5. Análisis de cúmulos

- OBJETIVO: Dividir a los individuos de una base de datos en grupos, llamados cúmulos (*clusters*), de tal manera que los individuos de un mismo cúmulo tengan características semejantes con respecto a las variables medidas.

5.1. Medidas de similaridad y disimilaridad.

- Para hacer un análisis de cúmulos es necesario medir de alguna manera la similaridad o disimilaridad entre dos observaciones multivariadas.

- TIPOS DE DISTANCIAS: Existen varias formas de medir la similaridad o disimilaridad entre observaciones. Las distancias más comunes son 3. Sean x_i y x_j dos observaciones multivariadas.

1) *Distancia métrica o euclidiana*. Es la norma del vector de diferencias de las dos observaciones,

$$d_{ij} = \left\{ (x_i - x_j)'(x_i - x_j) \right\}^{1/2} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2} \, .$$

2) *Distancia métrica o euclidiana estandarizada*. Es la norma del vector de diferencias de las dos observaciones estandarizadas,

$$d_{ij} = \left\{ (z_i - z_j)'(z_i - z_j) \right\}^{1/2},$$

donde z_i y z_j son las observaciones estandarizadas. Esta distancia es la más usada.

3) *Distancia de Mahalanobis*. Es una distancia euclidiana ponderada por la matriz de varianzas y covarianzas,

$$d_{ij} = \left\{ \left(x_i - x_j \right)' \Sigma^{-1} \left(x_i - x_j \right) \right\}^{1/2}.$$

- *Nota*: Si las características de un individuo no se pueden representar mediante variables (continuas), es posible mediar las similaridades entre individuos mediante la presencia o ausencia de cierta característica (variables binarias).

5.2. Métodos gráficos útiles.

- Existen varios algoritmos para formar cúmulos. De hecho, algoritmos diferentes pueden producir distintas agrupaciones. Más aún, el análisis de cúmulos puede detectar grupos que no existan en la realidad.

- Una forma de evaluar los resultados de los métodos de agrupación, es mediante métodos gráficos.

- DIAGRAMAS DE DISPERSIÓN. Cuando se tienen únicamente dos variables de interés ($p = 2$), un diagrama de dispersión entre ellas permitiría visualizar posibles agrupaciones entre los individuos.

- GRÁFICAS DE COMPONENTES PRINCIPALES. Cuando el número de variables de interés es mayor a dos ($p > 2$), se puede implementar un análisis de componentes principales. Si la proporción de la variabilidad explicada por las dos primeras componentes es significativa ($\approx 80\%$) se puede realizar un diagrama de dispersión de los marcadores de las dos primeras componentes y visualizar la posible existencia de cúmulos.

- NOTA: Si el número de variables es "grande" ($p > 10$), es más recomendable realizar primero un análisis de componentes principales y posteriormente aplicar las técnicas de análisis de cúmulos a las primeras *r* componentes, que aplicar directamente un análisis de cúmulos a las variables originales.

- ¡PRECAUCIÓN!: Los marcadores de las componentes principales nunca deben estandarizarse. Esto es porque los marcadores estandarizados no reflejan de manera realista las distancias entre individuos.

➢ DIAGRAMA DE ANDREWS. Este tipo de gráficas, aplicadas sobre las variables originales, son muy útiles para identificar cúmulos y para validar los resultados de un análisis de cúmulos. Los individuos en el mismo cúmulo deben de tener gráficas de Andrews similares.

➢ OTROS TIPOS DE DIAGRAMAS. Otros métodos gráficos como los diagramas de dispersión tridimensionales, diagrama de burbujas, las caras de Chernoff y los diagramas de estrellas son útiles para validar un análisis de cúmulos. Sin embargo, las caras de Chernoff y los diagramas de estrellas pierden sencillez e interpretación cuando el número de variables aumenta.

5.3. Métodos para realizar análisis de cúmulos.

➢ TIPOS DE MÉTODOS. Existen dos tipos de métodos para realizar un análisis de cúmulos: Métodos jerárquicos y métodos no jerárquicos.

- MÉTODOS JERÁRQUICOS: Este tipo de métodos consiste en una serie de uniones o una serie de divisiones sucesivas. Los resultados de estos métodos se muestran en un diagrama bidimensional llamado dendograma (*dendogram*).

