primera parte del Problema consistirá pues en obtener la formulación explícita de cada Clase a partir de la Muestra de Aprendizaje Supervisada con profesor, y una vez conocidas éstas definir, en la fase siguiente, un Identificador o Norma de Identificación que nos permita asignar cada individuo no clasificado a una de las Clases. Necesariamente con la Identificación se producirá, como en toda Decisión una posibilidad de error, que tendrá que ser estimada.

Planteado y resuelto el Problema de Identificación, en sus dos partes, la Clasificación y la Identificación propiamente dichas y a partir de la Muestra Supervisada con Profesor, se podría asignar cada individuo de la Muestra de Inspección (así llamada a la del Plan de Control de Calidad, para distinguirla de la Muestra de Aprendizaje) a su Clase, si bien con un riesgo de Mala-identificación que tendrá que estimarse; se estará pues en condiciones de tratar el Plan de Control de Calidad, una vez que se haya establecido la relación entre Riesgos de 1^a y 2^a Especie del Plan de Calidad, y Riesgos de 1^a y 2^a Especie en la Identificación.

3. TECNICAS ESPECIFICAS MULTIDIMENSIONALES EN CONTROL ESTADISTICO DE CALIDAD EN LA RECEPCION

3.1. Distribución de los Individuos dentro de cada una de las dos Clases según una Ley Normal multidimensional

Planteado el problema en el caso de que se tenga una Muestra de Aprendizaje supervisada con Profesor, y pueda suponerse que la Distribución de los individuos dentro de cada Clase sigue una Ley Normal multidimensional, tendrá que definirse explícitamente cada una de las Clases, explicitarse un Identificador y estimarse los riesgos de Mala-Identificación.

3.1.1. Definición de la Función Discriminante Lineal y su Ley de Distribución

Si la Muestra de Aprendizaje está compuesta de N_1 individuos pertenecientes a la Clase w_1, y de N_2 de la clase w_2, y dado que los individuos se distribuyen dentro de la Clase según una Ley Normal, definir explícita y matemáticamente las dos Clases será equivalente a estimar el vector Media y la Matriz de Covariancia, $-(x_i,\ i=1,2;S_i=1,2$, respectivamente)-, con lo cual quedarán totalmente definidas las Clases.

Definida la Regla de Asignación o de Identificación de Bayes para Error mínimo como

- Si $P(w_1/x) \geq P(w_2/x) \rightarrow x \epsilon w_1$,

- Si $P(w_1/x) < P(w_2/x) \rightarrow x \epsilon w_2$

y teniendo en cuenta que

$$P(w_i/x) = \frac{P(w_i).p(x/w_i)}{p(x)}$$

puede pasarse a

$$P(w_1) \cdot p(x/w_1) \gtrless P(w_2) \cdot p(x/w_2)$$

y extrayendo logaritmos neperianos a

$$\ln \frac{p(x/w_1)}{p(x/w_2)} \gtrless \ln \frac{P(w_2)}{p(w_1)}$$

y si $p(x/w_1)$ sigue una Ley Normal de media μ_i y Matriz de Covariancia común Σ, puede escribirse

$$\ln \frac{\exp\left[\frac{1}{2} - (x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right]}{\exp\left[-\frac{1}{2} (x-\mu_1)'\Sigma^{-1}(x-\mu_1)\right]} \gtrless$$

$$\gtrless \ln \frac{P(w_2)}{P(w_1)}$$

y si $P(w_1) = P(w_2) = 0,5$

$$\left[x - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1} (\mu_1-\mu_2) \gtrless \ln 1 = 0$$

y llamando

$$D_T(x) = \left[x - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1} (\mu_1-\mu_2),$$

siendo $D_T(x)$ la Función Discriminante lineal, y la Identificación mediante $D_T(x)$ sería

- $x \epsilon w_1$, si $D_T(x) \geq 0$
- $x \epsilon w_2$, si $D_T(x) < 0$

El caso práctico y real, donde el conocimiento de la población no es total, impide trabajar con los Parámetros de la Población, μ_i y Σ, siendo necesario trabajar con las Estimaciones, $\bar{x}_i$, S, obtenidas a partir de los individuos de la Muestra de Aprendizaje.
Puede definirse así la Función Discriminante Muestral, como

$$D_S(x) = \left[x - \frac{1}{2}(\bar{x}_1+\bar{x}_2)\right]' S^{-1} (\bar{x}_1-\bar{x}_2)$$

Por lo tanto definidas las distribuciones de los individuos dentro de cada Clase como Normales y Multidimensionales, y si la Probabilidad de cada Clase $P(w_i)=0,5$, $i = 1,2$, puede definirse el Identificador Función Discriminante como

$$D_T(x) = \left[x - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1} (\mu_1-\mu_2)$$

Si tal como se indicó interesa estimar la Probabilidad de Error de Mala-Identificación, mediante $D_T(x)$, debe tenerse en cuenta tal y como Lachenbruch demuestra en su "Discriminant Analysis", que la variable $D_T(x)$ es normal, y los valores de su media y su desviación tipo son:

$$E(D_T(x)/w_1) = \left[\mu_1 - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1} (\mu_1-\mu_2) =$$

$$= \frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1} (\mu_1-\mu_2) = \frac{1}{2}\delta^2$$

$$E(D_T(x)/w_2) = \left[\mu_2 - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1} (\mu_1-\mu_2) =$$

$$= -\frac{1}{2}(\mu_1-\mu_2)' \Sigma^{-1} (\mu_1-\mu_2) = -\frac{1}{2}\delta^2$$

$$E\{D_T(x) - D(\mu_i)\}^2 = (\mu_1-\mu_2)\ \Sigma^{-1} (\mu_1-\mu_2) =$$

$$= \delta^2$$

Si debido al caracter práctico interesa la variable $D_S(x)$, donde se encuentran los valores estimados, la $D_S(x)$ se distribuye condicionalmente (con respecto a $\bar{x}_1, \bar{x}_2$ y S) de manera normal, siendo

$$E\{D_S(x/x \epsilon w_i)\} = \left[-\frac{1}{2}(\bar{x}_1+\bar{x}_2)+\mu_i\right]'$$

$$S^{-1} (\bar{x}_1 - \bar{x}_2) = D_S(\mu_i)$$

$$\text{Var}\{D_S(x/x \epsilon w_i) =$$

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

y dado que μ_i y Σ son la media y la Matriz de Covariancia de la población, que son desconocidas

$$\hat{D}_i = \text{Estimación}\{E[D_S(x/x \epsilon w_i)]\} =$$

$$\sum_{\alpha=1}^{n_i} D_s(x\alpha)/n_i \qquad i=1,2$$

$$\hat{V}_{Di} = \text{Estimación}\{V[D_S(x/x \epsilon w_i)]\} =$$

$$\sum_{i=1}^{2} \sum_{\alpha=1}^{n_i} [D_s(x_i) - \hat{D}_i]^2 / (n_1+n_2)$$

Si bien tal como se ha indicado la Distribución Condicional de $D_s(x)$, -según valores de $\bar{x}_1, \bar{x}_2$, S-, sigue una Ley Normal, la Incondicional de $D_s(x)$ no sigue una Ley Normal, y Okamoto ha definido una expresión asimptótica para la distribución de $D_s(x)$, y Lachenbruch ha expresado las medias y las varian-cias incondicionales de $D_s(x)$ para muestras de tamaño n_1 y n_2

$$E[D_s(x), x \epsilon w_1] = \frac{1}{2} C_1 \left(\delta^2 - \frac{k(n_2-n_1)}{n_1-n_2}\right)$$

$$E[D_s(x), x \epsilon w_2] = \frac{1}{2} C_1 \left(-\delta^2 - \frac{k(n_2-n_1)}{n_1-n_2}\right)$$

siendo k = dimensión de la variable representativa del individuo estadístico

$$C_1 = \frac{n_1+n_2-2}{n_1+n_2-k-3}$$

$\delta^2 = (\mu_2-\mu_1)' \Sigma^{-1} (\mu_2-\mu_1)$ distancia de Mahalanobis

$$\text{Var}[D_s(x)] = C_2\left[\delta^2 + \frac{k(n_1+n_2)}{n_1+n_2}\right]$$

siendo

$$C_2 = \frac{(n_1+n_2-3)\quad (n_1+n_2-2)^2 \quad -}{(n_1+n_2-k-2)(n_1+n_2-k-3)\quad (n_1+n_2-k-5)}$$

