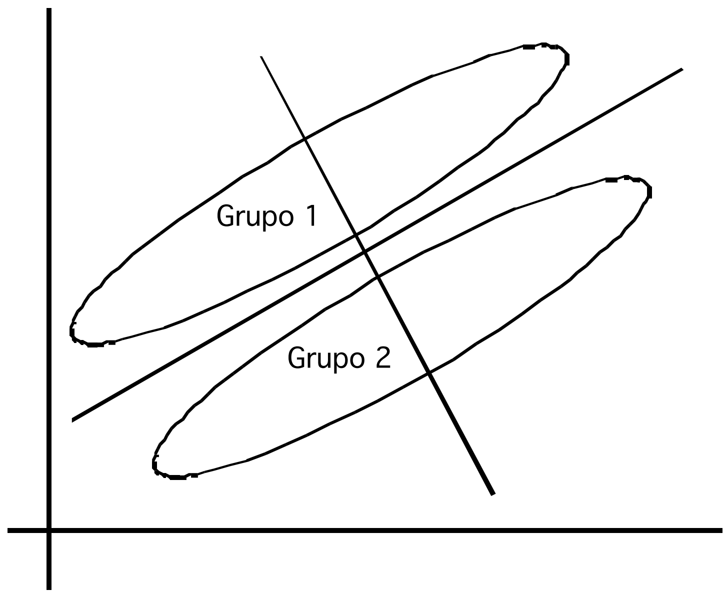


ANALISIS DISCRIMINANTE: INTRODUCCION

José Luis Vicente Villardón
Departamento de Estadística
Universidad de Salamanca

1.- EJEMPLO INICIAL

Consideremos el siguiente problema con datos reales.

Los vinos elaborados en áreas específicas y reconocidos con denominación de origen (DO) son de importancia significativa en las diferentes regiones productoras de vinos. La DO reconoce y garantiza calidad de los vinos fabricados. Consecuentemente, son necesarios una serie de parámetros específicos que permitan a los analistas clasificar distintos vinos en sus correspondientes denominaciones de origen. Entre las características que pueden usarse están la composición en ciertos metales, ácidos orgánicos, ciertos componentes polifenólicos, etc... Los valores de estas características dependen de diversos factores, tales como las variedades de uva empleadas en el proceso de elaboración, o la edad del vino.

Para la clasificación de los vinos en su correspondiente denominación de origen, es posible utilizar técnicas de Análisis Multivariante entre las que se incluyen las técnicas de Análisis Discriminante.

Los autores consiguen buenos resultados combinando ambas técnicas y con buenos porcentajes de clasificación para las denominaciones de origen de “Ribera de Duero y “Toro”.

Se ha realizado un estudio sobre las dos denominaciones de origen de vinos castellanos (Ribera de Duero y Toro) en dos años diferentes (1986, 1987), con el fin de distinguir las características diferenciales entre las dos denominaciones, mediante medidas objetivas obtenidas en laboratorio, de forma que pueda evitarse el fraude en las etiquetas de la denominación sustituyendo ambos vinos debido a su proximidad espacial.

Se han considerado 4 grupos diferentes procedentes de la combinación de denominaciones y años (RD1986, RD1987, T1986, T1987). Se ha considerado el año como posible factor de confusión en la clasificación de los vinos de las dos denominaciones.

Se han considerado 18 variables etiquetadas sobre al gráfico como:
Grad:Grado alcohólico, **AcVo**: Acidez Volatil **AcTo**:Acidez Total **AcFi**: Acid. Fija **pH** **Foli**: Fenoles tot (Folin) **Some**: Fenoles (Sommers) **SRV:**Sust. reactivas a la vanilina **Proc**: Procianidoles **ACRG:** Antocianos1 **ACSE**: Antocianos2 **ACHP:**Antocianos 3 **IC** : Indice de color 1 **IC2** : Indice de color 2 **Tono:** de color **IIm** : Indice de ionización. **EQ1:** Edad química **V/LA**

Disponemos ahora de una matriz de datos multivariantes cuyas filas están divididas en g grupos de tamaño n_i, definidos a priori.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_g \end{bmatrix}$$

2.- OBJETIVO GENERAL

ESTUDIAR LAS DIFERENCIAS ENTRE LOS GRUPOS Y CARACTERIZARLAS MEDIANTE TECNICAS MULTIVARIANTES.

Una primera alternativa al análisis de los datos, que es la que la mayor parte de los investigadores no expertos en estadística utilizan, sería el análisis de todas las variables por separado. Podemos hacer varias objeciones a dicha práctica:

- Las diferencias más importantes entre los grupos pueden no encontrarse en ninguna de las variables separadas sino en una combinación de las mismas. (Ver Figura 1).
- Algunas de las variables pueden contener información concomitante en relación a la separación de los grupos. Sería deseable extraer la información común para obtener una reducción de la dimensión que ayude a la interpretación de las diferencias.