  1) *Métodos de uniones*. Se inicia tomando a cada individuo como un cúmulo, los cúmulos (individuos) más similares se agrupan entre sí y así sucesivamente hasta que la disimilaridad entre distintos cúmulos va decreciendo. Eventualmente, todos los individuos quedan agrupados en un solo cúmulo. Los métodos de aglomeración más comunes son:
     a. Método del vecino más cercano (liga sencilla).
     b. Método del vecino más lejano (liga completa).
     c. Método de la distancia promedio (liga promedio).
     d. Método de la varianza mínima de Ward.

  2) *Métodos de divisiones*. Estos métodos trabajas en sentido opuesto a los anteriores. Se inicia tomando a todos los individuos en un solo cúmulo. Este cúmulo único se divide en dos subcúmulos de tal manera que los individuos en uno de los subcúmulos se encuentran lejos de los individuos en el otro subcúmulo. El proceso se continua hasta que hay el el mismo número de cúmulos que individuos.

- MÉTODOS NO JERÁRQUICOS: Este tipo de métodos consiste producir un número fijo de cúmulos, digamos K. El número K puede estar preestablecido o puede ser obtenido como parte del proceso. Este tipo de métodos puede iniciar con una partición inicial de individuos en cúmulos o

una selección inicial de puntos semilla que van a formar el centroide de los cúmulos. El método más común es:

a. Método de K-medias.

- NOTA: Los métodos no jerárquicos requieren de menos trabajo computacional, por lo que pueden aplicarse a bases de datos más grandes que los métodos jerárquicos.

➢ ALGORITMO GENERAL PARA EL MÉTODO JERÁRQUICO DE UNIONES.

Supongamos que el número total de individuos a agrupar es *n*.

1. Empieza con *n* cúmulos, cada uno conteniendo a un solo individuo.
2. Calcula la distancia entre cada uno de los cúmulos y determina los cúmulos con distancia mínima, digamos U y V (cuya distancia se denota como d_{UV}).
3. Une los cúmulos U y V y nombra al nuevo cúmulo (UV). Calcula de nuevo las distancias entre este nuevo cúmulo y los demás cúmulos.
4. Repite los pasos 2 y 3 un total de *n-1* veces, i.e. hasta que todos los individuos pertenezcan al mismo cúmulo. Registra los cúmulos que se van uniendo y las distancias a las que la unión ocurre.

➢ DEFINICIÓN DE LAS DISTANCIAS entre cúmulos:

- Para el *Paso 2*: $D = \{d_{ij}\}$, donde d_{ij} es cualquiera de las distancia definidas en la Sección 5.1.

- Para el *Paso 3*: Supongamos que los cúmulos (individuos) con menor distancia fueron U y V y se unieron para formar el cúmulo (UV). La distancia entre el nuevo cúmulo (UV) y otro cúmulo W es:
  a. *Método del vecino más cercano*: La distancia entre cúmulos se define como la distancia entre dos elementos (uno de cada cúmulo) que están más cercanos, i.e.,
  $$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}.$$
  b. *Método del vecino más lejano*: La distancia entre cúmulos se define como la distancia entre dos elementos (uno de cada cúmulo) que están más lejanos, i.e.,
  $$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}.$$
  Este método asegura que todos los elementos de un cúmulo están dentro de una distancia máxima uno del otro.
  c. *Método de la distancia promedio*. La distancia entre cúmulos se define como el promedio de todas las distancias entre dos elementos (uno de cada cúmulo), i.e.,
  $$d_{(UV)W} = \frac{\sum_i \sum_j d_{ij}}{n_{(UV)} n_W},$$
  donde d_{ij} es la distancia entre el elemento *i* del cúmulo (UV) y el elemento *j* del cúmulo W.

- *Método de la varianza mínima de Ward*. $D = \{d_{ij}\}$, donde d_{ij} es una distancia medida en términos de la varianza muestral de la unión de los cúmulos *i* y *j*. Es decir,

$$d_{ij} = \varpi^2_{(ij)} = \frac{1}{n_i + n_j} \sum_{k=1}^{n_i+n_j} \left(x_k - \bar{x}_{(ij)}\right)^2$$

Inicialmente, cada cúmulo consiste de una sola observación, por lo tanto $d_{ij} = \varpi^2_{(ij)} = 0$. Igual que en los tres métodos anteriores, los cúmulos U y V se unen si su distancia (varianza muestral de la unión) es la más pequeña de todas.

