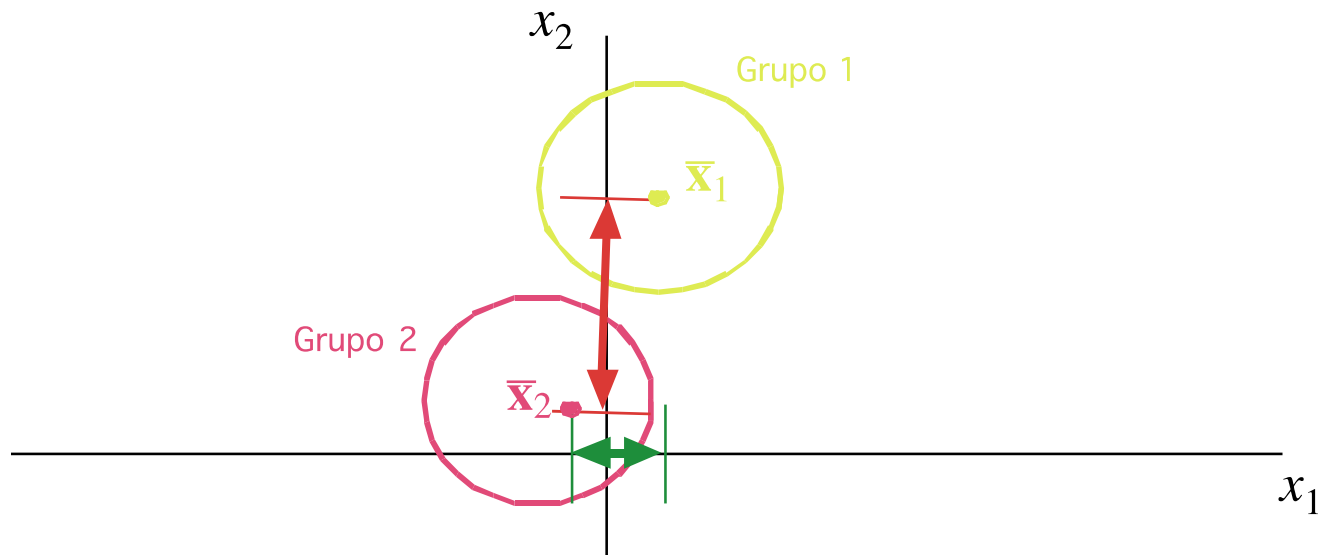


# Estudio de las diferencias entre grupos



Dpto. de Estadística  
Universidad de Salamanca

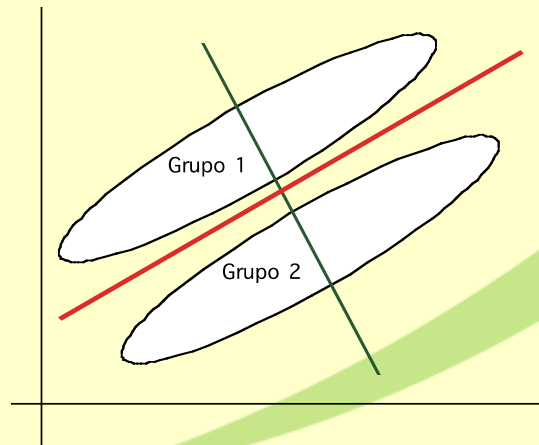
# Estudio de las diferencias entre grupos

- \* Análisis Discriminante
- \* Análisis Multivariante de la Varianza MANOVA
- \* Análisis Canónico
- \* Biplot Canónico

# Análisis de la estructura de grupos

## Objetivos particulares

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



# Análisis Discriminante Lineal

Supongamos que un conjunto de individuos está ya clasificado en una serie de grupos, es decir, se sabe previamente a qué grupos pertenecen. El Análisis Discriminante se puede considerar como un “análisis de regresión” donde la variable dependiente es categórica y tiene como categorías cada uno de los grupos, y las variables independientes son continuas y determinan a qué grupos pertenecen los objetos.

Se pretende **encontrar relaciones lineales** entre las variables continuas **que** mejor **discriminen** en los grupos a los individuos.

Un segundo objetivo es construir **una regla de decisión** que asigne un individuo nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados con un cierto grado de riesgo.

## Ejemplo inicial: Caracterización de vinos tintos jóvenes

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.

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.

# Ejemplo inicial 1 : Caracterización de vinos tintos jóvenes

Variables medidas:

**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**

$$n = 45$$

$$n_1 = 14$$

$$n_2 = 20$$

$$n_3 = 6$$

$$n_4 = 5$$

18 variables

Ribera de Duero 1986

Ribera de Duero 1987

Toro 1986

Toro 1987

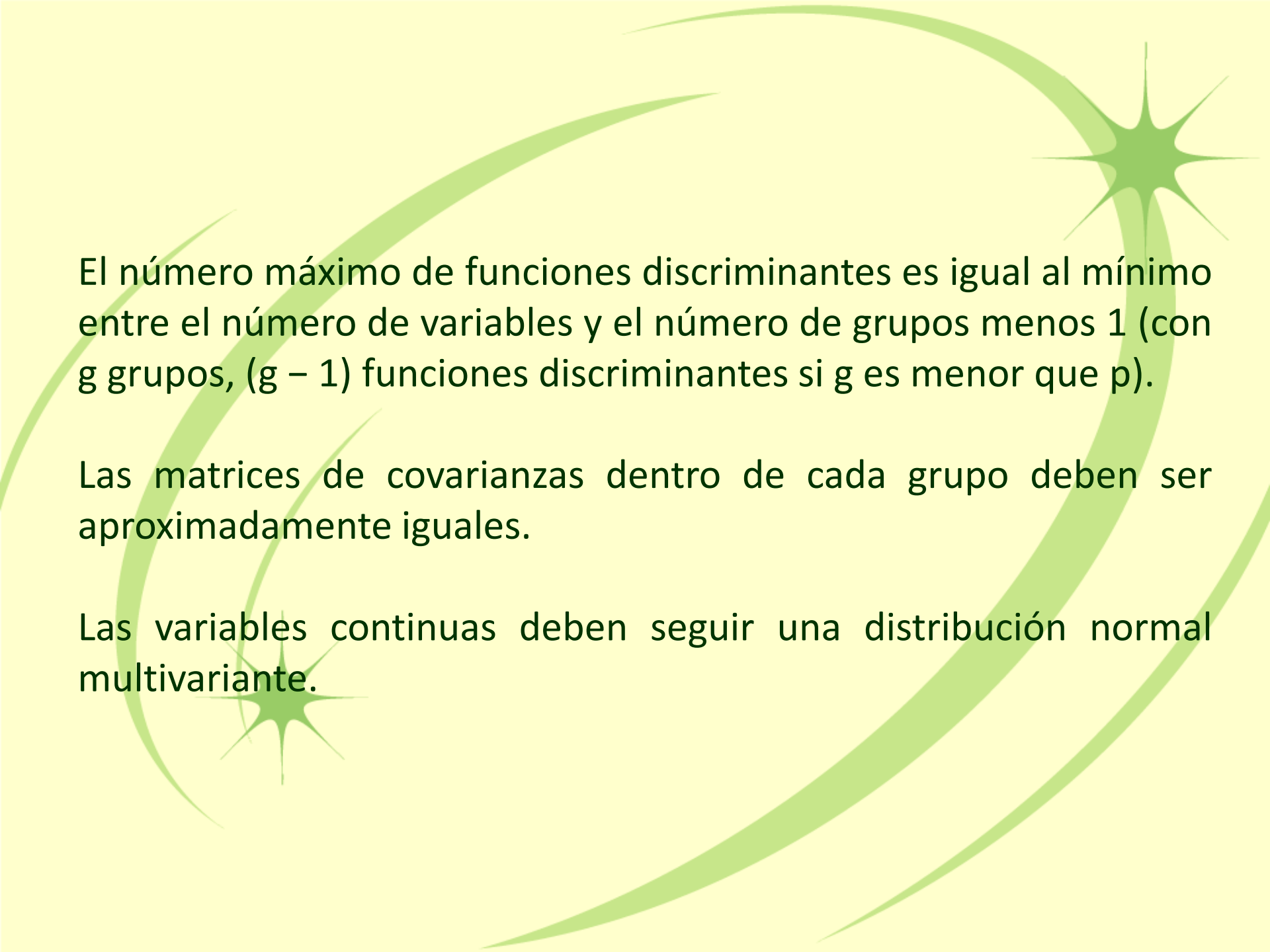
Es necesario considerar una serie de restricciones o supuestos:

Se tiene una variable categórica y el resto de variables son de intervalo o de razón.

Es necesario que existan al menos dos grupos, y para cada grupo se necesitan dos o más casos.

El número de variables discriminantes debe ser menor que el número de individuos menos 2:  $X_1, \dots, X_p$ , donde  $p < (n - 2)$  y  $n$  es el número de individuos.

Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.



El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 (con  $g$  grupos,  $(g - 1)$  funciones discriminantes si  $g$  es menor que  $p$ ).

Las matrices de covarianzas dentro de cada grupo deben ser aproximadamente iguales.

Las variables continuas deben seguir una distribución normal multivariante.



# Modelo matemático

Partiendo de  $g$  grupos donde se asignan una serie de individuos y de  $p$  variables medidas sobre ellos ( $X_1, \dots, X_p$ ), se trata de obtener una serie de nuevas variables que indican el grupo al que pertenecen ( $y_1, \dots, y_m$ ), de modo que sean funciones lineales de  $X_1, \dots, X_p$

$$y_1 = a_{11}X_1 + \dots + a_{1p}X_p + a_{10}$$

...

$$y_m = a_{m1}X_1 + \dots + a_{mp}X_p + a_{m0}$$

donde  $m = \text{mínimo}(g - 1, p)$ , tales que discriminen o separen lo máximo posible a los  $g$  grupos. Estas combinaciones lineales de las  $p$  variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos.

# Descomposición de la varianza

La variabilidad total de la muestra se puede descomponer en variabilidad dentro de los grupos y entre los grupos.

$$Cov(X_j, X_{j'}) = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})$$

Se puede considerar la media de cada variable  $X_j$  en cada uno de los grupos  $1, \dots, g$ .

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}$$

Sumando y restando las medias de cada grupo en cada término de la covarianza:

$$Cov(X_j, X_{j'}) = \frac{1}{n} \sum_{k=1}^g \sum_{i=1}^{n_k} (x_{ij} - \bar{x}_{kj} + \bar{x}_{kj} - \bar{x}_j)(x_{ij'} - \bar{x}_{kj'} + \bar{x}_{kj'} - \bar{x}_{j'})$$