- Si el propósito es establecer la significación estadística, se produce un considerable incremento en el riesgo tipo I al realizar repetidos contrastes

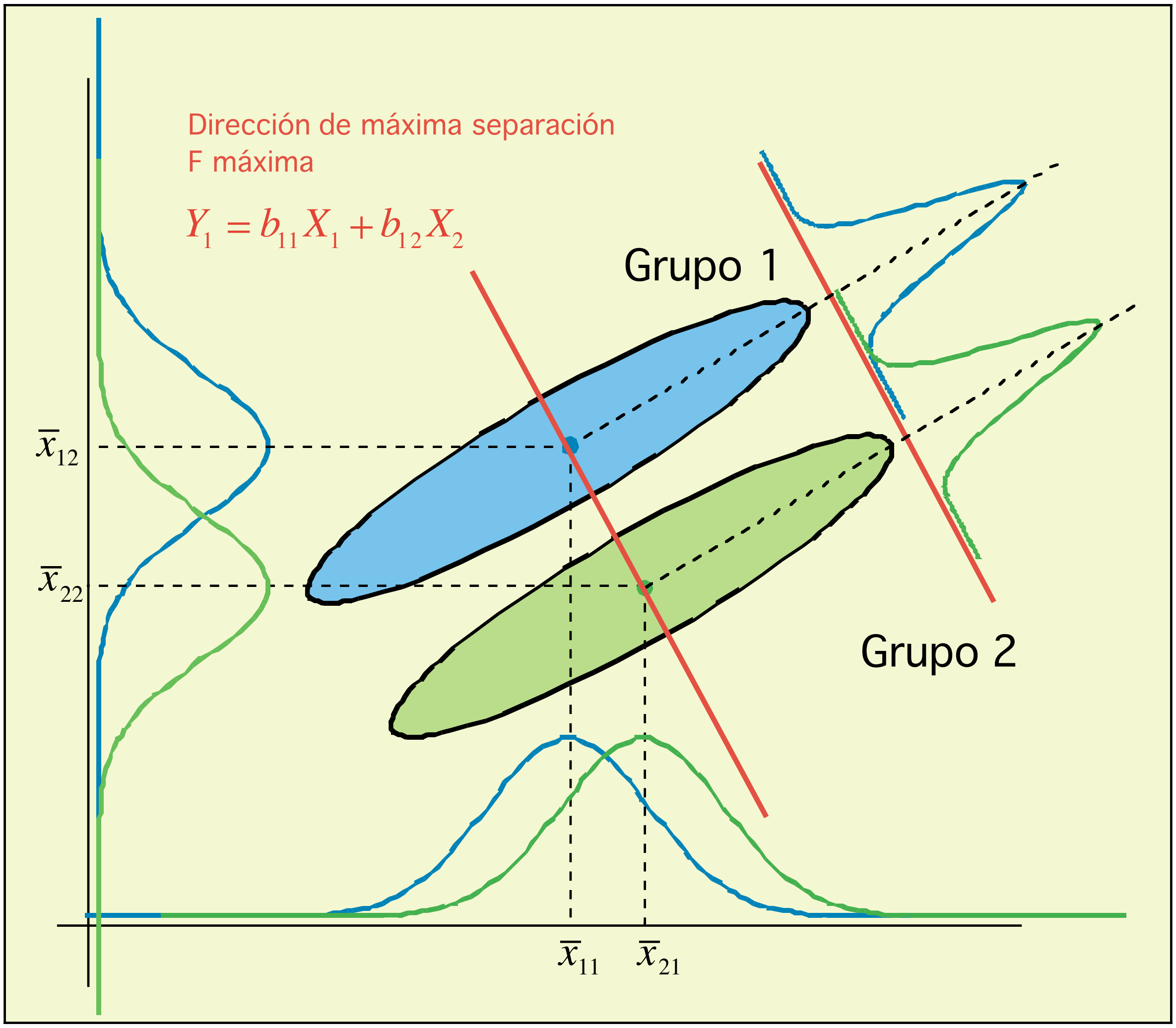


Figura 1: Representación esquemática de la diferenciación multivariante de varios grupos.

La técnica tradicional para la reducción de la dimensión utilizando la información común de grupos de variables es el Análisis de Componentes principales, sin embargo, es claro que necesitamos un nuevo grupo de técnicas ya que, en general, las direcciones de máxima variabilidad no coinciden con las direcciones de máxima separación entre grupos.