- MÉTODO DE K-MEDIAS (NO JERÁRQUICO). Este algoritmo asigna cada individuo al cúmulo que tenga el centroide más cercano. En general, el algoritmo se puede representar por los siguientes pasos:
  1. Particionar al conjunto de individuos en K cúmulos iniciales y calcula el centroide (media) de cada cúmulo.
  2. Calcula la distancia (euclidiana) de cada individuo a cada uno de los K centroides. Reasigna cada individuo al cúmulo cuya distancia al centroide sea la menor.
  3. Repite el Paso 2 hasta que ningún individuo sea reasignado a un cúmulo nuevo.

- COMETARIOS FINALES:

- El número de cúmulos óptimo se determina visualizando el dendograma y determinando una distancia para la cual los grupos están bien diferenciados.
- El método del vecino más cercano tiende a maximizar la distancia entre los cúmulos, produciendo un menor número de cúmulos que los demás métodos. En cambio, el método del vecino más lejano tiende a minimizar

las distancias dentro de cada cúmulo, por lo que produce un número más grande de cúmulos que los demás métodos.

- El método de K-medias es muy criticado porque fija de antemano el número K de cúmulos.
- La agrupación perfecta no es tan sencilla de obtener, por lo que es recomendable intentar con más de un método. Si varios métodos dan resultados semejantes, entonces se puede suponer que existe una agrupación natural de los individuos.
- Es importante realizar una evaluación gráfica de los métodos de análisis de cúmulos.
- *Nota*. Los métodos jerárquicos se pueden usar para formar cúmulos de variables, usando como medida de distancia el valor absoluto de la correlación muestral entre ellas.

❖ Splus: hclust, kmeans.

6. Escalamiento multidimensional

- DEFINICIÓN: El escalamiento multidimensional es una técnica que permite "mapear" (convertir, copiar) en un espacio de menos dimensiones las distancias originales entre individuos que se encuentran en un espacio de muchas dimensiones.

- UTILIDAD: Resulta de mucha utilidad mapear distancias de un espacio de muchas dimensiones a un espacio de dimensión 2, ya que en este caso los individuos se pueden representar en una gráfica de dos dimensiones y se puede apreciar visualmente la cercanía a lejanía entre ellos.

- En general, la idea del escalamiento multidimensional es representar las distancias entre individuos que originalmente se encuentran en un espacio p-dimensional a un espacio q-dimensional, donde $q < p$. Por sencillez, se acostumbra usar $q = 2$.

- La técnica de escalamiento multidimensional se puede ver como una técnica que nos permite hacer un análisis de cúmulos gráficamente.

- EXPLICACIÓN DEL ALGORITMO BÁSICO:

  - Calcular las distancias reales (D_{ij}) en el espacio p-dimensional entre los individuos *i* y *j*. La forma usual de calcular la distancia es mediante la distancia euclidiana estandarizada, i.e.,

$$D_{ij} = \left\{ (z_i - z_j)' (z_i - z_j) \right\}^{1/2},$$

para $i \neq j = 1,2,\ldots,n$.

¿Cuántas distancias hay que calcular?. $m = \binom{n}{2} = \frac{n(n-1)}{2}$.

- Ordenar las distancias en orden ascendente

$$D_{i_1 j_1} < D_{i_2 j_2} < \cdots < D_{i_m j_m}$$

donde, $D_{i_1 j_1}$ es la distancia entre los dos puntos más cercanos, $D_{i_2 j_2}$ la distancia entre los siguientes dos puntos más cercanos y finalmente, $D_{i_m j_m}$ la distancia entre los dos puntos más lejanos.

- La idea es encontrar un conjunto de m puntos en un espacio q-dimensional, cuyas distancias $\{d_{ij}\}$ preserven el orden de las distancias en el espacio original, i.e.,

$$d_{i_1 j_1} < d_{i_2 j_2} < \cdots < d_{i_m j_m} \qquad (6.1)$$

Nota: Lo más importante es el orden entre las nuevas distancias, no las magnitudes de las distancias.

- La forma de obtener la nueva configuración de los puntos en un espacio de *q* dimensiones es mediante un proceso iterativo.