grupo	año	denomina	grado	avol	atot	acfi	ph	folin	somers	srv	procian	acrg	acse	achplc	ic	ic2	tono	iim	eq1	vla
RD86	1986	RIBERA	12,8	1,2	6,7	5,2	3,7	2827	50,8	811	3794	386	287	181	7,81	8,95	0,72	18,4	0,49	0,21
RD86	1986	RIBERA	12,8	0,75	6,9	6	3,5	1818	37,8	968	1736	144	141	69	4,88	5,55	0,76	23,6	0,48	0,56
RD86	1986	RIBERA	12,5	1	7,2	6	3,6	1459	35,1	866	2306	225	132	78	5,52	6,35	0,46	36,8	0,6	0,38
RD86	1986	RIBERA	11,9	0,7	7,7	6,8	3,3	2054	32,1	978	3420	204	110	84	4,64	5,15	0,68	36,4	0,42	0,29
RD86	1986	RIBERA	12,5	0,95	7,7	6,3	3,6	2930	49,6	1128	3158	214	148	75	6,99	7,87	0,67	34,2	0,45	0,36
RD86	1986	RIBERA	12,1	0,5	5,8	5,2	3,2	1906	30,6	875	2931	167	95	74	3,98	4,36	0,72	38,1	0,43	0,3
RD86	1986	RIBERA	12,2	0,8	5,9	4,9	3,4	2071	35,6	754	3101	252	160	101	7,6	8,84	0,72	28,5	0,5	0,24
RD86	1986	RIBERA	12,6	0,4	5,4	4,9	3,3	1996	30,6	1202	3001	315	124	101	6,15	7,11	0,74	27,7	0,57	0,4
RD86	1986	RIBERA	13	0,4	4,6	4,1	3,6	2687	41,7	1354	4785	293	170	137	6,6	7,85	0,93	21,6	0,56	0,28
RD86	1986	RIBERA	12,4	0,35	5,5	5	3,3	2468	30	739	2800	152	67	56	5,49	6,23	0,75	30,3	0,69	0,26
RD86	1986	RIBERA	12,6	0,35	5,6	5,2	3,3	1965	30,4	839	2878	144	78	61	5,25	5,96	0,76	33,6	0,58	0,29
RD86	1986	RIBERA	12,2	0,4	4,8	4,4	3,5	1731	28,4	770	2649	193	107	75	5,86	6,65	0,77	25,6	0,61	0,29
RD86	1986	RIBERA	12,8	0,4	5,2	4,7	3,5	1827	29	960	2746	179	101	65	5,77	6,65	0,75	29,2	0,61	0,35
RD86	1986	RIBERA	13	0,6	5,3	4,5	3,6	2110	38,9	489	2587	358	238	172	6,29	6,97	0,68	22,8	0,41	0,19
RD87	1987	RIBERA	12,3	0,4	4,5	4,1	3,7	925	20,8	615	1792	231	177	168	1,95	2,11	0,89	22,4	0,23	0,34
RD87	1987	RIBERA	11,3	0,3	4,7	4,3	3,7	1637	30	881	2386	254	189	154	3,68	4,19	0,86	21,3	0,27	0,37
RD87	1987	RIBERA	10,8	0,3	4,8	4,4	3,7	1103	21,9	658	1906	158	120	89	2,2	2,44	0,88	20,7	0,32	0,35
RD87	1987	RIBERA	11,2	0,4	4,4	3,9	3,7	1528	28,2	735	2188	275	203	146	4,05	4,57	0,67	17,8	0,38	0,34
RD87	1987	RIBERA	11,9	0,35	4,3	3,9	3,6	1468	24,8	666	1889	222	169	132	3,31	3,74	0,71	16,9	0,38	0,35
RD87	1987	RIBERA	11,5	0,4	4,4	3,9	3,6	1462	25,7	612	1877	236	178	130	3,98	4,46	0,67	19,2	0,37	0,33
RD87	1987	RIBERA	11,7	0,25	5,7	5,4	3,7	1887	33,7	830	2634	290	223	186	3,17	3,53	0,7	22,8	0,26	0,32
RD87	1987	RIBERA	11,6	0,65	4,5	3,7	3,8	2071	32,1	906	2598	242	187	146	2,63	2,99	0,74	20,6	0,28	0,35
RD87	1987	RIBERA	12,4	0,4	4,8	4,3	3,7	2074	35,9	770	2702	220	167	101	5,34	6,01	0,63	26,1	0,44	0,29
RD87	1987	RIBERA	11,4	0,55	4,6	3,9	3,7	1481	27,5	726	2046	273	186	154	2,09	2,32	0,69	19,7	0,25	0,35
RD87	1987	RIBERA	12	0,5	4,8	4,2	3,6	1359	22	543	1771	184	111	87	3,46	3,96	0,64	24,7	0,38	0,31
RD87	1987	RIBERA	12,7	0,35	4,3	3,9	3,6	1981	28,6	853	2506	270	212	156	3,09	3,42	0,72	19,5	0,27	0,34
RD87	1987	RIBERA	11,9	0,55	4,5	3,8	3,6	1890	29,5	748	2570	309	230	179	2,82	3,04	0,64	22,2	0,27	0,29
RD87	1987	RIBERA	11,9	0,55	4,6	3,9	3,6	1706	32,2	659	2282	248	189	149	2,12	2,23	0,63	21,2	0,22	0,29
RD87	1987	RIBERA	12,7	0,8	5,6	4,6	3,8	1503	24,2	637	2036	384	200	159	3,83	4,31	0,68	18	0,4	0,31
RD87	1987	RIBERA	12,2	0,4	4	3,5	3,8	1778	25,2	670	2871	263	189	139	4,2	4,81	0,69	17,1	0,47	0,23
RD87	1987	RIBERA	12,4	0,6	5,2	4,5	3,5	2000	29,2	756	2298	289	199	160	5,13	5,74	0,61	24,9	0,39	0,33
RD87	1987	RIBERA	12,7	0,5	5,1	4,5	3,6	1444	23,9	606	2086	232	166	115	4,45	4,93	0,61	24,1	0,41	0,29
RD87	1987	RIBERA	12,6	0,35	5,1	4,7	3,6	1706	27,8	652	2509	300	205	163	4,23	4,62	0,58	22,8	0,38	0,26
RD87	1987	RIBERA	11,2	0,35	4,8	4,3	3,6	1515	20	528	1910	258	168	136	3,98	4,34	0,56	22,1	0,43	0,28
T86	1986	TORO	13	1,6	6,8	4,8	3,8	2153	38,4	716	2547	166	122	44	5,34	6,06	0,77	14,7	0,67	0,28
T86	1986	TORO	14	0,4	4,4	3,9	3,4	2481	44,3	1313	3392	288	197	146	5,76	6,64	0,93	16,7	0,56	0,39
T86	1986	TORO	13,2	0,55	4,7	4	3,6	1665	38	834	2956	206	204	82	5,15	5,85	0,79	14,2	0,55	0,28
T86	1986	TORO	13,4	1,1	4,9	3,5	3,6	3152	56,2	1561	5082	244	174	90	7,91	9,12	0,75	21,2	0,63	0,31
T86	1986	TORO	13,2	0,65	4,9	4,1	3,6	3109	54,2	1355	4536	280	189	115	7,88	9,3	0,83	16,1	0,62	0,3
T86	1986	TORO	13,9	0,8	4,6	3,6	3,7	1975	33,9	544	2390	254	182	113	4,42	5,02	0,82	18,6	0,44	0,23
T87	1987	TORO	14	0,6	4,1	3,4	3,4	2334	47,3	1210	3422	354	253	183	6,58	7,74	0,79	17,2	0,5	0,35
T87	1987	TORO	13,9	0,3	4,7	4,3	3,6	1915	39,8	1128	2970	229	188	89	5	5,67	0,72	18,3	0,45	0,38
T87	1987	TORO	12,1	0,7	5,5	4,6	3,4	2171	35,9	1177	3313	280	202	130	6,04	6,81	0,66	26,4	0,4	0,36
T87	1987	TORO	13,9	0,4	3,8	3,3	3,9	2668	47,1	1270	2922	345	268	148	7,07	8,37	0,86	12,9	0,53	0,43
T87	1987	TORO	12,3	0,65	4,8	4	3,6	3071	51,6	1496	4600	388	267	180	7,01	8,26	0,85	23	0,43	0,33