3.1.2. Errores de Clasificación utilizando como Identificadores la Función Discriminante

Definida la $D_T(x)$ como el logaritmo neperiano de la Norma que miniza la probabilidad total de Mala Clasificación, o sea

$$D_T(x) = \left[\bar{x} - \frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1}(\mu_1-\mu_2) \gtrless \ln \frac{P(w_2)}{P(w_1)}$$

se determinan las Probabilidades de Mala-Clasificación dentro de cada grupo como

$$E_1 = \int_{L_1} f_1(x/w_1)\,dx, \quad E_2 = \int_{L_2} f_2(x/w_2)\,dx$$

y la Proporción de Error como

$$E = P(w_1) \cdot E_1 + P(w_2)E_2$$

teniendo en cuenta la definición del Identificador y del Error, se tiene

$$E_1 = \int_{L_2} f_1(x/w_1)\,dx = P_r\left\{D_T < \ln \frac{1-P(w_1)}{P(w_1)}\right\} =$$

$$= P_r\left\{\frac{D_T(x) - \frac{1}{2}\delta^2}{\delta} < \ln \frac{\frac{1-P(w_1)}{P(w_1)} - \delta^2/2}{\delta}\right\}$$

y como se ha demostrado que la $D_T(x)$ sigue una Ley Normal de media $\frac{1}{2}\delta^2$ y una variancia δ^2, en la clase w_1, la Probabilidad buscada es la Acumulada en la Normal centrada y reducida, o sea

$$E_1 = P_r\left[\frac{D_T(x) - \frac{1}{2}\delta^2}{\delta} < \frac{\ln \frac{1-P(w_1)}{P(w_1)} - \frac{1}{2}\delta^2}{\delta}\right]$$

$$= \phi\left(\frac{\ln \frac{1-P(w_1)}{P(w_1)} - \frac{\delta^2}{2}}{\delta}\right) = E_1$$

y similarmente

$$E_2 = \phi\left(\frac{-\ln \frac{1-P(w_1)}{P(w_1)} + \frac{\delta^2}{2}}{\delta}\right) = E_2$$

o sea que en el caso de ser conocidos los Parámetros de las Distribuciones en w_1 y w_2, -o sea μ_1, μ_2, Σ-, pueden calcularse las Probabilidades de Error al utilizar el Identificador Función Discriminante.

3.1.2.1. Estimación del Error de Clasificación utilizando la Función Discriminante $D_s(x)$

En 1963 Okamoto en un artículo publicado en Ann. Math. Statis. dió una expresión asimptótica de la Ley que sigue la Distribución Incondicional de $D_s(x)$, que puede sintetizarse como:

Cuando se utiliza la $D_s(x)$ comparándola con el valor cero, de manera que x_0 se clasifica en w_1 si $D_s(x_0) \geq 0$ y en w_2 si $D_s(x_0) < 0$, las probabilidades de error vienen dadas por

$$P_r\{D_s(x) < 0 \mid w_1\} = \phi\{-\frac{\delta}{2}\} + \frac{a_1}{N_0} + \frac{a_2}{N_1} + \frac{a_3}{N_0 N_1 - 2} +$$

$$+ \frac{b_{11}}{N_0^2} + \frac{b_{22}}{N_1^2} + \frac{b_{12}}{N_0 N_1} + \frac{b_{13}}{N_0(N_0+N_1} +$$

$$+ \frac{b_{23}}{N_1(N_0+N_1^{-2}} + \frac{b_{33}}{(N_0+N_1-2)^2} + O_3$$

$$P_r\{D_s(x) \geq 0 \mid w_2\} = \phi\{-\frac{\delta}{2} + \frac{a_2}{N_0} + \frac{a_1}{N_1} + \frac{a_3}{N_0+N_1-2} +$$

$$+ \frac{b_{22}}{N_0^2} + \frac{b_{11}}{N_1^2} + \frac{b_{12}}{N_0 N_1} + \frac{b_{23}}{N_0(N_0+N_1-2)} + \frac{b_{13}}{N_1(N_0+N_1-2)} +$$

$$+ \frac{b_{33}}{(N_0+N_1-2)^2} + O_3$$

donde

$$a_1 = (2\delta^2)^{-1}(d_0^4 + 3pd_0^2)$$

$$a_2 = (2\delta^2)^{-1}(d_0^4 - (p-4)d_0^2)$$

$$a_3 = \frac{1}{2}(p-1)d_0^2$$

$$b_{11} = (8\delta^4)^{-1}\left[d_0^8 + 6(p+2)d_0^6 + (p+2)(9p+16)d_0^4 + 20p(p+2)d_0^2\right]$$

$$b_{22} = (8\delta^4)^{-1} \left[d_0^8 - 2(p-10)d_0^6 + (p-6)(p+16)d_0^4 + 4(p-4)(p-6)d_0^2\right]$$

$$b_{12} = (4\delta^4)^{-1} \left[d_0^8 + 2(p+8)d_0^6 - 3(p^2-10p-16)d_0^4 - 12(p-6)pd_0^2\right]$$

$$b_{13} = (4\delta^2)^{-1} \left[(p-1)\right]\left[d_0^6 + 3(p-4)d_0^4 + 6(p+4)\cdot d_0^2\right]$$

$$b_{23} = (4\delta^2)^{-1} (p-1)\left[d_0^6 - (p-8)d_0^4 - 2(p-4)d_0^2\right]$$

$$b_{33} = \frac{1}{8}(p-1)\left[(p+1)\, d_0^4 + 4\,(p-12)d_0^2\right]$$

siendo

p = dimensión de la variable multidimensional a clasificar

$$d_0^j = \left[\frac{d^j}{dc^j}\,\phi\,(c)\right]_{c=\delta/2} \quad ; \quad j = 2,4,6,8$$

Posteriormente Lachenbruch y Mickey en su artículo de Tecnométrica definen y comparan distintos tipos de estimadores para P_1, comparando las distintas técnicas mediante métodos Monte-Carlo.

Definido,

$$P_i = \phi\{(-1)^i D_s(\mu_i)/\sqrt{D_s}\} = \phi\left[(-1)^i \cdot \frac{(\mu_i - \frac{1}{2}(\bar{x}_1+\bar{x}_2))' \; S^{-1}(\bar{x}_1-\bar{x}_2)}{\sqrt{(\bar{x}_1-\bar{x}_2)' \; S^{-1}\Sigma\, S^{-1}(\bar{x}_1-\bar{x}_2)}}\right]$$

siendo $\phi(y) = \int_{-\infty}^{y} (2\pi)^{-1/2} \exp(-t^2)\, dt$,

μ_i = media de la Clase w_i, i = 0,1

Σ_i = Σ = matriz de Covariancia de la Clase w_i, i = 0,1

x_i = valor estimado de Media de w_i, obtenido a partir de los N_i individuos de la Muestra de Aprendizaje.

S = valor estimado de la Matriz de Covariancia Común

$$S = \frac{\sum_{i=0}^{1} N_i \hat{\Sigma}_i}{\sum_{i=1}^{0} N_i}$$

Σ_i = valor estimado de la Matriz de Covariancia de la Clase w_1, mediante los N_i individuos de la Muestra de Aprendizaje.

presenta el inconveniente de utilizar los valores de la Población, que en general y desde el punto de vista práctico no son conocidos. Como consecuencia se hace necesaria la estimación de P_i, o sea ha de utilizarse $\hat{P}i$.