1) Determina una configuración inicial de puntos en q dimensiones. Calcula las distancias entre puntos $d_{ij}^{(q)}$ y encuentra las cantidades $\hat{d}_{ij}^{(q)}$ que satisfacen la condición (6.1) y minimizan la función de Estrés definida como:

$$\text{Estrés}(q) = \frac{\sum_{i<j} \left(d_{ij}^{(q)} - \hat{d}_{ij}^{(q)} \right)^2}{\sum_{i<j} \left\{ d_{ij}^{(q)} \right\}^2}$$

2) Para $\hat{d}_{ij}^{(q)}$ fijos, encontrar una nueva configuración de puntos que minimicen la función de Estrés y regresar al Paso 1.
3) Repetir los pasos 1 y 2 hasta que se alcance un mínimo valor de la función de Estrés.

- Evaluación del escalamiento q-dimensional:

Estrés	*Ajuste*
20%	Pobre
10%	Regular
5%	Bueno
2.5%	Excelente
0%	Perfecto

❖ Splus: cmdscale.

7. Análisis de Factores

7.1. Introducción y objetivos.

- El análisis de factores es un procedimiento estadístico que crea un nuevo conjunto de variables no correlacionadas entre sí, llamadas *factores subyacentes o factores comunes*, con la esperanza de que estas nuevas variables proporcionen una mejor comprensión de los datos.

- Uno de los objetivos básicos del análisis de factores es determinar si las p variables respuesta exhiben patrones de relación entre sí, de tal manera que las variables se puedan dividir en m grupos, y que cada grupo conste de variables altamente correlacionadas entre sí, pero bajamente correlacionadas con variables de otros grupos.

- Los OBJETIVOS del análisis de factores son:
  1) Determinar si existe un conjunto más pequeño de variables no correlacionadas que expliquen las relaciones que existen entre las variables originales.
  2) Determinar el número de variables (diferentes) subyacentes.
  3) Interpretar estas nuevas variables.
  4) Evaluar a los individuos del conjunto de datos sobre estas nuevas variables.
  5) Usar estas nuevas variables en análisis estadísticos posteriores.

7.2. Modelo de factores ortogonales.

- Sea X un v.a. de dimensión *p*, con media μ y matriz de varianzas-covarianzas Σ. El modelo general de análisis de factores supone que existen *m* factores subyacentes, denotados por $F_1,F_2,...,F_m$, tales que

$$\begin{aligned} X_1 &= \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1m}F_m + \varepsilon_1 \\ X_2 &= \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \mu_p + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pm}F_m + \varepsilon_p \end{aligned}$$

- Los supuestos del modelo son:
  i) Los F_k tienen media cero y varianza 1, para k=1,...,m y además están no correlacionados.
  ii) Los ε_j tienen media cero y varianza ψ_j, para j=1,...,p.
  iii) F_k y ε_j son independientes para k=1,...,m y j=1,...,p.

- NOTACIÓN MATRICIAL: El modelo se puede expresar como

$$X = \mu + \Lambda F + \varepsilon, \qquad (7.1)$$

donde,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix},\ F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} \text{ y } \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & & \lambda_{pm} \end{pmatrix}$$

- En forma matricial los supuestos quedan como:
  i) $F \sim (0, I)$,
  ii) $\varepsilon \sim (0, \Psi)$, en donde $\Psi = \text{diag}(\psi_1,\psi_2,...,\psi_p)$, y

iii) F y ε son independientes.

- INTERPRETACIONES:

- Las nuevas variables F_k son llamadas factores subyacentes o factores comunes.
- Los términos ε_j son llamados factores específicos y describen la variación residual específica a la variable X_j.
- La cantidad ψ_j es llamada varianza específica de la variable X_j.
- Los coeficientes λ_{jk} son llamados pesos de la j-ésima variable en el k-ésimo factor. De hecho, $\lambda_{jk} = \mathrm{Cov}(X_j, F_k)$.

- COVARIANZA: El modelo (7.1) implica que

$$\mathrm{Var}(X) = \mathrm{Cov}(X) = \mathrm{Cov}(\mu + \Lambda F + \varepsilon),$$

por lo tanto,

$$\Sigma = \Lambda\Lambda' + \Psi. \qquad (7.2)$$

- OBSERVACIONES:

- Si existen Λ y Ψ de modo que la relación (7.2) se satisfaga, entonces los factores comunes explican con exactitud la covarianza entre las variables originales.
- La varianza de X_j se puede dividir de la siguiente manera:

$$\underbrace{\sigma_{jj}}_{} = \underbrace{\lambda_{j1}^2 + \cdots + \lambda_{jm}^2}_{} + \underbrace{\psi_j}_{}$$

$$\mathrm{Var}(X_j) = \text{Comunalidad} + \text{Varianza específica}$$

- Algunas covarianzas son:

$$\mathrm{Cov}(X_i, X_j) = \lambda_{i1}\lambda_{j1} + \cdots + \lambda_{im}\lambda_{jm}$$

