Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

TALLER DE ESTADÍSTICA PARA LA CIENCIA DE DATOS

R + Python

MÓDULO 4 : ANALISIS MULTIVARIADO I

Análisis Discriminente Lineal

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

TALLER DE ESTADÍSTICA PARA CIENCIA DE DATOS

EXPOSITORES

José Antonio Cárdenas Garro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Predictive Modelling Specialist — Scotiabank

Portfolio and Consumption Analyst — Scotiabank

Data Analyst — ONP (Oficina de Normalización Previsional)

Correo : josecardenasgarro@gmail.com
LinkedIn : www.linkedin.com/in/jos%C3%A9-antonio-c%C3%A1rdenas-garro-599266b0

André Omar Chávez Panduro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Data Scientist — Interbank

Customer Intelligence Analyst — Claro

Data Analyst — INEI (Instituto Nacional de Estadística e Informática)

Correo : andrecp38@gmail.com / 09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andré-chávez-a90078b9

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»

AGENDA

- Introducción
- Análisis Multivariable.
- Regla Discriminante lineal de Fisher.
- El problema general de clasificación para dos poblaciones.
- Clasificación general para g – poblaciones.
- Métodos de Selección de Variables.

¿Qué es el Análisis Multivariable?

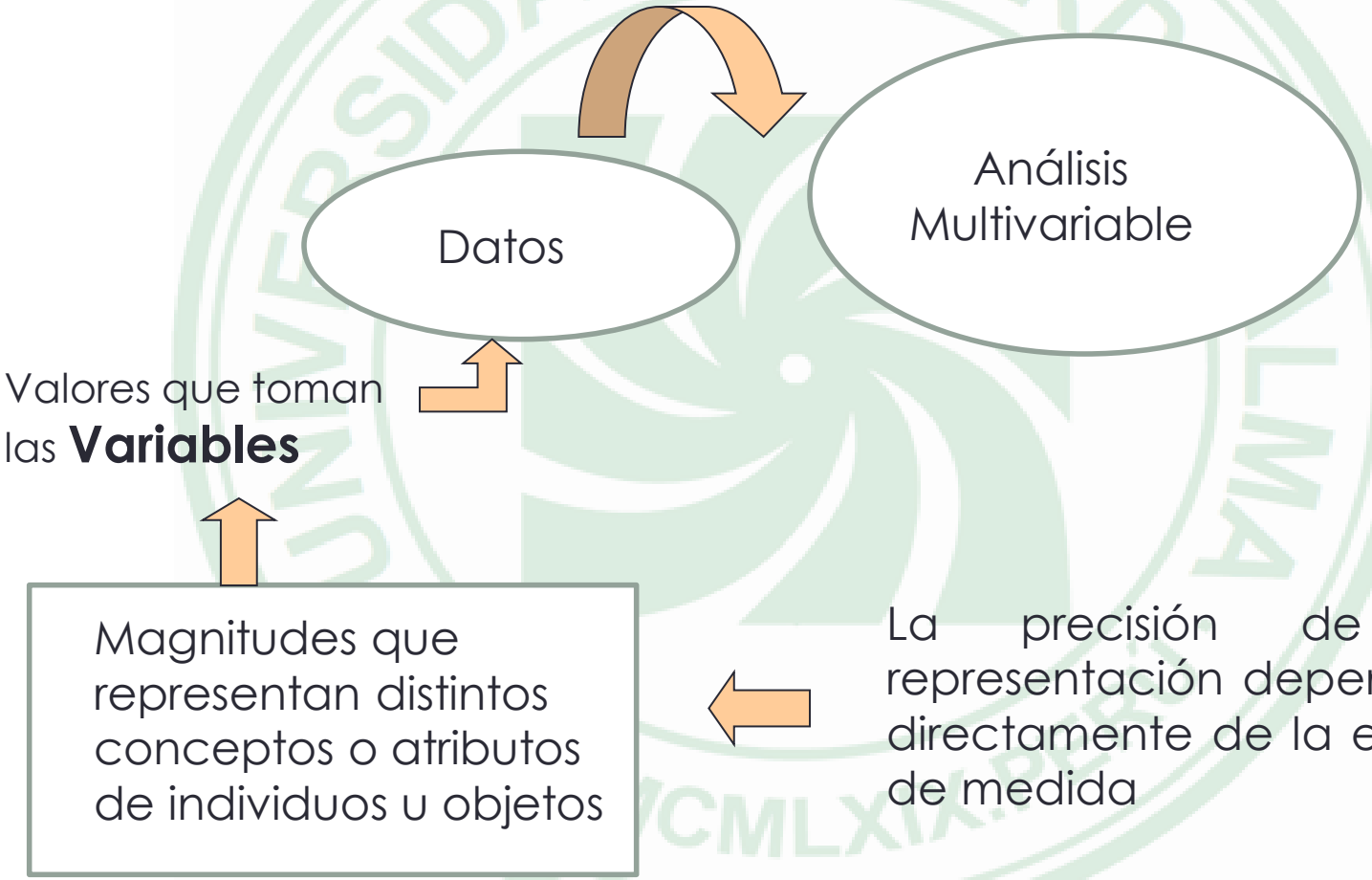
¿Qué es el Análisis Multivariable?