Mac-Lachlan define un Estimador de P_i que minimiza el error cuadrástico asimplótico cuando Σ es desconocido,

$Q_M = \phi(-D/2) + \hat{A} - \hat{B} - \hat{C}$, donde

$\hat{A}$ es el valor de Okamoto $P_r\{D_s(x) > 0 \quad w_2\}$ - en que se ha sustituido la δ^2 por D^2.

$$D^2 = (\bar{x}_2-\bar{x}_1)' \; S^{-1}(\bar{x}_2-\bar{x}_1)$$

$\hat{B}$ viene definida por Mac-Lachlan como

$$E\{\phi(-D/2)\} = B + \theta_3$$

$$\hat{B} = \frac{\phi(-D/2)}{16}\left[\left(\frac{1}{N_0}+\frac{1}{N_1}\right)\left(D - \frac{4(p-1)}{D}\right) + \frac{D}{2}\left[D^2 - 4(2p+1)\right]\frac{1}{N_0+N_1-2}\right] + \frac{\phi(-D/2)}{1024}\left\{D\left[D^2-4(2p+1)\right] + \frac{16(p-1)}{D}\left[p-1+\frac{4(p-3)}{D^2}\right]\right\}\left(\frac{1}{N_0}\,\frac{1}{N^2}\right)^2 + \frac{\phi(-D/2)}{1024}\left\{D^5 - 4(3p+7)D^3 + 16(2p^2+8p+15)\cdot D - \frac{64(p-1)(2p+1)}{D}\right\}\left(\frac{1}{N_0}+\frac{1}{N_1}\right)\cdot\left(\frac{1}{N_0+N_3-2} + \frac{D\cdot\phi(-D/2)}{12}\left[3D^6 - 4(12p+35)D^4 + 16(12p^2+72p+71)D^2 - 192(12p^2+12p+1)\right]\cdot\frac{1}{(N_0+N_1-2)^{-2}}\right]$$

- $\hat{C}$ se expresa según Mc. Lanenlan como

$$\hat{C} = \frac{p-1}{N_1} \frac{\phi(-D/2)}{64}\{\left[2D - \frac{8(p-2)}{D} - \frac{32(p-3)}{D^3}\right]\cdot$$

$$\cdot \left(\frac{1}{N_0}+\frac{1}{N_1}\right) + \left[D^3-8pD-\frac{16(2p-1)}{D}\right]\frac{1}{N_0+N_1-2}\}+$$

$$+ \frac{1}{32(N_0+N_1-2}\{4(4p-1)\frac{\phi(-D/2)}{64}\cdot$$

$$\cdot \left[\{2D^3-8(p+2)D+\frac{32(p-1)}{D}\}\cdot\right.$$

$$\cdot \left(\frac{1}{N_0}+\frac{1}{N_1}\right)+\{D^5-8(p-12)D^3+16(2p+1)D\}\cdot$$

$$\cdot\left.\frac{1}{N_0+N_1-2}\right] - \frac{\phi(-D/2)}{64}\cdot$$

$$\cdot \left[\{2D^5-8(p+6)D^3+96(p+1)D\}\left(\frac{1}{N_0}+\frac{1}{N_1}\right)+\right.$$

$$\left.+\{D^7-8(p+4)D^5+48(2p+3)D^3\}\frac{1}{N_0+N_1-2}\right]\} = \hat{C}$$

de manera que cuando N_0, N_1 aumentan la ex- - presión de la estimación del Error se acer-- ca a la $\phi(-D/2)$, $(-1)^i$.

3.2. Distribución desconocida de los individuos dentro de cada Clase

En los Problemas de Clasificación al comen-- tar la utilización del Clasificador de Ba-- yes debe hacerse constar la necesidad de co- nocer las Probabilidades de cada Clase $P(w_i)$, así como la densidad de Probabilidad de ca-- da individuo x, condicionado a pertenecer a una de las Clases. En el caso real y prác-- tico de Ley de Distribución conocida (Normal Multidimensional) y parámetros desconocidos (μ_i, Σ_i) se ha planteado la solución mediante el Clasificador o Identificador de la Fun--- ción Discriminante Muestral $D_s(x)$.

Para el caso que no se tenga ninguna infor-- mación sobre la Ley que sigue la Densidad de Probabilidad Condicional y se trate de esti- marla pueden utilizarse técnicas del k-Veci- no más próximo (k-NN, "k-Nearest Neighborg").

3.2.1. Definición del Estimador k-NN y su -- Norma de Decisión o Identificador del k-NN

Ante el desconocimiento que puede tomar la - Función Densidad de Probabilidad dentro de - una Clase, se plantea por Fukanaga una modi- ficación a los Estimadores de la "Ventana de Parzen", conocido como "Estimador del k-NN".

Dada una muestra de Aprendizaje integrada --- por N individuos clasificados, se determina - la distancia r entre el individuo no-clasifi- cado x, y el punto representativo del indivi- duo situado en orden de distancia k, perte- neciente a la Muestra de Aprendizaje. En defi- nitiva se trata de definir la distancia r en- tre x (individuo no-clasificado) y el indivi- duo de la Muestra de Aprendizaje que ocupa -- el lugar k-ésimo de vecindad. Para medir el concepto de Proximidad se podría utilizar con una métrica adecuada la expresión

$$\hat{P}_N(x) = \frac{k-1}{N} \frac{1}{A(k,N,x)}$$

en donde A(k,N,x) es el volumen del conjunto de puntos representantes de la Muestra de -- Aprendizaje cuya distancia a x es menor que r. Si se considera una métrica euclediana y de - acuerdo con lo expuesto por Fukunaga, A(k,N,x) se convierte en el volumen de una hiperesfe-- ra de radio r, de acuerdo con,

$$A(k,N,x) = \frac{2r^p \pi^{p/2}}{p.\Gamma(p/2)}$$

siendo: p la dimensión del individuo x
$\Gamma(p/2)$: Función Gamma $\Gamma(p/2)$

Loftsgaarden (1965) comprobó que si la k(N)-- satisface

$$\lim_{N\to\infty} k(N) = \infty \quad , \text{ y } \lim_{N\to\infty} k(N)/N = 0$$

$\hat{P}_N(x)$ es un estimador asintóticamente impar-- cial y consistente de p(x).

Se ve pues que la técnica del k-NN suministra un estimador muy sencillo de la Densidad Con- dicional de Probabilidad de x.

Sea

$\hat{P}_{N_1}(x/w_1)$ = Estimación de $p(x/w_1)$, obtenida a partir de una muestra de Aprendi- zaje de N_1 individuos pertenecien tes a w_1 y $N_2 \varepsilon w_2$.

$$\hat{P}_{N_1}(x/w_1) = \frac{k_1-1}{N_1} \frac{1}{A}$$

$$\hat{P}_{N_2}(x/w_2) = \frac{k_2-1}{N_2} \frac{1}{A},$$

siendo x un individuo que ha de clasificarse en w_1 ó w_2, sabiendo que los N individuos de la Muestra de Aprendizaje se encuentran repartidos en $N_1 \in w_1$ y $N_2 \in w_2$. La aplicación de la técnica del k-NN a la estimación de la $p(x/w_1)$ consistirá en determinar cuántos de los k-vecinos más próximos de x pertenecen a w_1, -o sea k_1 vecinos- y cuántos de ellos a la clase w_2, k_2.

Si se aplica entonces el test de Bayes:

$$P(w_1)p(x/w_1) \gtrless P(w_2)p(x/w_2) \rightarrow x \in \begin{cases} w_1 \\ w_2 \end{cases}.$$

sustituyendo

$$P(w_1) \text{ por } \hat{P}(w_i) = \frac{N_i}{N}$$

y $p(x/w_i)$ por $\hat{P}_{N_i}(x/w_i)$

se obtiene

$$\frac{N_1}{N} \frac{k_1-1}{N_1} \frac{1}{A} \gtrless \frac{N_2}{N} \frac{k_2-1}{N_2} \frac{1}{A}$$

siendo $A = A(p,N,x) = \frac{2r^p \pi^{p/2}}{p.r(p/2)}$,

donde = radio de la Hiperesfera centrada en x.
p = dimensión del individuo x
$r(p/2)$ = función Gamma $r(p/2)$

cuando se considera una Métrica Euclediana, y de acuerdo con lo expuesto por Fukunaga, se considera A como el volumen de una Hiperesfera.

De otra parte debe hacerse constar que Loftsgaarden comprobó en 1965 que si la k(N) satisface

$$\lim_{N \to \infty} k(N) = \infty, \text{ y } \lim_{N \to \infty} \frac{k(N)}{N} = 0,$$

$$\hat{p}N(x) = \frac{k-1}{N} \quad \frac{1}{A(p,N,x)}$$

es un estimador asimpóticamente imparcial y consistente de p(x).