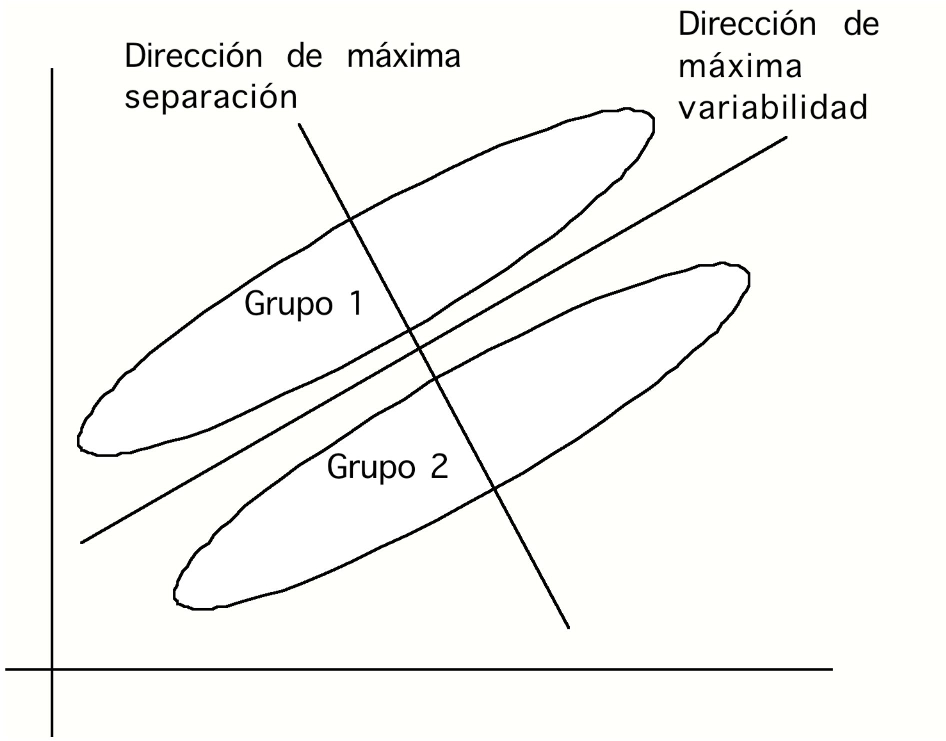


Figura 2 : Diferencia entre las direcciones de máxima variabilidad y máxima separación entre grupos.

3.- OBJETIVOS PARTICULARES

Dentro del objetivo general de caracterizar las diferencias entre los grupos, podemos tener objetivos específicos:

- Si el objetivo es la caracterización de las diferencias entre los vectores de medias mediante contrastes estadísticos, la técnica que utilizaremos es el Análisis Multivariante de la Varianza (MANOVA). El paso inicial consiste en contrastar si los vectores medios de las poblaciones de las que provienen los datos son iguales en todas ellas. El MANOVA es una extensión del ANOVA en la que se busca la combinación lineal de las variables observadas que produce una F mayor en la variable combinada. Los resultados se muestran en forma de tests estadísticos para el conjunto de todos los grupos y comparaciones por parejas de los mismos o contrastes que incluyan varios. Un error habitual es realizar ANOVAS para cada una de las variables por separado una vez que se ha decidido que los vectores medios son diferentes. El MANOVA es un Modelo Lineal General Multivariante (MLGM) en el que las variables continuas son las dependientes y los grupos son las regresoras.

- Si el objeto es más descriptivo y consiste en la representación de la estructura de los grupos en dimensión reducida, la técnica que utilizaremos será el denominado Análisis Canónico de Poblaciones (en otros contextos se denomina Análisis Discriminante Descriptivo, Análisis Factorial Discriminante, Análisis de Variables Canónicas o Coordenadas Discriminantes). La interpretación de la solución final se realiza a través de la representación de los centroides de las poblaciones en un diagrama de dispersión en dimensión reducida y de las correlaciones entre las variables canónicas y las variables originales de la misma manera que en el Análisis Factorial.

- El Análisis Canónico de Poblaciones (ACPo) no tiene información directa sobre las variables responsables de la separación. Si se desea añadir dicha información en el gráfico puede utilizarse un Biplot Canónico ó MANOVA Biplot.
- Si el propósito es la clasificación de un individuo en uno de los grupos, la técnica a utilizar es el Análisis Discriminante (Predictivo) - ADP. A veces se utiliza el ACPo como representación gráfica del ADP si bien los propósitos de ambas técnicas no son idénticos. En ADP los grupos pueden considerarse como las dependientes y las variables continuas como las regresoras por lo que también tiene analogías con la regresión en lo que se refiere a la interpretación de las soluciones aunque no es una regresión propiamente dicha. Dentro del ADP se consideran muchas técnicas diferentes como el Discriminante Lineal, el Discriminante Cuadrático, el Discriminante Logístico, los Métodos de Aprendizaje Supervisado y muchos otros.

Comparación de los grupos a través de sus vectores de medias	Análisis Multivariante de la varianza
Representación de la estructura de los grupos en dimensión reducida	Análisis Canónico (de poblaciones). (Análisis Discriminante Descriptivo).
Representación simultanea de la estructura de los grupos y de las variables responsables de la separación.	Biplot Canónico o MANOVA Biplot.
Clasificar un nuevo individuo en una de varias poblaciones	Análisis Discriminante (lineal, cuadrático, logístico)

El análisis Canónico sirve como representación gráfica de la hipótesis alternativa en el Análisis Multivariante de la Varianza y como representación de las direcciones de máxima discriminación en el Análisis Discriminante.