El análisis multivariable puede definirse como el conjunto de métodos o técnicas, diseñados con el fin de maximizar e interpretar la información contenida en un conjunto de variables, sin perder la interacción o grado en que se afectan unas con otras.

- El análisis multivariable permite llevar a cabo la resolución de problemas y la toma de decisiones con un enfoque analítico sobre todas las variables que llegan a influir sobre el o los problemas en cuestión.

Los Datos en el Análisis Multivariable

Variables y escalas de medida



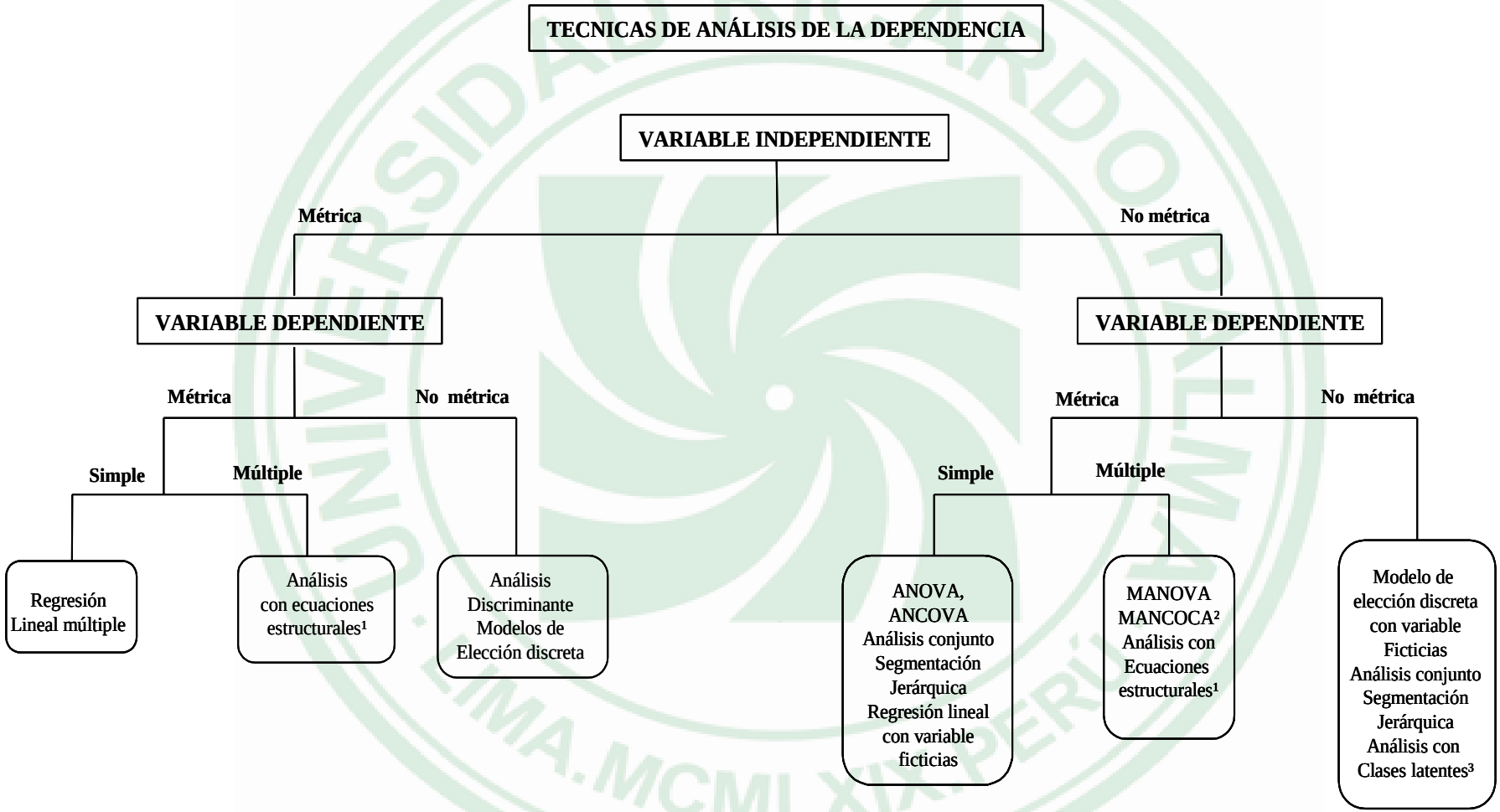
DATOS EN EL ANÁLISIS MULTIVARIABLE

Es fundamental inspeccionar:

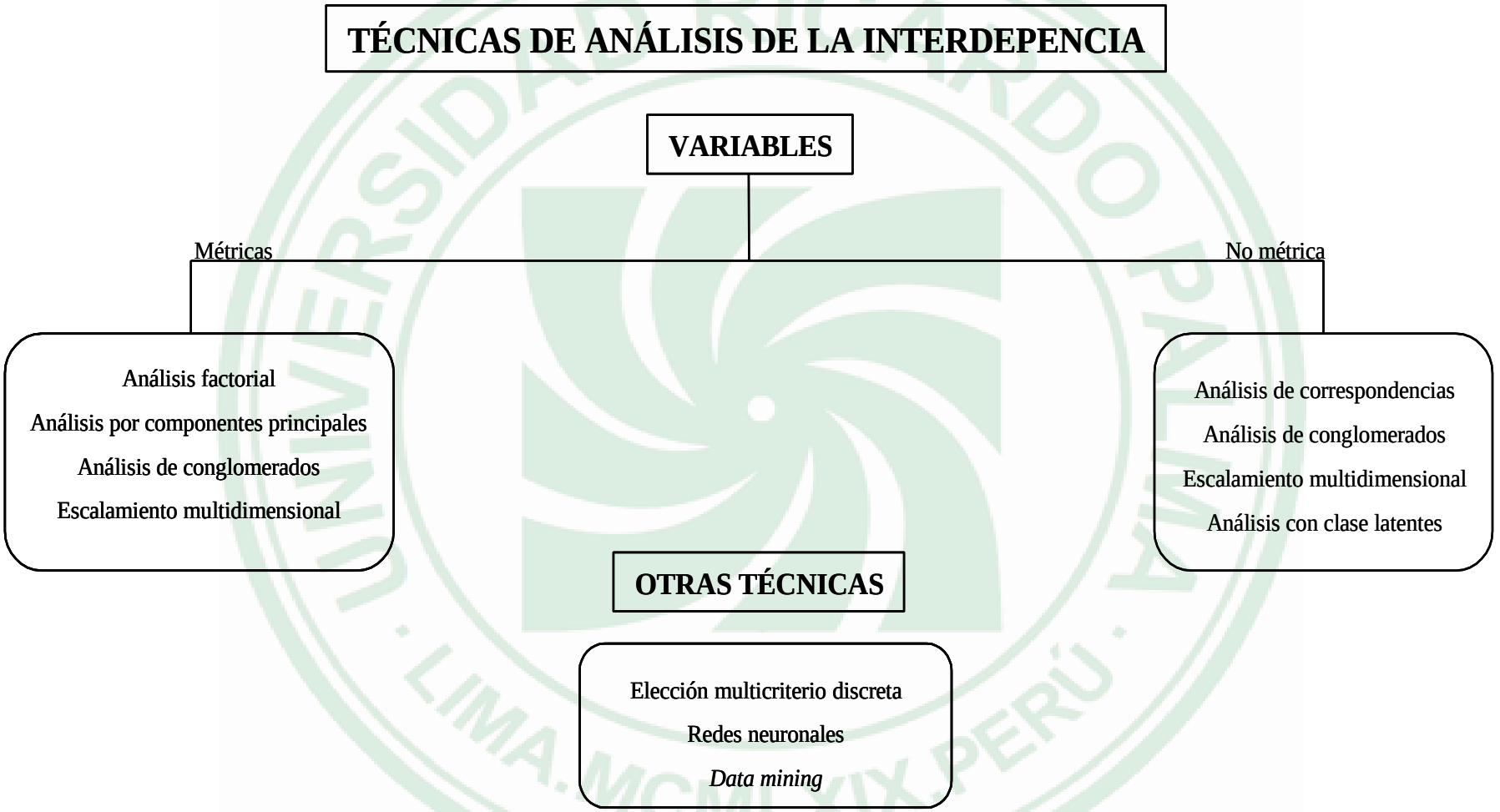
Supuestos subyacentes en los métodos multivariables