- NO UNICIDAD de los factores. Si m > 1, la matriz de pesos de los factores no es única, es decir, si existen Λ y Ψ que satisfacen (7.2), entonces

$$\Lambda\Lambda' = \Lambda TT'\Lambda' = (\Lambda T)(\Lambda T)' = \Lambda^*\Lambda^{*'},$$

  donde T es una matriz ortogonal, i.e., $TT' = I$. Por lo tanto $\Lambda^* = \Lambda T$ y Ψ también satisfacen (7.2).

  - Tomando ventaja de la no unicidad de la matriz de pesos, se pueden obtener distintas matrices rotadas $\Lambda^* = \Lambda T$, para distintas matrices ortogonales T, de tal manera que alguna de ellas produzca unos factores son una interpretación "adecuada".

- SOLUCIONES de la ecuación (7.2). Una solución inicial se puede obtener resolviendo el sistema de ecuaciones numéricamente. Dos de los métodos más comunes son: Método de factores principales y Método de máxima verosimilitud.

- MÉTODOS DE ROTACIÓN. La idea de los métodos de rotación es que se tengan factores fácil de interpretar. Para ello, el objetivo es que las variables originales no tengan peso alto en más de un factor. El método más común es el VARIMAX.

- ¿CUÁNTOS factores son necesarios?. Recuerda que el número de factores comunes o subyacentes es un número fijo que, en principio, se determina a-priori. Una posible elección inicial sería tomar a *m* como el número de componentes significativas en un análisis de componentes principales.

- MARCADORES de los factores. Si los factores resultantes del análisis de factores se van a usar posteriormente, es necesario calcular el valor o marcador de cada factor para cada individuo. Para cada individuo se tiene, $x_i = \Lambda F_i + \varepsilon_i$, en donde la matriz de pesos Λ se estima y las cantidades ε_i son no observables y por lo tanto no se conocen. Existen dos métodos principalmente, el método de Bartlett o de mínimos cuadrados y el método de Thompson o de regresión.

7.2. Cometarios y notas finales.

- DIFERENCIAS entre un *análisis de componentes principales* (ACP) y un *análisis de factores* (AF).
  1) El ACP produce una transformación ortogonal de las variables y no depende de un modelo subyacente, mientras que el AF sí depende de un modelo estadístico.
  2) En el ACP el objetivo es explicar la varianza de las variables originales, mientras que en el AF el objetivo es explicar la estructura de covarianza (correlación) entre las variables.

- NOTA 1: El AF crea un nuevo conjunto de variables no correlacionadas a partir de un conjunto de variables correlacionadas, por lo que si las variables originales son no correlacionadas entonces no tiene sentido aplicar un AF.

- NOTA 2: Algunos estadísticos creen que el análisis de factores no es una técnica estadística válida y útil, esto se debe a la no unicidad de sus

resultados y a la subjetividad relacionada con sus numerosos aspectos (determinación del número de factores, interpretación de los factores, etc.).

- NOTA 3: La presentación de las ideas de AF supone explicar la matriz de varianzas y covarianzas Σ, pero en la práctica este tipo de análisis se hace sobre la matriz de varianzas y covarianzas de las variables estandarizadas, es decir, sobre la matriz de correlaciones de las variables originales P.

8. Análisis discriminante

8.1. Introducción y objetivos.

- El análisis discriminante es también conocido como análisis de clasificación.

- Suponga que se tienen varias poblaciones de las cuales fueron tomadas observaciones. Suponga además que se tiene una nueva observación que proviene de una de estas poblaciones, pero no se sabe cuál. El OBJETIVO básico del análisis discriminante es producir una regla o un esquema de clasificación que nos permita predecir la población más probable de la cual proviene la nueva observación.

- EJEMPLO: Un anestesiólogo necesita determinar si un anestésico es seguro para una persona que están operando del corazón. Con base en ciertas características del paciente como edad, sexo, presión sanguínea, peso, etc., el anestesiólogo tomar una decisión. ¿Cuál sería la probabilidad de equivocarse?