4.- Notación previa

X matriz de datos con n individuos y p variables, la suponemos centrada con respecto a las medias globales. Las filas de la matriz están divididas en g grupos exhaustivos y mutuamente excluyente, es decir, cada uno de los individuos pertenece a uno de los grupos y solo a uno. Los tamaños muestrales de cada grupo se denotan con n_i (i= 1, …, g) y $n = n_1 + \ldots + n_g$. La matriz diagonal con los tamaños muestrales se denota con $\mathbf{D}_g = diag(n_1, \ldots, n_g)$ y los vectores de medias de cada uno de los grupos como $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \ldots, \bar{x}_{1p})'$. $\bar{\mathbf{X}} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_g)'$ es la matriz de g filas y p columnas con las medias de los g grupos en las p variables.

$\mathbf{Q}_h = \bar{\mathbf{X}}'\mathbf{D}_g\bar{\mathbf{X}}$ Sumas de cuadrados entre los grupos. Es el análogo a la suma de cuadrados entre grupos del Análisis de la Varianza univariante.

$\mathbf{Q}_t = \mathbf{X}'\,\mathbf{X}$ Suma de cuadrados total

$\mathbf{Q}_e = \mathbf{Q}_t - \mathbf{Q}_h$ Sumas de cuadrados dentro de los grupos o residual

$\mathbf{S}_t = \frac{1}{n-1}\mathbf{Q}_t$ Matriz de covarianzas total

$\mathbf{S}_h = \frac{1}{g-1}\mathbf{Q}_h$ Matriz de covarianzas entre grupos

$\mathbf{S}_e = \frac{1}{n-g}\mathbf{Q}_e = \mathbf{S}$ Matriz de covarianzas dentro de los grupos

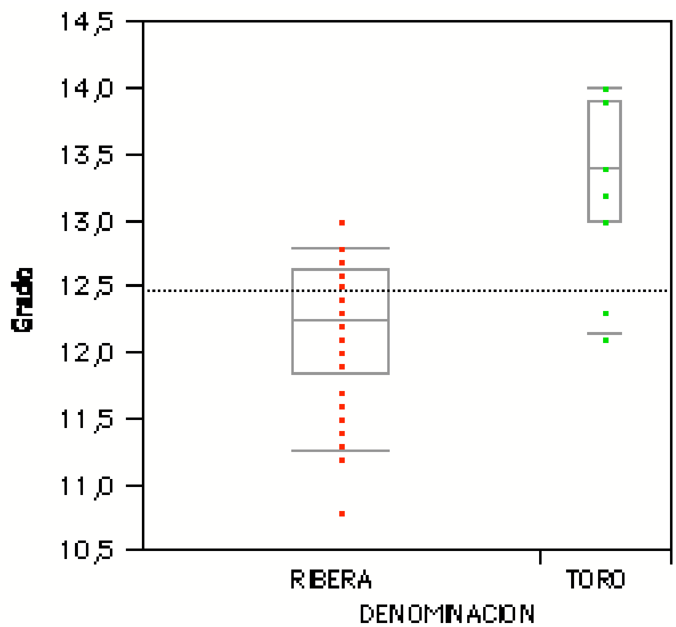
Es la media ponderada de las matrices de covarianzas de cada grupo por separado, y estima la matriz de covarianzas común a todos ellos.

$(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_i)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_i)$ Distancia de Mahalanobis entre dos grupos

5.- ANALISIS DISCRIMINANTE CON UNA VARIABLE

El problema consiste primero, en determinar si existen diferencias entre las medias y después, en obtener una función que nos permita la clasificación de los vinos de las dos denominaciones, es decir, dada una nueva observación para las 18 variables y con la denominación desconocida, se trata de encontrar una función a partir de la cual podamos predecir la denominación.

Vamos a tomar inicialmente una sola variable, por ejemplo el grado alcohólico. En el gráfico siguiente se muestran los cuantiles de las



distribuciones de frecuencias para las dos denominaciones.

Obsérvese como parece bastante claro que, en general, el grado alcohólico de ambas denominaciones es bastante diferente, mayor para la denominación de toro. Como regla básica tendremos, entonces, que si el grado alcohólico es alto, clasificaremos la observación como de Toro, mientras que si es bajo la clasificaremos como de Ribera de Duero. El problema fundamental es encontrar el punto de corte, es decir, a partir de que valor clasificamos a una nueva observación como de Toro.

Para la selección del punto de corte podemos pensar en varias formas alternativas, pero vamos a utilizar una basada en argumentos puramente descriptivos.

Como representante de cada uno de los grupos elegimos la media. Incluso si realizamos un contraste t veremos que las medias son significativamente diferentes.

Unpaired t-test for Grado
Grouping Variable: DENOMINACION
Hypothesized Difference = 0

	Mean Diff.	DF	t-Value	P-Value
RIBERA, TORO	-1,184	43	-5,708	<,0001

F Test for Grado
Grouping Variable: DENOMINACION
Hypothesized Ratio = 1

	Var. Ratio	Num. DF	Den. DF	F-Value	P-Value
RIBERA, TORO	,713	33	10	,713	,5159

Group Info for Grado
Grouping Variable: DENOMINACION

	Count	Mean	Variance	Std. Dev.	Std. Err
RIBERA	34	12,171	,327	,572	,098
TORO	11	13,355	,459	,677	,204

Además, la información de la que disponemos es compatible con la hipótesis de que la variabilidad es la misma en ambas poblaciones.