denomina	grado	avol	atot	acfi	ph	folin	somers	srv	procian	acrg	acse	achplc	ic	ic2	tono	iim	eq1	vla
RIBERA	12,17	0,52	5,26	4,61	3,58	1834,32	30,70	787,76	2551,44	246,00	165,47	122,88	4,49	5,07	0,70	24,44	0,42	0,32
TORO	13,35	0,70	4,84	3,95	3,60	2426,73	44,25	1145,82	3466,36	275,82	204,18	120,00	6,20	7,17	0,80	18,12	0,53	0,33
TOTAL	12,46	0,57	5,16	4,45	3,58	1979,13	34,01	875,29	2775,09	253,29	174,93	122,18	4,90	5,58	0,73	22,89	0,44	0,32

# Descomposición de la varianza

$$Cov(X_j, X_{j'}) = \frac{1}{n} \sum_{k=1}^g \sum_{i=1}^{n_k} (x_{ij} - \bar{x}_{kj})(x_{ij'} - \bar{x}_{kj'}) + \sum_{k=1}^g \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'})$$

$$Cov(X_j, X_{j'}) = d(X_j, X_{j'}) + e(X_j, X_{j'})$$

Es decir, la covarianza total es igual a la covarianza dentro de grupos más la covarianza entre grupos.

$$t(X_j, X_{j'}) = d(X_j, X_{j'}) + e(X_j, X_{j'})$$

En notación matricial:

$$T = D + E$$

T = matriz de covarianzas total

E = matriz de covarianzas entre grupos

D = matriz de covarianzas dentro de grupos.

# Funciones Discriminantes

La idea del Análisis Discriminante consiste en obtener a partir de  $p$  variables observadas  $X_1, \dots, X_p$  en  $g$  grupos,  $m$  funciones  $y_1, \dots, y_m$  de forma

$$y_j = a_{j1}X_1 + \dots + a_{jp}X_p + a_{j0}$$

donde  $m = \text{mínimo}(g-1, p)$ , tales que  $\text{Corr}(y_j, y_{j'}) = 0$  para todo  $j \neq j'$ .

Si las variables  $X_1, \dots, X_p$  están tipificadas, entonces las funciones

$$y_j = a_{j1}X_1 + \dots + a_{jp}X_p$$

para  $j=1, \dots, m$ , se denominan funciones discriminantes canónicas.

# Funciones Discriminantes

Las funciones  $y_1, \dots, y_m$  se extraen de modo que

(i)  $y_1$  sea la combinación lineal de  $X_1, \dots, X_p$  que proporciona la mayor discriminación posible entre los grupos.

(ii)  $y_2$  sea la combinación lineal de  $X_1, \dots, X_p$  que proporciona la mayor discriminación posible entre los grupos, después de  $y_1$ , tal que  $\text{Corr}(y_1, y_2) = 0$ .

(iii) ...

# Procedimiento matricial

Se busca una función lineal de  $x_1, \dots, x_p$ :  $y = a'x$ , de modo que

$$Var(y) = a'Ta = a'Ea + a'Da$$

la variabilidad es igual a la variabilidad entre grupos más la variabilidad dentro de grupos.

Se maximiza la variabilidad entre los grupos para discriminarlos mejor y esto equivale a hacer

$$\text{máximo} \left( \frac{a'Ea}{a'Ta} \right)$$

es decir, maximizar la varianza entre grupos en relación al total de la varianza.

# Procedimiento matricial

Se calcula su derivada y se iguala a cero

$$\frac{\partial L}{\partial a} = 2Ea - 2\lambda Ta = 0$$

Se observa que

$$Ea = \lambda Ta \quad y \quad (T^{-1}E)a = \lambda a$$

Por tanto, el autovector asociado a la primera función discriminante lo es de la matriz  $T^{-1}E$ .

Como  $Ea = \lambda Ta$

$$a'Ea = \lambda a'Ta = \lambda$$

Luego si tomo el vector asociado al máximo autovalor, se obtendrá la función que recoge el máximo poder discriminante. El autovalor asociado a la función discriminante indica la proporción de varianza total explicada por la m función discriminante.