- Se puede decir que el análisis discriminante es semejante al análisis de regresión en el sentido de que una variable respuesta es explicada por varias variables explicativas. La diferencia sería que en el análisis de regresión la variable respuesta es continua, en cambio en el análisis discriminante la variable respuesta es discreta.

8.2. Análisis discriminante para dos poblaciones normales.

- DESCRIPCIÓN del problema. Sean Π_1 y Π_2 dos poblaciones. Cada población está caracterizada por las variables $X_1 \sim N_p(\mu_1, \Sigma_1)$ y $X_2 \sim N_p(\mu_2, \Sigma_2)$ respectivamente, donde $X_k' = (X_{k1}, \ldots, X_{kp})$, para k=1,2. Sea x_F un nuevo vector de observaciones que se sabe proviene de Π_1 o de Π_2. La idea es encontrar una regla de decisión para predecir de cuál de las dos poblaciones es más probable que provenga x_F.

- SOLUCIONES al problema. Existen 4 reglas propuestas para solucionar el problema.

  - *Regla de verosimilitud*:

$$RD_1(x) = \begin{cases} \Pi_1, \text{ si } L(x;\mu_1,\Sigma_1) > L(x;\mu_2,\Sigma_2) \\ \Pi_2, \text{ si } L(x;\mu_1,\Sigma_1) \leq L(x;\mu_2,\Sigma_2) \end{cases},$$

    donde $L(x;\mu_k,\Sigma_k)$ es la función de verosimilitud para la k-ésima población evaluada en x.

  - *Regla de la función discriminante lineal*:

    Cuando dos poblaciones normales multivariadas tienen matrices de varianzas-covarianzas iguales ($\Sigma_1 = \Sigma_2 = \Sigma$), la regla de verosimilitud se simplifica a,

$$RD_2(x) = \begin{cases} \Pi_1, \text{ si } b'x - c > 0 \\ \Pi_2, \text{ si } b'x - c \leq 0 \end{cases},$$

    donde $b = \Sigma^{-1}(\mu_1 - \mu_2)$ y $c = \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$. La función $b'x$ es llamada función discriminante lineal de x.

- *Regla de la distancia de Mahalanobis*:

  Cuando dos poblaciones normales multivariadas tienen matrices de varianzas-covarianzas iguales, la regla de verosimilitud también es equivalente a,

$$RD_3(x) = \begin{cases} \Pi_1, \text{ si } d_1 < d_2 \\ \Pi_2, \text{ si } d_1 \geq d_2 \end{cases},$$

  donde $d_k = (x - \mu_k)' \Sigma^{-1} (x - \mu_k)$, para k=1,2. La cantidad d_k es una medida de la distancia entre x y la media de la k-ésima población.

- Regla de la probabilidad posterior:

  Cuando las matrices de varianza-covarianzas son iguales, una regla de decisión sería,

$$RD_4(x) = \begin{cases} \Pi_1, \text{ si } P(\Pi_1|x) > P(\Pi_2|x) \\ \Pi_2, \text{ si } P(\Pi_1|x) \leq P(\Pi_2|x) \end{cases},$$

  donde

$$P(\Pi_k|x) = e^{-\frac{d_k}{2}} \Big/ e^{-\frac{d_1}{2}} + e^{-\frac{d_2}{2}}$$

  es llamada probabilidad posterior de la población Π_k dado x, para k=1,2. En realidad la probabilidad posterior no es una probabilidad verdadera porque no se está considerando ningún evento aleatorio. La aleatoriedad proviene de tomar la decisión correcta. Por ejemplo, la decisión no se tomaría con tanta confianza si $P(\Pi_1|x) = 0.53$ y $P(\Pi_2|x) = 0.47$, que si $P(\Pi_1|x) = 0.96$ y $P(\Pi_2|x) = 0.04$.

- NOTA 1: Las 4 reglas discriminantes anteriores son equivalentes cuando las matrices de varianzas-covarianzas son iguales en las dos poblaciones. Es decir, las cuatro reglas asignarán a un nuevo individuo al mismo grupo.

- REGLAS DISCRIMINANTES MUESTRALES. Si no se conoce el valor poblacional de μ_1, μ_2, Σ_1, y Σ_2, estos parámetros se pueden estimar mediante los estimadores insesgados correspondientes $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\Sigma}_1$ y $\hat{\Sigma}_2$ y proceder de igual manera. Si se cree que las matrices de varianzas-covarianzas poblacionales son iguales, entonces una estimación combinada de la matriz común Σ sería,

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2}{n_1 + n_2 - 2},$$

  en donde n_1 y n_2 son los tamaños de las muestras de Π_1 y Π_2.