Teniendo en cuenta que la Hiperesfera utilizada en el Identificador es la misma en ambos casos, pues tienen el mismo centro y radio, este queda

$$k_1 \gtrless k_2 \rightarrow x \begin{cases} w_1 \\ w_2 \end{cases}$$

con lo que el Identificador del k-NN consistirá en:

- Dado el individuo x, a clasificar, se considera a éste como centro de una hiperesfera que contenga k individuos (ya clasificados) de la Muestra de Aprendizaje, de los cuales los k_1-individuos pertenecerán a la clase w_1 y k_2-individuos a la Clase w_2, de manera que

$$k_1+k_2 = k$$

- Si $k_1 \geq k_2$, el individuo inclasificado x se asigna a la Clase w_1.

- Si $k_1 < k_2$, el individuo inclasificado x se asigna a la Clase w_2.

3.2.2. Errores de Clasificación utilizando la Norma de Identificación del k-Vecino más próximo (k-NN)

Fukunaga en sus trabajos junto con Kesell y Hosteltler trata de resolver el Error de Identificación (Asignación o Clasificación) utilizando la Norma del k-NN.

Se plantea la Norma del 1-Vecino más próximo, para continuar luego con el k-NN de carácter general.

Define el Riesgo Condicional o Error condicional, $\sqrt{(x)}$, al utilizar el 1-NN como la probabilidad de que el vecino de x pertenezca a la Clase w_1, -por ser su vecino de esta clase- aunque realmente el x pertenezca a la clase w_1, o la probabilidad de que ocurra lo contrario, y por lo tanto

$$r(x,x_k) = P(w_1/x_k).P(w_2/x) + P(w_2/x_k).P(w_1/x)$$

si puede considerarse que x y el individuo clasificado están lo suficientemente cercanos

$P(w_i/x_j) \tilde{=} P(w_i/x)$, ya que N será grande y r pequeño, con lo cual la proximidad entre x y x_k se dá, con lo que

$$r(x,x_k) = r(x) \tilde{=} 2P(w_1/x)P(w_2/x) = 2P(w_1/x)[1-P(w_1/x)]$$

Teniendo en cuenta el Error Condicional, -- Riesgo Condicional o Probabilidad Condicional de Error cuando se utiliza el Identificador Bayes de Error Mínimo, $r^*(x)$, o sea

- $x \in w_1$, si $P(w_1/x) \geq P(w_2/x)$

- $x \in w_2$, si $P(w_1/x) < P(w_2/x)$

$r^*(x) = \min \left[P(w_1/x), P(w_2/x)\right]$, entonces

$$r(x) = 2\left[P(w_1/x).(1-P(w_1/x)\right] =$$

$$2P(w_2/x)\left[1-P(w_2/x)\right] =$$

$$= 2r^*(x)\left[1-r^*(x)\right] = r(x)$$

o sea el Error Condicional de Identificación utilizando la Norma del 1-NN vecino más -- próximo es inferior a dos veces el óptimo - del Error Condicional, que es el Error Condicional utilizando el Identificador de Bayes, cuando la densidad condicional es conocida.

En lo que se refiere a la Probabilidad de - Error, ε, -a veces definido como R-, valor - esperado del Error Condicional $r(x)$,

$$\varepsilon = E\{r(x)\} = E\{2r^*(x)\left[1-r^*(x)\right]\} =$$

$$2\varepsilon^*(1-\varepsilon^*) - 2\mathrm{Var}\left[r^*(x)\right] = \varepsilon$$

o sea la Probabilidad de errór ε, al asignar individuos mediante la Norma del 1-NN, es menor que dos veces la Probabilidad de Error - utilizando el Identificador Bayes con densidad condicional conocida. Además se demuestra que $\varepsilon \geq \varepsilon^* = E\{r^*(x)\} = R^*$ por lo tanto

$$\varepsilon^* \leq \varepsilon \leq 2\varepsilon^*$$

o sea la Probabilidad de Error según el Identificador del 1-NN está comprendido entre -- una y dos veces la Probabilidad de Error mediante el Identificador de Bayes para Distribución de densidad condicional conocida.

Cuando se utiliza el Identificador del k-NN, y llamando

$r_k(x)$ = Error condicional de Clasificación - mediante el Identificador del k-NN,

$r^*(x)$ = Error condicional mediante Identificador de Bayes, se demuestra igual-- mente que

$$r_k(x) = P(w_i/x) \sum_{l=0}^{(k-1)/2} \binom{k}{l} (w_i/x)^l \left[1-P(w_i/x)\right]^{k-l} +$$

$$+ \left[1-P(w_i/x)\right] \sum_{l=(k+1)/2}^{k} \binom{k}{l} P(w_i/x)^l \cdot$$

$$\cdot \left[1-P(w_i/x)\right]^{k-l}$$

y teniendo en cuenta el Error Condicional mediante Identificador de Bayes

$$r_k(x) = r^*(x) \sum_{l=0}^{(k-1)/2} \binom{k}{l} \left[r^*(x)\right]^l \left[1-r^*(x)\right]^{k-l} +$$

$$+ \left[1-r^*(x)\right] \sum_{l=(k+1)/2}^{k} \binom{k}{l} \left[r^*(x)\right]^l \left[1-r^*(x)\right]^{k-l} =$$

$$= r_k(x)$$

Si se trata de expresar la relación entre Probabilidades de Error, ε_k, tomando las $E\{r_k(x)\}$ se obtiene, y demuestra que

$$\varepsilon^* \leq \varepsilon_k \leq 2\varepsilon^* \ (1-\varepsilon^*)$$

o sea que la Proabilidad de Error, o Error, según asignación por la Norma del k-NN está - comprendida entre el Error según Bayes y algo menos que el doble de la Probabilidad de Error o Error, según Bayes.

3.2.1.2. Estimación del Error de mala-Identificación según la Norma del k-NN

Según la relación entre Probabilidades dada - por el Teorema de Bayes,

$$\hat{P}(w_i/x) = \frac{\hat{P}(w_i)\ \hat{p}(x/w_i)}{\hat{P}(w_1)\hat{p}(x/w_1) + \hat{P}(w_2)\hat{p}(x/w_2)}$$

y de acuerdo con los valores obtenidos para - las estimaciones

$$\hat{P}(w_i) = \frac{N_i}{N_1+N_2}, \quad \hat{p}(x/w_i) = \frac{k_i-1}{N_i} \ \frac{1}{A(r,k,N)}$$

se obtiene

$$\hat{P}(w_i/x) = \frac{k_i-1}{k_1+k_2-2} = \frac{k'_i}{k'_1+k'_2}, \text{ si } k'_i = k_i - 1$$

y en definitiva

$$\hat{P}_{N_1}(w_2/x) = \frac{k'}{k'_1+k'_2}, \ \hat{P}_{N_2}(w_2/x) = \frac{k'_2}{k'_1+k'_2}$$

por lo tanto

$\hat{P}_{N_1}(w_1/x) \gtrless P_{N_2}(w_2/x)$ pasa a ser

$k'_1 \gtrless k'_2$ y los Errores condicionales

$$\sqrt{}_k(x) = P(w_1/x) \sum_{l=0}^{(k-1)/2} \binom{k}{l} [P(w_i/x)]^l .$$

$$. [1-P(w_i/x)]^{k-l} + [1-P(w_i/x)] .$$

$$. l=(\sum_{k+1)/2}^{k} \binom{k}{l} [P(w_i/x)]^l [1-P(w_i/x)]^{k-l}$$

$$= \sqrt{}_k(x),$$

$r^*(x) = \min P(w_1/x)$ cuando se utiliza Bayes

La relación entre ambos Errores condicionales puede escribirse,

$$\sqrt{}_k(r) = \sqrt{r^*(x)} \sum_{l=0}^{(k-1)/2} \binom{k}{l} [r^*(x)]^l [1-r^*(x)]^{k-l} +$$

$$+ [1-r^*(x)] \sum_{(k+1)/2}^{k} \binom{k}{l} . [r^*(x)]^l .$$

$$. [1-r^*(x)]^{k-l}$$

Fijada una X = X, o sea dado un individuo, - se tienen de la Muestra de Aprendizaje k_1 vecinos pertenecientes a la Clase w_1 y k_2 vecinos pertenecientes a w_2

$$\sqrt{}_k(x) = \min \{\hat{P}(w_1/X), \hat{P}(w_2/X)\} = \min$$

$$\{\frac{k_1}{k_1+k_2}, \frac{k_2}{k_1+k_2}\}$$

$$\sqrt{}_k(X) = \{0/k, 1/k, 2/k, \ldots {}^{k/2}/k,\} \text{ si k es par}$$

$$\sqrt{}_k(X) = \{0/k, 1/k, 2/k, \ldots {}^{(k-1)/2}/k\} \text{ si k es un par}$$