El punto de corte que elegimos para la clasificación es el punto medio entre las dos medias muestrales

$$c = \frac{\overline{x}_1 - \overline{x}_2}{2}$$

es decir, (12,171+13,355)/2=12,763.

La regla es ahora simple, clasificamos una observación nueva como de Toro si su grado alcohólico es mayor o igual que 12,763 y como de Ribera de Duero en caso contrario.

Es obvio que la regla no es perfecta, ya que, en contraremos vinos de Ribera con un grado por encima de ese valor y vinos de Toro con grados por debajo. Tenemos que valorar entonces la bondad de la regla obtenida, la valoración la haremos a través del porcentaje de mal clasificados.

FRECUENCIAS

DENOMINACION REAL	PREDICCION RIBERA	PREDICCION TORO	Totals
RIBERA	29	5	34
TORO	2	9	11
Totals	31	14	45

PORCENTAJES

DENOMINACION REAL	PREDICCION RIBERA	PREDICCION TORO	Totals
RIBERA	85,294	14,706	100,000
TORO	18,182	81,818	100,000

Con el criterio desarrollado, 5 de las 34 observaciones de Ribera (14,706%) y 2 de las 11 (18,182%) de Toro, se han clasificado incorrectamente, es decir 7 de las 45 observaciones se han clasificado incorrectamente, el 15,556% de las observaciones se ha clasificado incorrectamente.

Si tenemos más de dos variables el problema es un poco más complejo ya que se necesita utilizar una técnica multivariante.

6.- ANALISIS DISCRIMINANTE PARA DOS GRUPOS

Buscamos una nueva variable **y**, combinación lineal de las variables observadas **y = Xa**, que muestre las mayores diferencias entre las medias de los dos grupos de forma que nos permita la clasificación de uno de ellos con la máxima resolución posible.

Las media de los valores de la nueva variable para cada grupo son

$$\bar{y}_1 = \mathbf{a}' \, \bar{\mathbf{x}}_1$$

$$\bar{y}_2 = \mathbf{a}' \, \bar{\mathbf{x}}_2$$

La diferencia de las medias es, entonces,

$$\bar{y}_1 - \bar{y}_2 = \mathbf{a}'\,\bar{\mathbf{x}}_1 - \mathbf{a}'\,\bar{\mathbf{x}}_2 = \mathbf{a}'(\,\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Se trata, por tanto, hacer máximo

$$\left|\mathbf{a}'\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right|$$

sujeto a la restricción $\mathbf{a}'\,\mathbf{S}\,\mathbf{a} = 1$ para evitar las indeterminaciones en los coeficientes producidas por la indeterminación en la escala de la variable combinada. La restricción significa que la variabilidad dentro de los grupos en la nueva variable es la unidad.

La solución viene dada por

$$\mathbf{a} = \mathbf{S}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

y la función discriminante lineal es

$$\mathbf{y} = \mathbf{Xa} = \mathbf{XS}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Los elementos de **a** dan la importancia relativa de cada variable. Para hacerlos de magnitud comparable pueden estandarizarse de la misma forma que se hace en regresión.

La función discriminante puede usarse para clasificar nuevos individuos en uno de los dos grupos, de la misma forma que lo hacíamos con una sola variable.

Después de obtener los coeficientes, los valores medios de la función discriminante para los dos grupos serán

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

$$\bar{y}_2 = (\,\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\bar{\mathbf{x}}_2.$$

Y el punto medio de ambos

$$\bar{y} = (\bar{y}_1 + \bar{y}_2) = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

que puede ser utilizado como punto de corte para la clasificación.

Es decir:

Asignamos el individuo al grupo 1 si

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{x} > \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

y al grupo 2 si

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{x} \le \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

ó bien si tomamos

$$W = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

la regla es, asignar a la población 1 si $W > 0$ y si no, asignar a la población 2.

7.- DISTANCIA DE MAHALANOBIS

La distancia de Mahalanobis (al cuadrado) entre dos individuos con vectores de observaciones $\mathbf{x}$ y $\mathbf{z}$, es

$$d_M^2 = d_M^2(\mathbf{x},\mathbf{z}) = (\mathbf{x}-\mathbf{z})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{z})$$

La distancia de Mahalanobis de un individuo al grupo i es la distancia al centroide del grupo

$$d_M^2 = d_M^2(\mathbf{x},\bar{\mathbf{x}}_i) = (\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)$$

y la distancia entre dos grupos es la distancia entre sus centroides

$$d_M^2 = d_M^2(\bar{\mathbf{x}}_i,\bar{\mathbf{x}}_j) = (\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_j)$$

Propiedades

- La distancia de Mahalanobis tiene en cuenta las correlaciones entre las variables utilizando sólo la información d cada variable no redundante.
- Es Invariante por transformaciones lineales no singulares, en particular por cambios de escala.