# Procedimiento matricial

Si consideramos la función

$$f(a) = \frac{a' E a}{a' T a}$$

Se observa que

$$f(\mu a) = f(a)$$

para todo  $\mu$

Esto implica que obtener el máximo equivale a calcular

$$\text{máximo}(a' E a) \text{ tal que } a' T a = 1$$

Como este es el esquema habitual de los multiplicadores de Lagrange, se define

$$L = a' E a - \lambda(a' T a - 1)$$

## Ejemplo inicial 2 : Confianza en las instituciones de las elites parlamentarias de América Latina

Confianza de los políticos hispanoamericanos en las instituciones y grupos de su país.

En relación a las siguientes personas, grupos e instituciones, me gustaría saber, ¿que grado de confianza, mucha, bastante, poca o ninguna, le merece su actuación en la vida pública de su país?

- El Poder Judicial.
- Los partidos políticos
- Las organizaciones de empresarios
- Los sindicatos
- Las Fuerzas Armadas
- La Iglesia Católica
- El Parlamento
- El presidente de la República
- Los funcionarios
- Los medios de comunicación
- La policía
- El Tribunal Supremo Electoral

-La valoración numérica de las respuestas es : mucha (4), bastante (3), poca (2) o ninguna (1).

## Ejemplo 2 : Confianza en las instituciones de las elites parlamentarias de América Latina

$n = 86$

12 variables



**PDC**  $n_1 = 18$   
**Partido Demócrata Cristiano**

**RN**  $n_2 = 16$   
**Renovación Nacional**

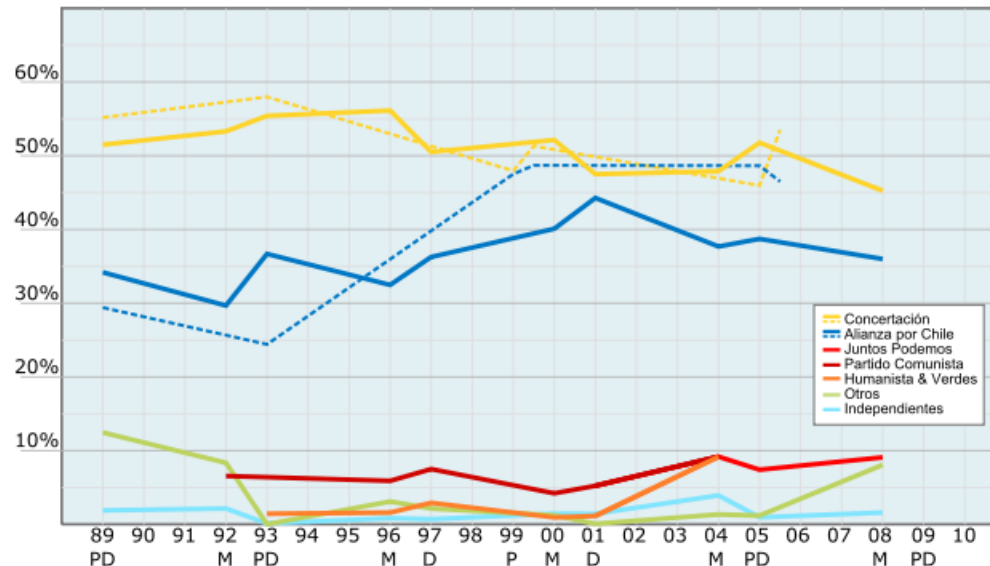
**UDI**  $n_3 = 23$   
**Unión Demócrata Independiente**

**PPD**  $n_4 = 15$   
**Partido por la Democracia**

**PS**  $n_5 = 9$   
**Partido Socialista de Chile**

**PRSD**  $n_6 = 5$   
**Partido Radical Social Demócrata**

Elecciones en Chile (1989-2008)



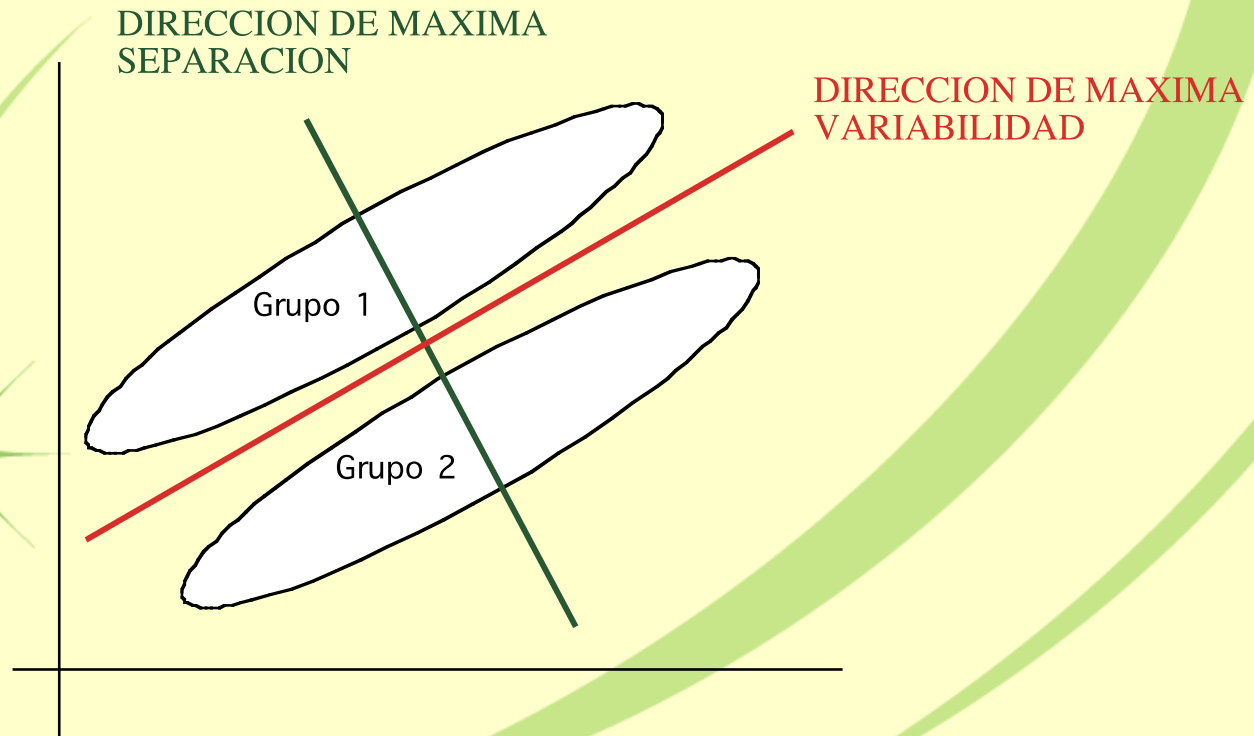
RN + UDI : Alianza por Chile.  
PDC + PPD + PS + PRSD:  
Concertación de partidos por la democracia