A partir de aquí, y teniendo en cuenta:

a) La probabilidad de que el Error Condicional $\sqrt{}_k(x)$ tome un valor i/k, significa que

$$\text{Min } [\hat{P}(w_1/X), \hat{P}(w_2/X)] = i/k$$

que puede expresarse como

$$\text{Prob}\{\sqrt{}_k(X) = \frac{i}{k}\} \text{Prob} \left\{ \begin{array}{l} \text{De que al contar los k-} \\ \text{Vecinos próximos a X,} \\ \text{hayan i pertenecientes} \\ \text{a } w_1 \end{array} \right\} +$$

$$+\text{Prob}\left\{ \begin{array}{l} \text{De que al contar los k-vecinos próximos} \\ \text{a X, hayan i pertenecientes a } w_2 \end{array} \right\}$$

b) Aplicando la Ley Binomial con

$$\frac{k_i}{k} = \hat{P}(w_i/X)$$

se obtiene $\forall k=2n$, (o sea k par)

$$\text{Prob}\{\sqrt{}_k(X) = \frac{i}{k}\} = \binom{k}{i} [P(w_1/X)]^i [\bar{P}(w_2/X)]^{k-l} +$$

$$+ \binom{k}{k-i} [\bar{P}(w_2/X)]^i [\bar{P}(w_2/X)]^{k-l}$$

para los valores de i comprendidos entre 0 y $\frac{k}{2}-1$

Cuando i = k/2

$$\text{Prob}\{\sqrt{(X)} = \frac{i}{k} = \frac{1}{2}, X=x\} = \binom{k}{k/2} .$$

$$. [\bar{P}(w_1/X)]^{k/2} [\bar{P}(w_2/X)]^{k/2}$$

si k es impar, k = 2n+1

$$\text{Prob}\{\sqrt{}_k(x)/X=x\} = \binom{k}{i} [\bar{P}(w_1/X)]^i [\bar{P}(w_2/X)]^{k-l} +$$

$$+ \binom{k}{i} [\bar{P}(w_1/X)]^{k-l} [\bar{P}(w_2/X)]^i$$

y la Esperanza Condicional de $\sqrt{}_K(x)$ es

$$r_K(X) = E\{\sqrt{}_K(X)/X=\} \Sigma \sqrt{}_K(X) . \text{Prob}\{\sqrt{}_K(x)/X=X\} =$$

$$= \sum_{i=1}^{\{k/2\}} \frac{1}{i} \binom{2i-2}{i-1} [\bar{P}(w_1/x)]^i [\bar{P}(w_2/x)]^i$$

siendo $\{k/2\} = \begin{cases} k/2 = \frac{2n}{2}, & \text{si } k = 2n \\ = \frac{2n+1-1}{2}, & \text{si } k = 2n+1 \end{cases}$

y al definir

$R_k = E\{r_k(x)\}$, Fukunaga obtiene

$$\hat{R}_k = \frac{1}{N_t} \sum_{l=1}^{N} r_k(X_i)$$

o sea dada una Muestra Test $X_1, X_2, \ldots, X_{Nt}$, - se utiliza la k-NN de cada X_i para calcular - $\min(k_1/k, k_2/k)$, y se establece el valor promedio para todas las X_i, obteniéndose una estimación de la Esperanza matemática del Riesgo Condicional cuando se utiliza el Identificador del k-NN.

Debe hacerse constar que la $\hat{R}_k$ es la estimación obtenida a partir de una Muestra de Test, definida como una parte del conjunto de la --

Muestra de Aprendizaje, sobre la cual se aplica el criterio de Identificación del k-NN.

Igualmente debe tenerse en cuenta que los métodos de estimación analizados plantean el Error Condicional de Clasificación, o su esperanza matemática, sin distinguir entre los Errores de Clasificación dentro de la Clase w_1 y w_2, $\bar{r}_1$ y $\bar{r}_2$, que son en definitiva los interesantes para el Control Estadístico de Calidad que es el centro de interés de este trabajo.

En cuanto a la Estimación de Errores de Clasificación debe señalarse la actuación de Fukunaga y Hostletler, quienes han trabajado sobre la estimación del R de Bayes en Clasificación mediante Muestras Test no-clasificadas previamente, utilizando lo que llaman Estimación Agrupada (Grouped-Estimates) y Estimación Conjunta (Pooled Estimates). Ni por las características de la Muestra Test, ni por variable estimada (R), son de aplicación al tema del Control Estadístico de Calidad.

4. PROPUESTA DE PLAN DE MUESTREO EN LA RECEPCION DE INDIVIDUOS ESTADISTICOS MULTIDIMENSIONALES

4.0 Generalidades

Planteada la idea del C.E.C. en la Recepción mediante variables multidimensionales, se ha visto la necesidad de resolver con caracter previo el Problema de Identificación, explicitando claramente las dos Clases en que se ha peticionado el conjunto, definiendo el Identificador, y calculando los Errores de Clasificación medios.

En definitiva a partir de la Muestra de Aprendizaje Supervisada con Profesor, se integra toda la información existente sobre los individuos, resolviéndose el Problema de Identificación. Llegados aquí puede plantearse el C.E.C. en la Recepción por Atributos, distinguiéndose del unidimensional únicamente en la Fase de Identificación.

Se comprende fácilmente que la Puesta en Marcha de C.E.C. tenga cinco fases completamente diferenciales:

- Fase de Aprendizaje

Tiene por objeto fijar el tamaño de la Muestra de Aprendizaje, y su extracción, actuando el Supervisor (Profesor) que definirá los N_1 individuos de la MA pertenecientes a w_1 y los N_2 pertenecientes a w_2. Por lo tanto se tendrá registrados:

- Los N_1 individuos de la MA pertenecientes a w_1
- Los N_2 individuos de la MA pertenecientes a w_2
- Los p (ó k) variables o componentes de cada uno de los $N_A = N_1 + N_2$ individuos de la MA.

- Fase de Identificación (Cálculos)

A partir de las características de la Población (Lote) de donde va a extraerse la Muestra de Inspección, y teniendo en cuenta las características de la Muestra de Aprendizaje, se definirá la Norma de Clasificación o Identificador, y se estimarán los Errores de Clasificación medios.

- Fase de Diseño del Plan C.E.C. en la Recepción

Teniendo en cuenta la Teoría Estadística del Control de Recepción por Atributos, y lo calculado en la Fase de Identificación (Cálculos), se diseña el Plan de Muestreo,

- El tamaño de la Muestra de Inspección (n-individuos p-dimensionales) a extraer del Lote integrado por N-individuos.

- La definición del estadístico del test, que mediante la aplicación del Identificador permita la Aceptación o Rechazo del Lote recepcionado, y de acuerdo con las características de los n-individuos extraidos como M.I. de Lote.

- Fase de Comprobación del Plan de Muestreo.

Diseñado el Plan, esta fase de caracter experimental previo, tiene como objetivo evaluar la eficacia del Plan de Muestreo comparándolo con los resultados obtenidos con una Inspección al 100%, -mediante utilización del Profesor-.

Si de la comparación resultasen diferencias

estadísticamente significativas, se reajustará el Plan, revisando en primer lugar el tamaño de la Muestra de Aprendizaje y el Ajuste de la Fase de Cálculos de Identificación; en el momento que las diferencias estadísticas no tuviesen significación, se pasaría, ya, a la Fase de Trabajo u Operativa.