- PROBABILIDADES DE CLASIFICACIÓN ERRÓNEA. Cuando se realiza un análisis discriminante, es necesario determinar o estimar la probabilidad de que la regla de clasificación clasifique erróneamente a un nuevo individuo. Lo ideal sería que este valor fuera cercano a cero. Existen 3 formas de estimar esta probabilidad.

  - *Estimador de resustitución*: Consiste en aplicar la regla discriminante a los mismos datos con los que se construyó la misma regla y determinar la proporción de individuos clasificados erróneamente.

  - *Estimador con una muestra de prueba*: Consiste en dividir a la muestra en dos subconjuntos de observaciones. El primer subconjunto llamado

muestra de prueba servirá para construir la regla de clasificación. Esta regla se aplica al segundo subconjunto de observaciones y se determina la proporción de individuos mal clasificados.

- *Estimador de validación cruzada*: Este método consiste en lo siguiente: Elimine la primera observación de los datos, construya una regla discriminante basada en los datos restantes, use esta regla para clasificar la primera observación y observe si ésta fue clasificada correctamente o no. Reemplace la primer observación al conjunto de datos y elimine la segunda y haga lo mismo que con la primera observación y así sucesivamente con todas las observaciones. Finalmente cuente cuántas observaciones fueron clasificadas erróneamente y divídalas entre el número total de observaciones.

➢ NOTA 2: Existen reglas discriminantes generales para dos poblaciones que toman en cuenta que las consecuencias (costos) de clasificar erróneamente a un individuo de una población u otra son diferentes.

8.3. Análisis discriminante para varias poblaciones.

➢ FUNCIONES DISCRIMINANTES CANÓNICAS. Este método también es conocido como *análisis discriminante de Fisher*. La idea es crear funciones discriminantes como combinaciones lineales de las variables, de tal manera que contengan la mayor cantidad de información posible.

- *Descripción del problema*. Sean $\Pi_1,...,\Pi_m$ m poblaciones definidas por el vector de variables X_k con media μ_k y matriz de varianzas-covarianzas Σ

(las m poblaciones tienen matrices de varianzas-covarianzas iguales). Suponga que se tiene además una muestra de tamaño n_k de cada población.

- *Solución del problema*. Sean

$$B = \sum_{k=1}^{m} n_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})' \quad y \quad W = \sum_{k=1}^{m} \sum_{i=1}^{n_k} (x_{ki} - \hat{\mu}_k)(x_{ki} - \hat{\mu}_k)',$$

donde $\hat{\mu} = \frac{1}{n} \sum_{k=1}^{m} n_k \hat{\mu}_k$ y $n = \sum_{k=1}^{m} n_k$. La matriz B es llamada matriz de varianzas muestrales entre poblaciones y W es llamada matriz de varianzas dentro de las poblaciones (muestras). La idea es encontrar el vector b que maximice

$$\frac{b'Bb}{b'Wb}.$$

Se puede demostrar que el vector que maximiza el cociente anterior es el primer eigenvector a_1 correspondiente al primer eigenvalor λ_1 de la matriz $(W^{-1}B)$. Un vector ortogonal al anterior que maximiza el cociente anterior es el segundo eigenvector a_2 correspondiente al segundo eigenvalor λ_2 de la matriz $(W^{-1}B)$ y así sucesivamente. El número máximo de eigenvalores es min(p,m-1).

- *Regla discriminante*.

Si se usa únicamente la primer función canónica, se calcula $d_k = \left| b_1' x - b_1' \hat{\mu}_k \right|$, para k=1,2,...,m y se asigna x a la población cuyo valor d_k sea el más pequeño.

Si se usan las primeras dos funciones canónicas, se calcula $d_k^2 = \left(b_1'x - b_1'\hat{\mu}_k\right)^2 + \left(b_2'x - b_2'\hat{\mu}_k\right)^2$, para k=1,2,...,m y se asigna x a la población cuyo valor d_k^2 sea el más pequeño.

- ÁRBOLES DE CLASIFICACIÓN (CART). La idea es construir un árbol de clasificación de tal manera que los nodos (puntos) terminales del árbol definan una clase. Las ramas se bifurcan con la respuesta afirmativa o negativa a preguntas formadas a partir de las variables originales.