# Análisis de la estructura de grupos

## OBJETIVO GENERAL:

ESTUDIAR LAS DIFERENCIAS ENTRE LOS GRUPOS Y CARACTERIZARLAS MEDIANTE TECNICAS MULTIVARIANTES

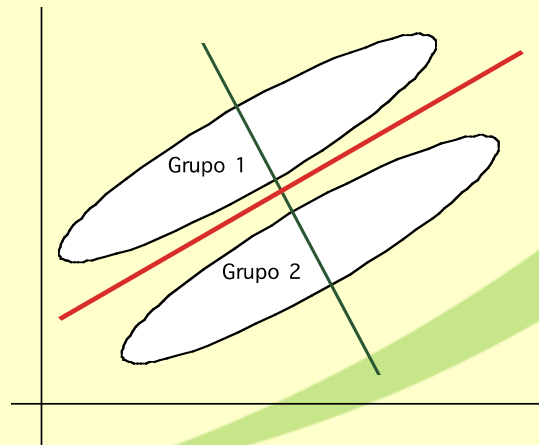
Necesitamos un nuevo grupo de técnicas ya que, en general, las direcciones de máxima variabilidad no coinciden con las direcciones de máxima separación entre grupos



# Análisis de la estructura de grupos

## Objetivos particulares

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)

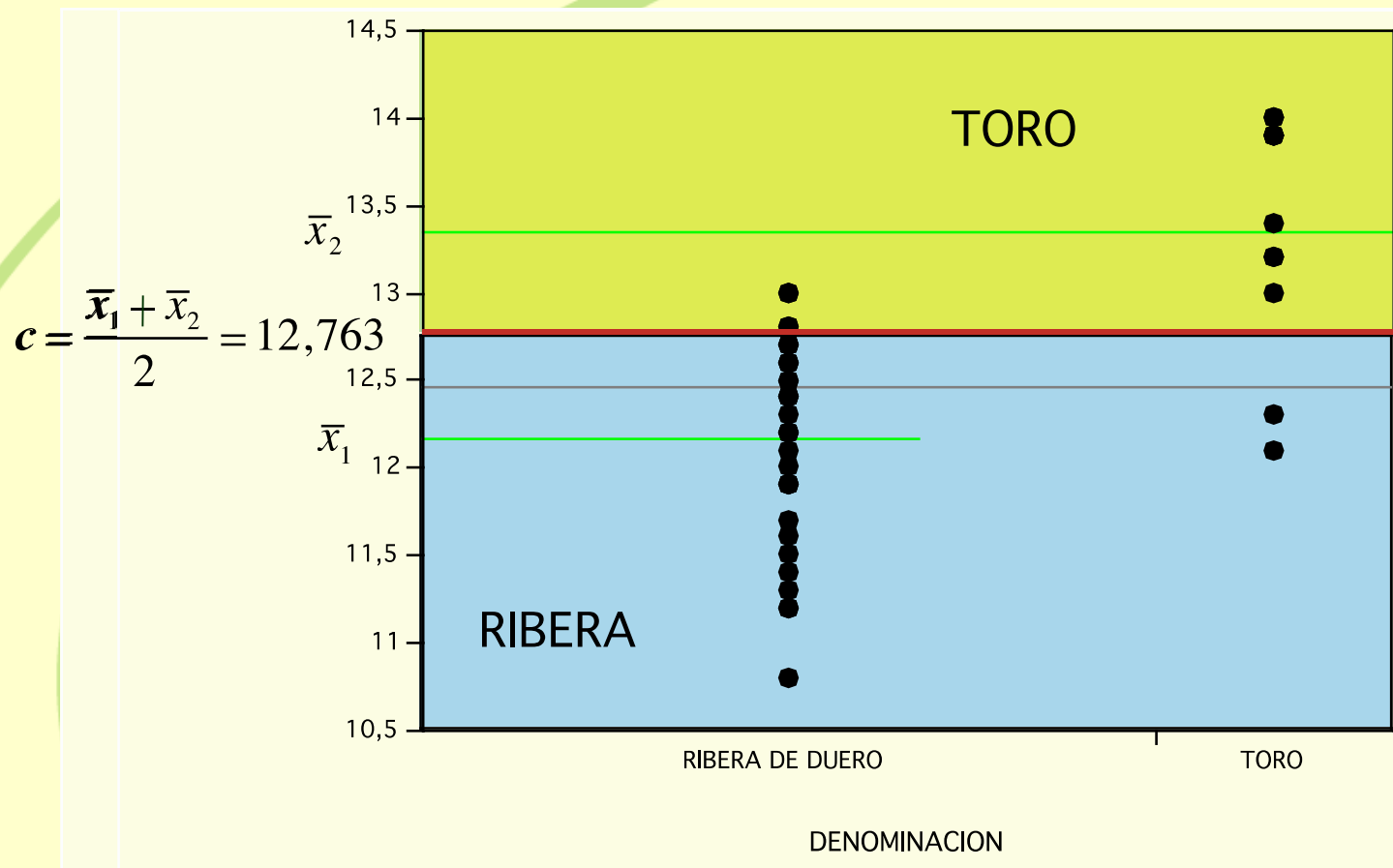


# Análisis Discriminante Lineal

- \* Ejemplo previo con una variable
- \* Análisis para dos grupos
- \* Análisis para varios grupos
- \* Otros métodos discriminantes

# Ejemplo previo con una sola variable

Tomamos como referencia la variable “grado alcohólico” el propósito es clasificar una observación en su denominación de origen a partir de esta variable.