- Normalidad de las variables
- Linealidad (existencia de asociaciones lineales entre variables)
- Homocedasticidad (Varianza de los errores es constante)

TÉCNICAS MULTIVARIABLES DE DEPENDENCIA



TÉCNICAS MULTIVARIABLES DE INTERDEPENDENCIA



TÉCNICAS MULTIVARIABLES DE DEPENDENCIA

Técnica	Variable dependiente	Variables independientes
Análisis de la varianza y la covarianza	Métrica	No métricas
Análisis discriminante	No métrica	Métricas
Regresión lineal múltiple ídem con variables ficticias	Métrica Métrica	Métricas No métricas
Modelos de elección discreta ídem con variables ficticias	No métrica No métrica	Métricas No métricas
Análisis conjunto	Métrica o no métrica	No métricas
Segmentación Jerárquica	No métrica o métrica	No métricas
Análisis de ecuaciones estructurales	Métrica	Métricas o no métricas
Análisis con clases latentes	No métrica latente	No métricas observables

TÉCNICAS MULTIVARIABLES DE INTERDEPENDENCIA

Técnica	Variable	Forma grupos de :
Análisis factorial y por componentes principales	Métrica	Variables
Análisis de correspondencias	No métrica	Categorías de variables
Análisis de conglomerados	Métrica y no métrica	Objetos
Escalamiento multidimensional	Métrica y no métrica	Objetos
Análisis con clases latentes	No métricas	Objetos y categorías de variables

ANÁLISIS DISCRIMINANTE

Objetivo:

- Explicar la pertenencia de cada individuo a un grupo (variable categórica) según la variable aleatoria p-dimensional del objeto (variable explicativa).
- Predecir a qué grupo pertenece un individuo nuevo, del que conocemos el valor de la variable p dimensional clasificadora o explicativa.