8.- INTERPRETACIÒN GEOMÉTRICA DEL ANÁLISIS DISCRIMINANTE

Geométricamente el criterio consiste en asignar el individuo a la población mas cercana, midiendo la cercanía a partir de la distancia de Mahalanobis.

Geométricamente el criterio consiste en asignar el individuo a la población mas cercana, midiendo la cercanía a partir de la distancia de Mahalanobis. La regla es, asignamos la observación a la población 1 si

$$d_M^2(\mathbf{x},\bar{\mathbf{x}}_1) < d_M^2(\mathbf{x},\bar{\mathbf{x}}_2) \text{ ó } d_M^2(\mathbf{x},\bar{\mathbf{x}}_2) - d_M^2(\mathbf{x},\bar{\mathbf{x}}_1) > 0$$

La regla puede expresarse como

$$\begin{aligned} d_M^2(\mathbf{x},\bar{\mathbf{x}}_2) - d_M^2(\mathbf{x},\bar{\mathbf{x}}_1) &= (\mathbf{x}-\bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2) - (\mathbf{x}-\bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1) \\ &= \mathbf{x}'\mathbf{S}^{-1}\mathbf{x} + \bar{\mathbf{x}}_2'\mathbf{S}^{-1}\bar{\mathbf{x}}_2 - 2\mathbf{x}'\mathbf{S}^{-1}\bar{\mathbf{x}}_2 - \mathbf{x}'\mathbf{S}^{-1}\mathbf{x} - \bar{\mathbf{x}}_2'\mathbf{S}^{-1}\bar{\mathbf{x}}_2 + 2\mathbf{x}'\mathbf{S}^{-1}\bar{\mathbf{x}}_2 \\ &= (\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_2+\bar{\mathbf{x}}_1) + 2\mathbf{x}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)' > 0 \end{aligned}$$

que es idéntica a la regla desarrollada anteriormente.

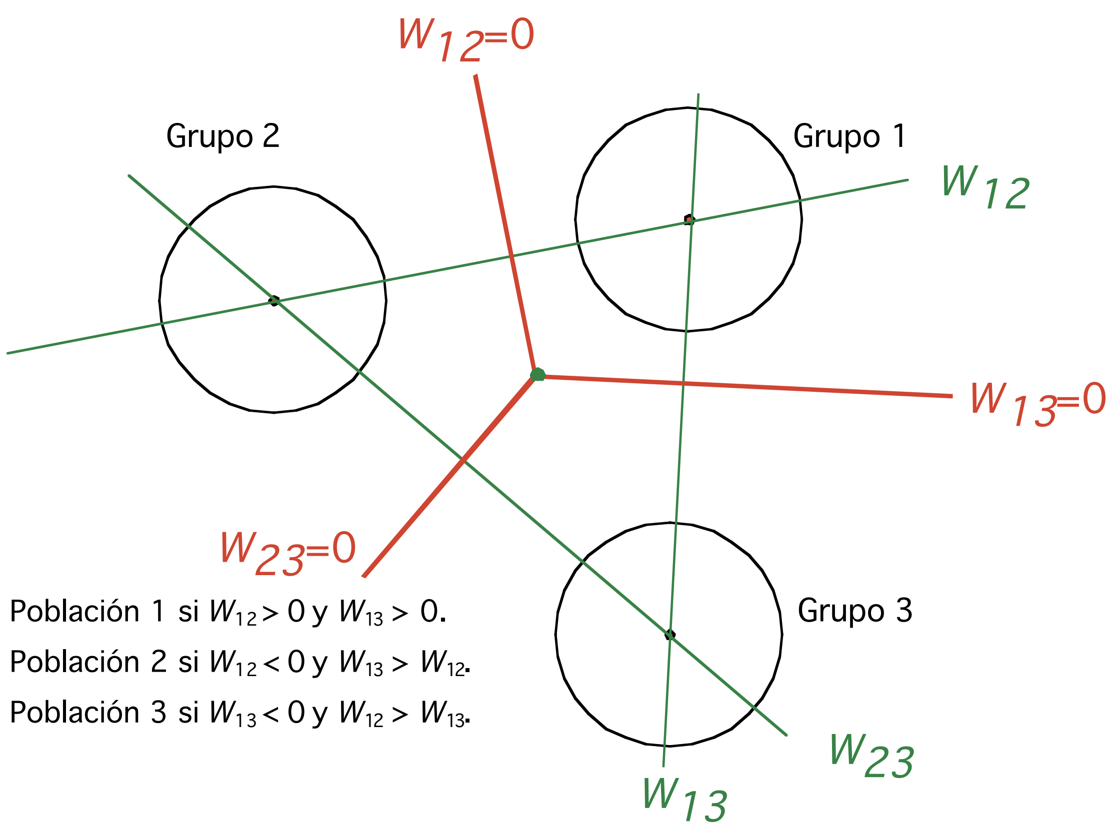
9.- CLASIFICACION PARA VARIOS GRUPOS

Cuando disponemos de varios grupos tenemos varias posibles reglas de clasificación por parejas

$$W_{ij} = (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)'\mathbf{S}^{-1}\mathbf{x} - \tfrac{1}{2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i + \overline{\mathbf{x}}_j)$$

aunque una de ellas es redundante.