## FRECUENCIAS

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

## PORCENTAJES

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

# Análisis Discriminante para dos grupos (1)

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  $\bar{y}_1 = a' \bar{x}_1$  e  $\bar{y}_2 = a' \bar{x}_2$ .

La diferencia de las medias es, entonces,  $\bar{y}_1 - \bar{y}_2 = a' \bar{x}_1 - a' \bar{x}_2 = a'(\bar{x}_1 - \bar{x}_2)$ .

Se trata, por tanto, hacer máximo  $|a'(\bar{x}_1 - \bar{x}_2)|$  sujeto a la restricción  $a' S a = 1$  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  $a = S^{-1}(\bar{x}_1 - \bar{x}_2)$

y la función discriminante lineal es

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



## Análisis Discriminante para dos grupos (2)

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

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

Y el punto medio de ambos  $\bar{y} = (\bar{y}_1 + \bar{y}_2) = \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$

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

Es decir, asignamos el individuo al grupo 1 si

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

y al grupo 2 si

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

ó bien si tomamos  $W = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$  la regla es, asignar a la población 1 si  $W > 0$  y si no, asignar a la población 2.

# 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 de cada variable no redundante.
- Es Invariante por transformaciones lineales no singulares, en particular por cambios de escala.

# Interpretación Geométrica del Análisis Discriminante

Geométricamente el criterio de clasificación 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

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

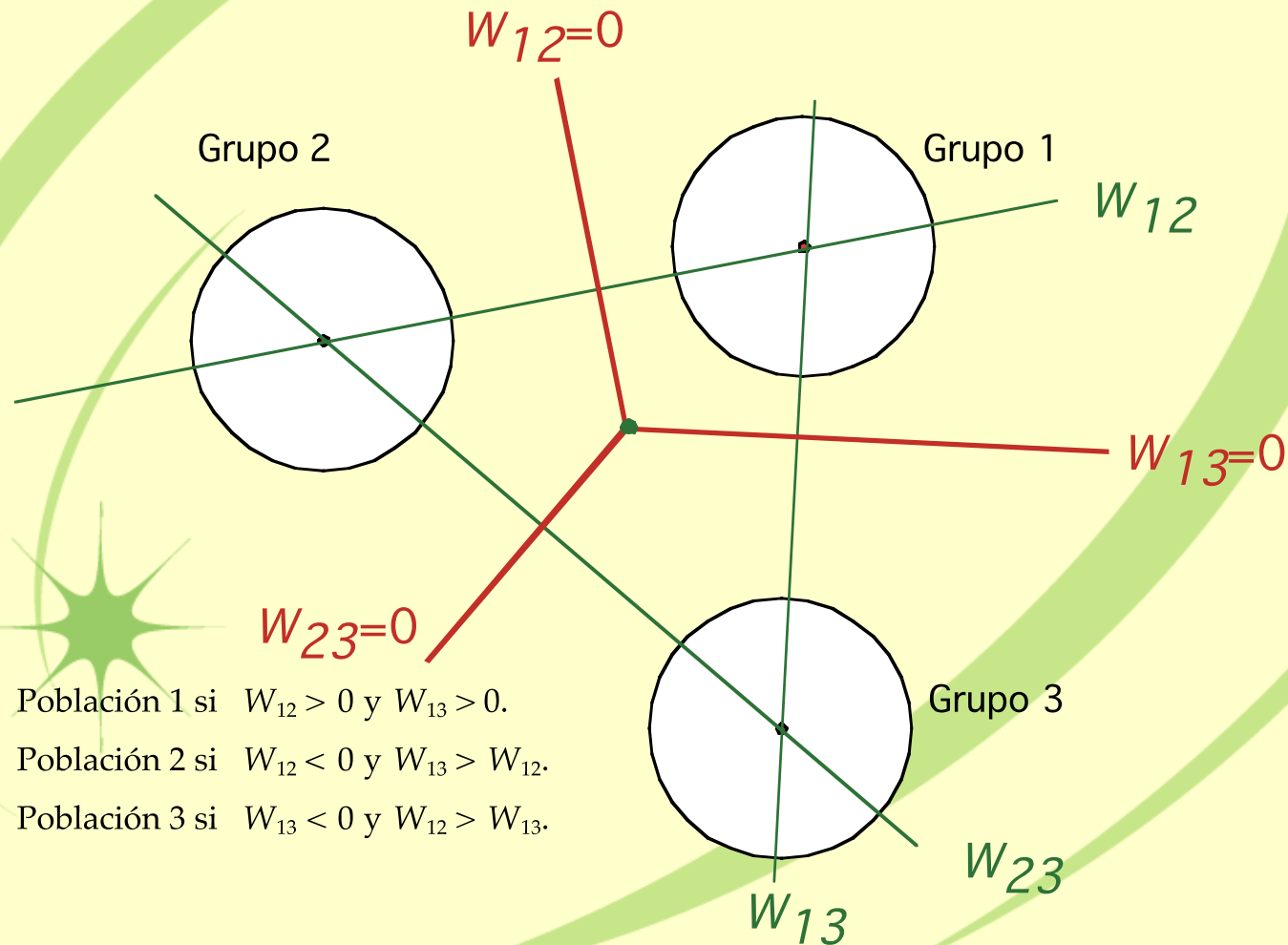
que es idéntica a la regla desarrollada anteriormente.