- Fase Operativa del C.E.C. en la Recepción

Verificada la eficacia del Plan de Muestreo Multidimensional, ya puede operarse con el Plan diseñado, de manera que la Aceptación o Rechace del Lote sólo responda a los resultados del Plan, -donde se utilizará para identificar los individuos de la MI, el Identificador obtenido de la MA, sin exigir la supervisión del Profesor en el Control propiamente dicho.

4.1 Control de Calidad Multidimensional en la Recepción mediante identificación de cada uno de los individuos de la M. Inspección, utilizando un Identificador del tipo Función Discriminante

4.1.1. Plan de Muestreo

- Se extraen n-individuos p-dimensionales del Lote de N-individuos a recepcionar.

- Se clasifican los n-individuos, aplicando el Identificador de la Función Discriminante muestral, o sea

$$x \in w_i, \text{ si } D'_s(x) \geq 0$$
$$x \notin w_i. \text{ si } D'_s(x) < 0 \qquad \forall i = 1,2$$

siendo

$$D'_s(x) = \left[\bar{x} - \frac{1}{2}(\bar{x}_1+\bar{x}_2)\right]^t S^{-1}(\bar{x}_1-\bar{x}_2), \text{ donde}$$

x = individuo de la M. Inspección que ha de identificarse

$\bar{x}_i$ = media de la clase w_i, a partir de los N_i individuos supervisados mediante Profesor de la Muestra de Aprendizaje.

S = matriz de covariancia común de las dos clases w_i.

- Se determina la proporción de individuos rechazables o clasificados como defectuosos en la M. Inspección al aplicar el Identificador, $\hat{P}$.

$$\hat{P} = \frac{\text{Card}\{x_j;\ j=1,2,\ldots n; x_j \notin \text{Clase Aceptable}\}}{n}$$

- Se acepta el Lote, si $\hat{P} \leq P_c$
  Se rechaza el Lote, si $\hat{P} > P_c$

4.1.2. Cálculo de los valores que intervienen en el Plan de Muestreo

4.1.2.1. Riesgos de 1ª y 2ª Especie de Error de Clasificación media mediante Función Discriminante

Se trata de establecer la relación entre los Riesgos de Proveedor y de Cliente -existentes en todo C.E.C. en la Recepción-, con el Error Medio de Clasificación de cada uno de los individuos que componen la M. de Inspección.

Definido:

$$P_i = \text{Prob}\{(w_i/w_j);\ i, j=0,1, i \neq j\} =$$

= Probabilidad{de clasificar un x en w_i cuando realmente pertenece a w_j, $i \neq j$.}

que de acuerdo con las características de $D_s(x)$, y cuando N_i, $i=1,2$ eran
te grandes, puede enunciarse como

$$P_i = \phi\left\{(-1)^i D_s(\mu_i)/\sqrt{V_{D_S}}\right\} = \phi\left\{(-1)^i \cdot \frac{\left[\mu_i - \frac{1}{2}(\bar{x}_1-\bar{x}_2)^t S^{-1}(x_1-x_2)\right]}{\sqrt{(\bar{x}_i-\bar{x}_2)^t S^{-1}\Sigma_i S^{-1}(\bar{x}_1-\bar{x}_2)}}\right\}$$

siendo:

μ_i = Media de la Clase w_i
Σ_i = Matriz de Covariancia de la Clase w_1
$\bar{x}_i$ = Valor estimado, a partir de los N_i individuos de la M.A., de la Media de la Clase w_i.
S = Valor estimado de la Matriz de Covariancia

común definido por

$$S = \frac{\sum_{i=1}^{2} N_i \hat{\Sigma}_i}{\sum_{i=1}^{2} N_i}$$

siendo $\hat{\Sigma}_i$: valor estimado de la Matriz de Covariancia de cada Clase w_i

Como los valores μ_i, Σ_i, de cada clase son desconocidos, ha de utilizarse un estimador de P_i, $\hat{P}_i$. Ya se ha definido anteriormente el Estimador $\hat{P}_i$ de Mc Lachlan que minimiza el Error Cuadrático medio cuando Σ es desconocido,

$$\hat{P}_1 = Q_M = \phi(-D/2) + \hat{A} - \hat{B} - \hat{C}$$

Teniendo en cuenta la definición los riesgos α, β y la Hipótesis de Normalidad de $p(X/w_i)$, -utilizada en la propia definición de la Función Discriminante-, se tendría que

$$(P_1)^N \leq {}^{\alpha}\text{Lote} \qquad (P_2)^N \leq {}^{\beta}\text{Lote}$$

por lo tanto la relación entre Errores Medios de Clasificación y Riesgos del Comprador y del Vendedor, no dependen del tamaño de la M. Inspección, mientras que si dependen del Tamaño del Lote, y de la Muestra de Aprendizaje, -a través de los Errores de Clasificación-.

4.1.2.2. Cálculo de n y P_c

Al plantear un C.E.C. multidimensional mediante la Identificación individual de los componentes de la Muestra de Inspección, existe la posibilidad de Error en la Clasificación, que actuará sobre la eficacia del Control, debido a que éste se efectua según los resultados de la Clasificación sobre la M. de Inspección.

Esta posibilidad de Error se ha estimado mediante $\hat{P}_1$ y $\hat{P}_2$ y por lo tanto

Prob $(H_1/P) \cong C(n,P,N).(1-\hat{P}_2-\hat{P}_1)^n$ y aplicando

$$C_1(n,P_c,N).(1-\hat{P}_1-\hat{P}_2) = \left[\sum_{1=0}^{c} C_n^1\, p_1^1\, (1-p_1)^{n-1}\right].$$

$$.\ (1-\hat{P}_o-\hat{P}_1) \geq (1-\alpha)$$

$$C_2(n,P_c,N)(1-\hat{P}_1-\hat{P}_2)^n = \left[\sum_{1=0}^{c} C_n^1\, p_1^1\, (1-p_2)^{n-1}\right].$$

$$.\ (1-\hat{P}_o-\hat{P}) < \beta$$

de donde se obtiene n y c, definiéndose

$P_c = \frac{c}{n}$, donde c= nº de defectuosos en la M. Inspección

n= tamaño M. Inspección

Si se desea trabajar al máximo nivel de exactitud posible debe tenerse en cuenta que la Estimación de la Proporción de individuos P, obtenida a partir de la utilización del identificador de la F. Discriminante, resulta de dos circunstancias:

- Clasificación, mediante la $D_s(x)$, de individuos realmente defectuosos como defectuosos.
- Clasificación, mediante la $D_s(x)$, de individuos realmente no-defectuosos como individuos defectuosos,

por lo tanto, el número de defectuosos observados $\hat{P}$, será:

- Rechazables por definición, que son rechazados por el Identificador

  $\bar{P} = (1-P_2)^n$

- Aceptables en realidad que son rechazados por el Identificador

  $(1-P)(1-P_1)^n$, y además

  $\hat{P} = \bar{P}\,(1-P_2)^n + (1-\bar{P})(P_1)^n$;

$\bar{P} = \dfrac{\hat{P} + (P_1)^n}{(1-P_2)^n - (P_1)^n}$, por lo tanto en el caso de la máxima exactitud - no se compararía $\hat{P}$ con P_c, sino el valor $\bar{P}$ - con P_c de manera que,

Si $\bar{P} \leq P_c$, se acepta el Lote
Si $\bar{P} > P_c$, se rechaza el Lote

4.2. Control de Calidad en la Recepción mediante identificación de cada uno de los Individuos de la Muestra de Inspección, utilizando un Identificador del k-NN (o k-Vecino más próximo)

4.2.1. Plan de Muestreo

- Se extraen n-individuos p-dimensionales del Lote de individuos a recepcionar.
- Se clasifican los n-individuos, aplicando el Identificador del k-vecino más próximo (k-NN), o sea

$x \in w_i$, si $k_{i-1} > k_{i+1}^{-1}$

- Se determina la proporción de defectuosos - P en la Muestra,

$$\hat{P} = \{\text{Card}[x_i,\ i=1,2,\ldots n;\ \hat{p}(x_i/w_0) < \hat{p}(x_i/w_1)]\}/n$$

- Se acepta el lote si $\hat{P} \leq P_c$, y

- Se rechaza el lote si $\hat{P} > P_c$

4.2.2. Cálculo de los valores que intervienen en el Plan de Muestreo

4.2.2.1. Riesgos de 1ª y 2ª Especie y Error de Clasificación Medio

Tal como se ha indicado, existirán unas relaciones entre los Riesgos del Comprador y Vendedor (Riesgos de 1ª y 2ª Especie en el Plan de Control de Calidad en la Recepción), y con la Estimación del Error Condicional de Clasificación (Identificación) mediante la Técnica del k-NN.