Puede aplicarse para:

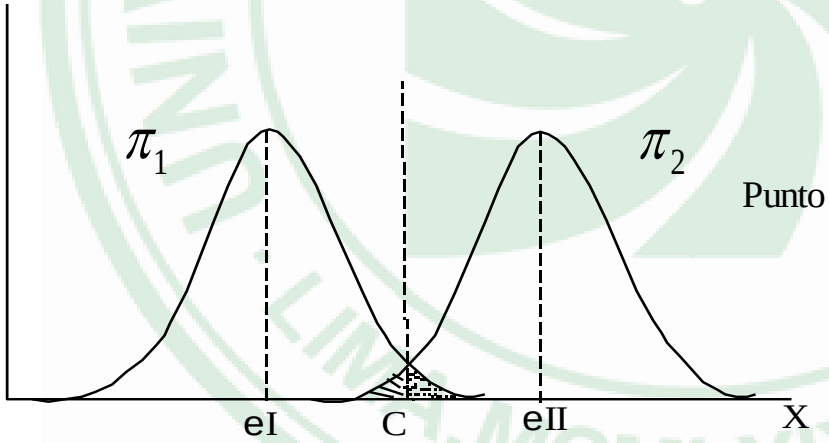
- **Describir:** Explicar la diferencia entre los distintos tipos de objetos.
- **Tomar de decisiones:** Decidir donde clasificar un objeto.

SUPUESTOS

- Se tiene una variable categórica y el resto de variables son de intervalo o de razón y son independientes respecto de ella.
- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.
- Las matrices de covarianzas dentro de cada grupo deben de ser aproximadamente iguales.
- Las variables continuas deben seguir una distribución normal multivariante.

Idea Intuitiva : Clasificación con sólo 2 grupos

- Clasificar a cada individuo en el grupo correcto, según la variable clasificadora. Gráficamente, podríamos representar las hipotéticas funciones de distribución de la variable X para cada uno de los dos grupos. Las distribuciones de frecuencias y la varianza son iguales en los dos grupos, coincidiendo en todo menos en su media.



$$\text{Punto de corte discriminante} : C = \frac{\overline{X}_I + \overline{X}_{II}}{2}$$

Criterios y errores de clasificación

- Los errores de clasificación se encuentran en :

- **Si Xi < C,** se clasifica al individuo i en el grupo I.