# Análisis Discriminante para más de dos grupos

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

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

aunque una de ellas es redundante.



# Bondad del ajuste: Probabilidad de clasificación errónea

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.

# Matrices de covarianzas distintas: Discriminante cuadrático

Cuando las matrices de covarianzas no son las mismas en los dos grupos y suponemos que las poblaciones son normales multivariantes, el método de máxima verosimilitud proporciona el siguiente discriminador cuadrático

$$Q(\mathbf{x}) = \frac{1}{2} \mathbf{x}'(\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1})\mathbf{x} + \mathbf{x}'(\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1 - \mathbf{S}_2^{-1}\bar{\mathbf{x}}_2) \\ + \frac{1}{2} \bar{\mathbf{x}}_2' \mathbf{S}_2^{-1} \bar{\mathbf{x}}_2 - \frac{1}{2} \bar{\mathbf{x}}_1' \mathbf{S}_1^{-1} \bar{\mathbf{x}}_1 + \frac{1}{2} \log |\mathbf{S}_2^{-1}| - \frac{1}{2} \log |\mathbf{S}_1^{-1}| > 0$$

# Otras técnicas discriminantes : 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  $\mathbf{x}$  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  $\pi_1$  y  $\pi_2$  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.

## Otras técnicas discriminantes : Discriminante Logístico

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)

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

Naturalmente

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

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 para la que la probabilidad sea mayor, es decir, asignar a la población 1 si  $P(i \in \text{Pob 1} / \mathbf{x}_i) > 0.5$  y a la Población 2 en caso contrario.



# Denominaciones de origen : Resultados Discriminante

	DENOMINACION	
	RIBERA	TORO
GRADO	61,314	63,020
AVOL	-795,610	-749,559
ATOT	550,120	525,438
ACFI	-589,819	-565,147
PH	681,218	652,512
FOLIN	,000	-,005
SOMERS	-6,728	-6,504
SRV	-1,949	-1,779
PROCIAN	,631	,584
ACRG	,366	,298
ACSE	-1,744	-1,454
ACHPLC	2,730	2,483
IC	433,055	427,308
IC2	-369,036	-365,675
TONO	151,250	170,096
IIM	19,369	18,485
EQ1	368,911	378,473
VLA	5299,921	4897,042
(Constante)	-2712,378	-2528,759

Funciones discriminantes lineales de Fisher

Resultados de la clasificación(a)

		DENOMI NACION	Grupo de pertenencia pronosticado		Total
			RIBERA	TORO	
Original	Recuento	RIBERA	34	0	34
		TORO	0	11	11
	%	RIBERA	100,0	,0	100,0
		TORO	,0	100,0	100,0

a Clasificados correctamente el 100,0% de los casos agrupados originales.

Classification Function Coefficients						
	Partido político					
	PDC	RN	UDI	PPD	PS	PRSD
El poder judicial	4,917	5,679	6,429	5,468	3,883	6,274
Los partidos políticos	2,782	2,566	3,104	2,533	2,386	2,866
Las organizaciones de empresarios	3,935	7,064	6,855	5,233	3,253	3,876
Los sindicatos	1,431	1,335	1,223	2,111	2,702	1,835
Las Fuerzas Armadas	2,612	4,262	3,747	2,687	2,813	3,533
La Iglesia Católica	1,94	0,791	1,519	0,949	0,907	-0,483
El Parlamento	5,076	5,41	4,431	4,162	4,288	4,975
El Presidente de la República	5,029	3,061	3,131	5,185	5,98	4,196
Los funcionarios	4,853	2,777	3,196	3,879	3,103	4,527
Los medios de comunicación	2,13	1,912	2,469	2,734	1,521	3,967
La policía	2,235	3,685	3,722	2,258	2,114	4,809
El Tribunal Supremo Electoral	1,13	1,964	1,418	0,7	1,394	2,061
(Constant)	-57,018	-62,966	-64,423	-55,129	-48,218	-70,718
Fisher's linear discriminant functions						

Classification Results(b),(c)

		Predicted Group Membership							
		Partido político	PDC	RN	UDI	PPD	PS	PRSD	Total
Original	Count	PDC	14	0	3	1	0	0	18
		RN	0	11	4	1	0	0	16
		UDI	1	5	13	3	0	0	22
		PPD	3	1	2	8	1	0	15
		PS	2	0	0	0	7	0	9
		PRSD	1	0	1	0	1	2	5
	%	PDC	77,8	0	16,7	5,6	0	0	100
		RN	0	68,8	25	6,2	0	0	100
		UDI	4,5	22,7	59,1	13,6	0	0	100
		PPD	20	6,7	13,3	53,3	6,7	0	100
		PS	22,2	0	0	0	77,8	0	100
		PRSD	20	0	20	0	20	40	100
Cross-validateda	Count	PDC	9	0	3	3	2	1	18
		RN	0	6	8	1	0	1	16
		UDI	1	8	8	3	0	2	22
		PPD	4	1	2	6	1	1	15
		PS	4	0	0	0	5	0	9
		PRSD	1	0	2	0	1	1	5
	%	PDC	50	0	16,7	16,7	11,1	5,6	100
		RN	0	37,5	50	6,2	0	6,2	100
		UDI	4,5	36,4	36,4	13,6	0	9,1	100
		PPD	26,7	6,7	13,3	40	6,7	6,7	100
		PS	44,4	0	0	0	55,6	0	100
		PRSD	20	0	40	0	20	20	100

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 64,7% of original grouped cases correctly classified.

c. 41,2% of cross-validated grouped cases correctly classified.