Sea

$$E_x\{\Gamma_k(x)\} = E_X\{\Gamma^*(x) \sum_{l=0}^{(k-1)/2} \binom{k}{l} [r^*(x)]^l \cdot [1-r^*(x)]^{k-l} + [1-r^*(x)] \sum_{l=(k+1)/2}^{k} \binom{k}{l} [r^*(x)]^l [1-r^*(x)]^{k-l}$$

que puede escribirse, teniendo en cuenta la - concavidad de la función como,

$$E_x\{r_k(x)\} \leq \sum_{l=0}^{(k-1)/2} \binom{k}{l} \{E_x [r^*(x)]^{l+1} \cdot [1-E_x(r^*(x)]^{k-l}\} + \sum_{l=(k+1)/2}^{k} \binom{k}{l} \cdot \{E_x [r^*(x)]^l [1-E(r^*(x)]^{k-l+1} = \bar{\Gamma}_1 + \bar{\Gamma}_2$$

y por lo tanto, dado que $E\{r^*(x)\}$ no es conocida, -por no serlo las leyes de Distribución de las densidades dentro de las clases - w_i, y en consecuencia no estar determinado el $r^*(x)$ de Bayes, ni su Esperanza Matemática-, habrá dificultad en estimar y conocer $\bar{\Gamma}_1$ y - $\bar{\Gamma}_2$.

Si como suele ocurrir en los procesos de fabricación semiautomáticos con intervención - humana, puede suponerse que las variables multidimensionales de los Individuos del Lote - sigan una Ley Normal multidimensional, entonces

$$[E_x(\widehat{r^*(x)})] = \phi(-D/2)$$

siendo

$$D^2 = (\bar{x}_1-\bar{x}_2)^t \ S^{-1} (\bar{x}_1-\bar{x}_2)$$

$$x_1 = \frac{\sum_{m_1}^{N_1} x_1^m}{N_1} \qquad x_2 = \frac{\sum_{m_1}^{N_2} x_2^m}{N_2}$$

$$S = \frac{\sum_{m=1}^{N_1} (x_1^m-\bar{x}_1)^t (x_1^m-\bar{x}_1) + \sum_{m=1}^{N_2} (x_2^m-x_2)^t (x_2^m-\bar{x}_2)}{N_1+N_2-2}$$

con lo que

$$E_x (r^*_k(x)) \leq \bar{\Gamma}_1+\bar{\Gamma}_2 =$$

$$= \sum_{l=0}^{(k-1)/2} \binom{k}{l} [G(-D/2)]^{l+1} [1-G(-D/2)]^{k-l} + \sum_{l=(k+1)/2}^{k} \binom{k}{l} [G(-D/2)]^{l+1} [1-G(-D/2)]^{k-\ell+1}$$

y teniendo en cuanta lo expresado en 4.1.2.1,

$$(\bar{\Gamma}_1)^N \leq \alpha_{Lote},\ (\bar{\Gamma}_2)^N \leq \beta_{Lote}$$

4.2.2.2. Cálculo de n y P_c

Estimadas las posibilidades de Error mediante el Error de Clasificación medio $E_x (r^*(x)) \leq \bar{\Gamma}_1+\bar{\Gamma}_2$, y calculados $\bar{\Gamma}_1$ y $\bar{\Gamma}_2$, se cumplirá

$$\text{Prob}\ (H_1/P) \cong C(n,P,N) \cdot (1-\bar{\Gamma}_1-\bar{\Gamma}_2)^n \quad \text{y}$$

teniendo en cuenta que

$$C_1(n,P_c,N) \cdot (1-\bar{r}_1-\bar{r}_2)^n = [\sum_{i=0}^{I} C_n^i\ p_1^i\ (1-p_1)^{n-i}] \cdot (1-\bar{\Gamma}_1-\bar{\Gamma}_2)^n \geq (1-\alpha)$$

$$C_2(n,P_c,N)\ (1-\bar{r}_1-\bar{r}_2)^n = [\sum C_n^I\ p_2^l\ (1-p_2)^{n-\ell}]\ (1-\bar{r}_1-\bar{r}_2)^n \leq \beta$$

se obtendrá n = número de Individuos de la -- M.I.

c = número de Individuos defectuosos en la M.I.

con lo que quedaría definido el Plan de Muestreo.

Si al igual que en la 4.1.2.2. interese plantear el Control con la máxima exactitud deberá plantearse como allí, la relación entre -- defectuosos observados y defectuosos reales,

$$\bar{P} = \frac{\hat{P} - (r_1)^n}{(1-r_2)^n - (r_1)^n}$$

siendo $\bar{P}$ = defectuosos reales (estimación de la Proporción real)

$\hat{P}$ = defectuosos observados

4.2. Propuesta de Programa de Control de Calidad en la Recepción, mediante variable multidimensional, y con Muestra de Aprendizaje

Teniendo en cuenta lo expuesto en el desarrollo de este trabajo, se trata de plantear un Plan de Control de Recepción en el campo de - Variables Multidimensionales, utilizando una Muestra de Aprendizaje supervisada con Profesor, en el que se aplique un concepto equivalente al Control Cualitativo o por Atributos de la Recepción en el caso de Variable Unidimensional.

Se justifica la necesidad de definir un Identificador calculado como Clasificador que particione el conjunto de Individuos de la M.A. en sus Clases, que permitirá identificar (Asignar) cada elemento de la Muestra de Inspección, extraída del Lote a recepcionar, en la Clase Buena o Defectuosa. Según la proporción de Individuos de la M.I. que se identifiquen en la Clase Buena o Defectuosa, y su comparación con la obtenida en el Diseño del -- Plan, se Acatará o Rechazará el Lote recepcionado.

Según lo anterior se considera oportuno subdividir el Programa de Control de Calidad en la Recepción en dos sub-Programas. En el primero se obtienen los parámetros de la función u Operador de Clasificación a partir de la -- Muestra de Aprendizaje, se define el Operador de Clasificación, y se estima el Error de Mala Identificación. En el segundo sub-Programa, y a partir de lo obtenido en el primero, y de los Riesgos de 1^a y 2^a especie (o del -- Comprador y del Vendedor), se calculan los parámetros del Plan de Control de Calidad (Diseño del Plan) realizándose posteriormente la identificación de la Clase a la cual pertenece cada individuo de la M.I. extraída del Lote recepcionado, efectuándose el Contado de los pertenecientes a cada Clase, para mediante comparación con el parámetro calculado en el Diseño del Plan, Aceptar o Rechazar el Lote recepcionado.

El primer Sub-Programa, LCM-PCC1, obtiene, -- a partir de los Individuos de la Clase w_1 y de la w_2 en que se ha dividido la M.A.-, los valores estimados de la Media y de la Desviación tipo de cada Clase.

Se efectua entonces la comprobación de la igualdad de ambas matrices de Covariancia, ---mediante técnicas del Análisis de Covariancia que aquí no se explicitan-, y que de no cumplirse interrumpe la realización del Programa, escribiendo "NO EXISTE MATRIZ UNICA - DE COVARIANZA".

Verificada la posiblidad de considerar una sola Matriz de Covariancia para los individuos de las dos Clases, se calculan los valores - constantes del Clasificador (aquí la Función Discriminante), y la Estimación de la Distancia de Malhanobis.

Posteriormente y mediante una Subrutina de - Cálculo ESP1, totalmente explicitada, y otra de carácter semejante ESP, se estiman -utilizando el Estimador de Okamoto-, los valores del Error de Mala-Identificación.