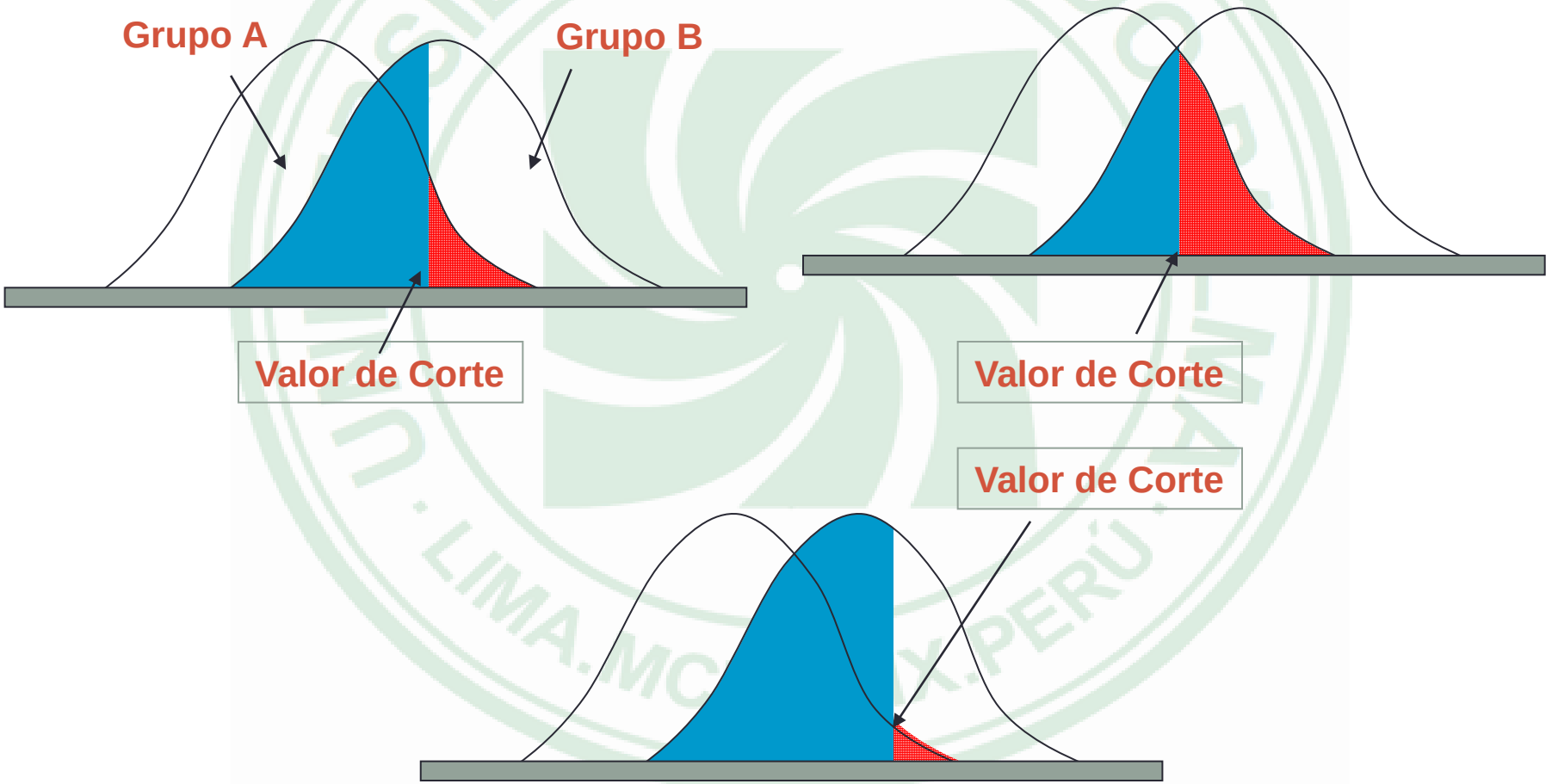
- **Si Xi > C**, se clasifica la individuo i en el grupo II

- Los errores de clasificación se encuentran en :

- **Área a la derecha de C :** Casos del grupo I en los que Xi > C, es decir, son casos del grupo I mal clasificados en el grupo II.

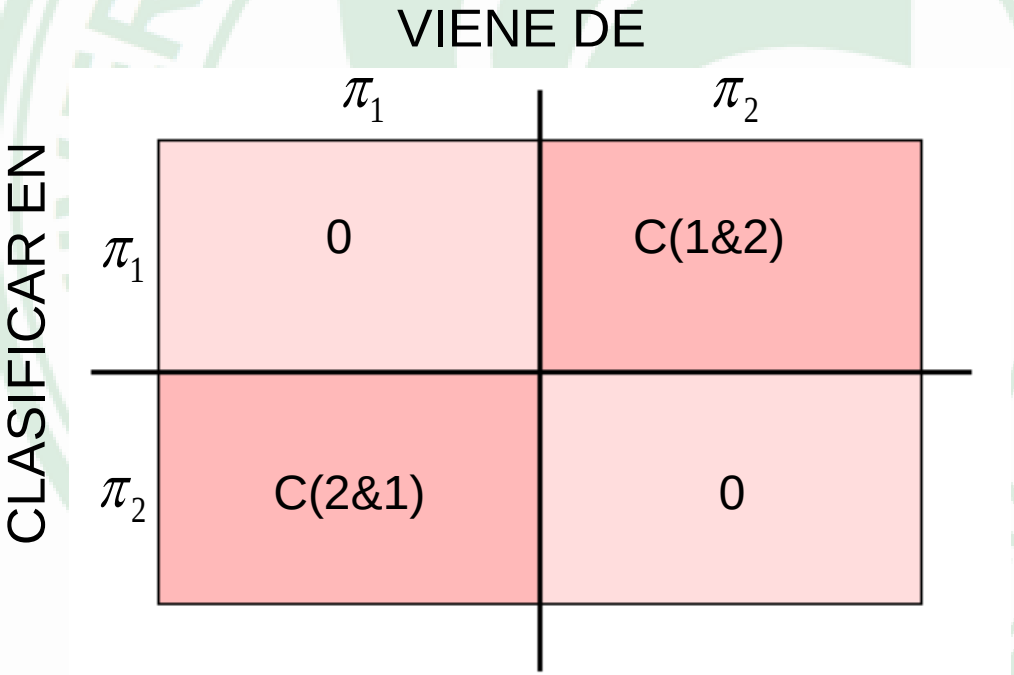
- **Área a la izquierda de C :** Casos del grupo II en los que Xi < C, es decir, son casos del grupo II mal clasificados en el grupo I.