Por ejemplo, con tres grupos, tenemos 3 reglas posibles, pero solamente 2 son necesarias



10.- MEDIDA DE LA BONDAD DE LA CLASIFICACION: PROBABILIDAD DE CLASIFICACION ERRONEA.

Como medida de la bondad de la clasificación se suele utilizar la probabilidad de clasificación errónea, es decir, el número de individuos mal clasificados dividido por el número total de individuos.

La probabilidad de clasificación errónea queda subestimada cuando se realiza sobre el mismo conjunto de individuos que se utilizó para estimar la función discriminante. Para evitar esto, pueden utilizarse dos conjuntos de individuos, uno para estimar la función y otro para valorar la clasificación.

Otra forma de valoración puede realizarse clasificando cada individuo a partir de la función calculada con el resto.

La valoración puede realizarse también asignando distribuciones de probabilidad a cada una de las poblaciones y utilizándolas para calcular las probabilidades.

11.- OTRAS TÉCNICAS DISCRIMINANTES

11.1 DISCRIMINANTE BASADO EN DISTRIBUCIONES DE PROBABILIDAD

Cuando es posible asignar distribuciones de probabilidad f_1 y f_2 a cada una de las poblaciones, la regla discriminante para una observación **x** es

Asignar a la población 1 si

$$f_1(\mathbf{x}) > f_2(\mathbf{x})$$

ó bien

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1$$

ó también

$$\log\left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right] > 0$$

Si se conocen las probabilidades a priori π_1 y π_2 de que los individuos pertenezcan a cada una de las poblaciones, la regla sería

Asignar a la población 1 si

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_1}{\pi_2}$$

ó

$$\log\left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right] > \log\left[\frac{\pi_1}{\pi_2}\right]$$

En definitiva se trata de asignar la observación a aquella población que tenga la verosimilitud más alta.

En este caso la probabilidad de mala clasificación puede realizarse utilizando distribuciones normales, ya que como la función discriminante es una combinación de variables normales, también tiene una distribución normal.

La probabilidad de clasificación errónea sería

$$\pi_1 P(2/1) + \pi_2 P(1/2)$$

donde $P(2/1)$ y $P(1/2)$ son las probabilidades de clasificación errónea en cada población.

Generalmente, los parámetros de las funciones probabilidad son desconocidos, por lo que es necesario estimarlos a partir de una muestra.

Cuando se supone que las dos poblaciones tienen distribuciones normales multivariantes, que la matriz de covarianzas es la misma en todos los grupos y que las probabilidades a priori se estiman a partir de los datos muestrales como la proporción muestral en cada grupo, entonces el criterio utilizando la función discriminante lineal coincide con el criterio basado en la distribución de probabilidad.

La media de la distribución se estima a partir del los vectores de medias muestrales y la matriz de covarianzas común a todos los grupos se estima mediante la matriz de covarianza dentro de los grupos.

12.- RESULTADOS DEL ANALISIS DISCRIMINANTE DEL EJEMPLO DE LOS VINOS DE RIBERA Y TORO

```
 - - - - - - - -   D I S C R I M I N A N T   A N A L Y S I S   - - - -

On groups defined by DENOMINA  DENOMINACION
            45 (Unweighted) cases were processed.
             0 of these were excluded from the analysis.
            45 (Unweighted) cases will be used in the analysis.

Number of cases by group
                Number of cases
  DENOMINA  Unweighted     Weighted  Label
         1          34         34,0  RIBERA
         2          11         11,0  TORO
     Total          45         45,0

Group means
  DENOMINA         GRADO           AVOL           ATOT           ACFI
         1      12,17059         ,52206        5,26471        4,61471
         2      13,35455         ,70455        4,83636        3,95455
     Total      12,46000         ,56667        5,16000        4,45333

  DENOMINA            PH          FOLIN         SOMERS            SRV
         1       3,57647     1834,32353       30,70000      787,76471
         2       3,60000     2426,72727       44,24545     1145,81818
     Total       3,58222     1979,13333       34,01111      875,28889

  DENOMINA       PROCIAN           ACRG           ACSE         ACHPLC
         1    2551,44118      246,00000      165,47059      122,88235
         2    3466,36364      275,81818      204,18182      120,00000
     Total    2775,08889      253,28889      174,93333      122,17778

  DENOMINA            IC            IC2           TONO            IIM
         1       4,48647        5,06618         ,70194       24,43824
         2       6,19636        7,16727         ,79736       18,11818
     Total       4,90444        5,57978         ,72527       22,89333

  DENOMINA           EQ1            VLA
         1        ,41738         ,31529
         2        ,52545         ,33091
     Total        ,44380         ,31911
```