Se obtiene así, en el caso de una sola Matriz de Covariancia, y escrito en el listado del - Programa, el Tamaño de la Muestra de Aprendizaje, el tamaño de la dimensión de cada Individuo, la Parte constante de la Función Discriminante, la Distancia de Malhanobis y los valores Estimados de los Errores de Mala Identificación, al mismo tiempo que queda constancia de la Media Estimada de la Clase w_1 y de la w_2, así como la Matriz de Covariancia común a ambas Clases

El segundo Sub-Programa, LCM-PP2, al que se - le ha concedido entidad propia, -ya que podría utilizar un microprocesador situado en - la zona geográfica de Recepción de Lotes-, -- realizaría el cálculo del Plan de Muestreo, - la extracción automática (y aleatoria de individuos del Lote recepcionado, hasta alcanzar

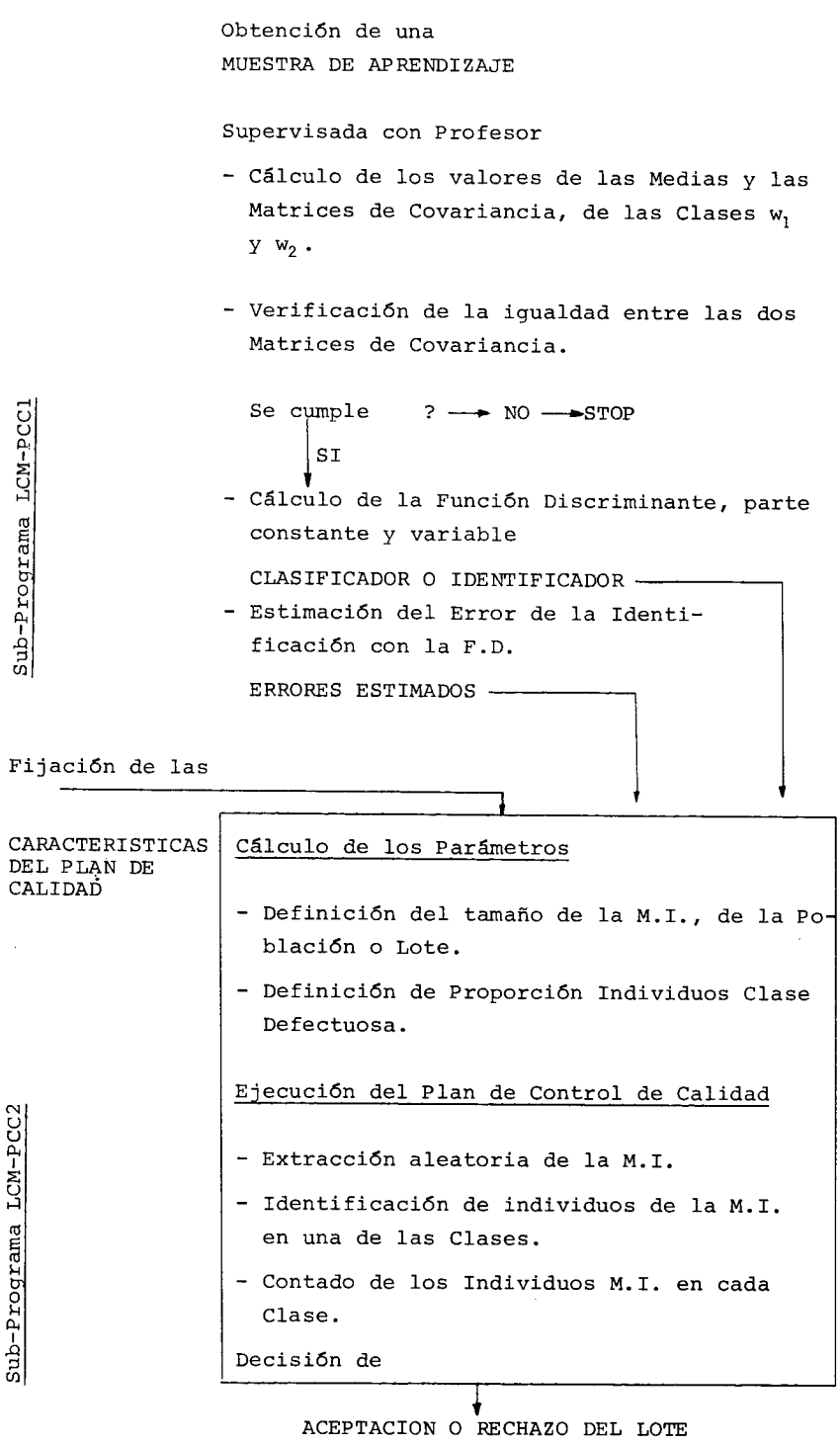

Fig.1.- ESQUEMA DEL PROGRAMA DE CONTROL DE CALIDAD MULTIDIMENSIONAL EN LA RECEPCION DE LOTES

el tamaño de la Muestra de Inspección exigido por el Plan) y la obtención (mediante sensores o captores) de las distintas magnitudes integrantes de la Variable Multidimensional, Individuo. Posteriormente realizarían la Asignación de cada uno de los individuos de la M.I. a las Clases w_1 y w_2, realizando un Proceso de contado que finalizaría con la Aceptación o Rechace del Lote.

En este Sub-Programa se utilizan como Datos de Entrada, las Estimaciones de los Errores de Mala-Identificación, -efectuadas en el -- LCM-PCC1-, denominadas aquí DP0 y DP1, los -- Riesgos de 1^a y 2^a Especie ALF y BET, así como la dimensión de los Individuos, T.

Como resultado del Sub-Programa y dentro del Diseño del Plan de Muestreo de Calidad, se ob

tienen TA (tamaño de la Muestra de Inspec-- ción), ZS (proporción de Defectuosas que como máximo pueden haber en la M.I.) DEFT (número máximo de Individuos pertenecientes a la Clase Mala o Defectuosa, que pueden haber en la M.I. cuando se acepta el Lote).

Definidos estos valores, el Microprocesador ordenaría la extracción de los T.A. individuos (componentes de la M.I.) obtendría mediante los erroes o captores las magnitudes componentes de cada Individuo, los identificaría dentro de cada Clase w_1 o w_2, contará las pertenecientes a cada Clase, según el número de la Clase Mala o Defectuosa y su comparación con DEFT, Aceptaría o Rechazaría el Lote, escribiendo la decisión adoptada sobre el Lote recepcionado.

5. BIBLIOGRAFIA

/1/ ESCUDERO, L.F.: "Reconocimiento de Patrones", Paraninfo 77 - Madrid.

/2/ FUKUNAGA,K.: "Introduction to statistical Pattern Reconition", Acedemic Press Inc. 1972, New York.

/3/ LACHENBRUCH : "Discriminant Analysis", - Hafner Press New York, 1975.

/4/ PAU, L.F. : "Diagnostic des Pannes dans les systemes. Approche par le Reconnaissance des Formes", Cepadues Editions - - Toulouse 1976.

/5/ FUKUNAGA & KESELL :"Estimación del Error de Clasificación", IEE Transactions on - Computer - December 1971.

/6/ FUKUNAGA & KESSELL : "Estimación no paramétrica del Error de Bayes, usando muestras no-clasificadas", IEE Trans. of Inf. Theory July, 1973.

/7/ LACHENBRUCH : "Estimaciones de las Proporciones de Error en el Análisis Discrimi-- nante", Tecnométrica, Febrero 1968.

/8/ LARIO, Francisco-Cruz: "Métodos Estadísticos en Control de Calidad Multidimensional por Medidas y en el Diagnóstico de - Averias Estimación y Análisis de Datos" - (Tesis Doctoral). No Publicado.

/9/ MC LACHLAN,: "Una técnica asimptótica-- mente insesgada para estimar las proporciones de Error en el Análisis Discriminante", Biometrics June, 1974.

/10/ OKAMOTO,: "Una expresión asimptótica de la distribución de la Función Discrimi-- nante Lineal"

/11/ SORM, "Estimación de la Probabilidad - Condicional de Mala Clasificación", Tecnometrics, May 1971.