INFORMACIÓN DE CONTROL, VECTORES DE MEDIAS PARA CADA UNO DE LOS GRUPOS Y MEDIAS GLOBALES

```
Wilks' Lambda (U-statistic) and univariate F-ratio
with 1 and 43 degrees of freedom

  Variable   Wilks' Lambda         F         Significance
  --------   -------------   -------------   ------------
  GRADO          ,56896        32,5764          ,0000
  AVOL           ,91510         3,9894          ,0521
  ATOT           ,96065         1,7616          ,1914
  ACFI           ,86245         6,8582          ,0121
  PH             ,99560          ,1900          ,6651
  FOLIN          ,75615        13,8670          ,0006
  SOMERS         ,59658        29,0771          ,0000
  SRV            ,67477        20,7253          ,0000
  PROCIAN        ,75294        14,1093          ,0005
  ACRG           ,95961         1,8097          ,1856
  ACSE           ,88797         5,4252          ,0246
  ACHPLC         ,99904          ,0412          ,8401
  IC             ,79888        10,8253          ,0020
  IC2            ,78710        11,6310          ,0014
  TONO           ,82431         9,1647          ,0042
  IIM            ,79674        10,9702          ,0019
  EQ1            ,85751         7,1451          ,0106
  VLA            ,98831          ,5086          ,4796
```

ANALISIS DE LA VARIANZA PARA CADA UNA DE LAS VARIABLES OBSERVADAS

```
On groups defined by DENOMINA  DENOMINACION
Analysis number       1
Direct method:  all variables passing the tolerance test are entered.
     Minimum tolerance level..................  ,00100

Canonical Discriminant Functions
     Maximum number of functions.............       1
     Minimum cumulative percent of variance...  100,00
     Maximum significance of Wilks' Lambda....  1,0000

Prior probability for each group is  ,50000
                       Canonical Discriminant Functions
                 Pct of   Cum  Canonical  After  Wilks'
 Fcn Eigenvalue Variance  Pct     Corr      Fcn  Lambda  Chi-square  df  Sig
                                         :    0 ,139449     66,982    18  ,0000
  1*    6,1711  100,00  100,00    ,9277 :
   * Marks the 1 canonical discriminant functions remaining in the analysis.
```

```
Standardized canonical discriminant function coefficients

             Func  1

GRADO        -,18054
AVOL        -2,14666
ATOT         4,06411
ACFI        -3,17326
PH            ,79052
FOLIN         ,37804
SOMERS       -,28773
SRV         -6,81762
PROCIAN      5,76982
ACRG          ,76702
ACSE        -2,45466
ACHPLC       1,79418
IC           1,52369
IC2         -1,05661
TONO         -,30308
IIM           ,86023
EQ1          -,19725
VLA          4,50072
```

COEFICIENTES DE LA FUNCIÓN DISCRIMINANTE ESTANDARIZADOS

```
Structure matrix:

Pooled within-groups correlations between discriminating variables
                                  and canonical discriminant functions
(Variables ordered by size of correlation within function)
             Func  1
GRADO        -,35038
SOMERS       -,33103
SRV          -,27947
PROCIAN      -,23059
FOLIN        -,22860
IC2          -,20936
IIM           ,20333
IC           -,20198
TONO         -,18584
EQ1          -,16409
ACFI          ,16076
ACSE         -,14299
AVOL         -,12261
ACRG         -,08258
ATOT          ,08148
VLA          -,04378
PH           -,02676
ACHPLC        ,01246
```

CORRELACIONESENTRE LAS VARIABLES ORIGINALES Y LAS VARIABLES DISCRIMINANTES

```
Canonical discriminant functions evaluated at group means (group centroids)

   Group      Func   1

       1       1,38123
       2      -4,26925



Classification results -

                      No. of    Predicted Group Membership
   Actual Group        Cases          1          2
--------------------  ------   --------   --------

Group       1             34         34          0
RIBERA                             100,0%        ,0%

Group       2             11          0         11
TORO                                  ,0%     100,0%

Percent of "grouped" cases correctly classified: 100,00%
```

MEDIAS SOBRE LAS VARIABLES CANONICAS Y PORCENTAJE DE CLASIFICACION CORRECTA

13.- DISCRIMINANTE LOGISTICO

Cuando no se verifican las condiciones de aplicación del análisis discriminante (distribuciones normales y varianzas iguales) puede utilizarse el denominado discriminante logístico basado en la regresión logística.

En este análisis tratamos de estimar la probabilidad de que un individuo pertenezca a cada uno de los grupos cuando tiene una combinación concreta de variables explicativas, mediante un modelo de respuesta logística de la forma (para dos poblaciones)

$$P(i \in Pob\,1/\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\beta}}{1+e^{\mathbf{x}_i'\beta}} = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}$$

Naturalmente

$$P(i \in Pob\,2/\mathbf{x}_i) = 1 - P(i \in Pob\,1/\mathbf{x}_i) = \frac{1}{1+e^{\mathbf{x}_i'\beta}}$$

Una vez que se han estimado los parámetros y se han calculado la probabilidades de pertenencia a cada una de las poblaciones, el individuo será asignado a aquella población parta la que la probabilidad sea mayor, es decir

Asignar a la población 1 si $P(i \in Pob\,1/\mathbf{x}_i) > 0.5$ y a la Población 2 en caso contrario.

El resto del análisis sería análogo al realizado en el Discriminante Lineal.