ERROR DE CLASIFICACIÓN: 1 VARIABLE PREDICTORA Y 2 GRUPOS)



EL PROBLEMA GENERAL DE CLASIFICACIÓN PARA DOS POBLACIONES

- El objetivo es encontrar la mejor regla de clasificación, que proporcionará las regiones que minimicen el coste esperado por mala clasificación.



$$CEMC = C(1\&2)\cdot P(1|2)\cdot p_2 + C(2\&1)\cdot P(2|1)\cdot p_1$$

CLASIFICACIÓN CON DOS O MÁS GRUPOS Y DOS O MÁS VARIABLES CLASIFICADORAS

- **Criterio:**

  Buscar el eje que separe lo más posible los centros de los grupos, de forma que los individuos de cada grupo sean lo más homogéneos posibles. Hay que maximizar la dispersión entre grupos respecto a la dispersión dentro de los grupos.

- **Generalizar a K grupos:**

  Habrá más de un eje discriminante. El objetivo es representar a los n individuos de K grupos predefinidos, en un espacio de dimensión reducida (ejes discriminantes) de forma que los grupos proyectados en ese espacio estén bien diferenciados.

Regla discriminante lineal de Fisher

- Sea la variable $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ y dos poblaciones π_1 y π_2.
- Sean $E_{\pi_1}(X) = \mu_1$ y $E_{\pi_2}(X) = \mu_2$

$$V_{\pi_1}(X) = V_{\pi_2}(X) = \Sigma.$$

- Se busca una combinación lineal de la forma

$$Y = l'X = l_1X_1 + l_2X_2 + \cdots + l_pX_p$$

que sea óptima para clasificar una observación en alguna de las dos poblaciones.

REGLA DISCRIMINANTE LINEAL DE FISHER

- Se tiene que

$$E_{\pi_1}(Y) = E_{\pi_1}(l'X) = l'\mu_1 = \mu_{Y1}$$

$$E_{\pi_2}(Y) = E_{\pi_2}(l'X) = l'\mu_2 = \mu_{Y2}$$

$$V_{\pi_1}(Y) = V_{\pi_1}(l'X) = l'\sum l = \sigma_Y^2 = V_{\pi_2}(l'X) = V_{\pi_2}(Y)$$

- Hay que buscar *l* que optimice la separación entre las dos poblaciones: se maximiza la separación entre las medias:

$$\max_{l \in \Re^p} (\mu_{Y1} - \mu_{Y2})^2 = \max_{l \in \Re^p} (l'\mu_1 - l'\mu_2)^2$$

REGLA DISCRIMINANTE LINEAL DE FISHER

Si se maximiza sin restricciones, el máximo puede no ser finito: se maximiza dividiendo por la varianza

$$\max_{l \in \Re^p} \frac{(\mu_{Y1} - \mu_{Y2})^2}{\sigma_Y^2} = \max_{l \in \Re^p} \frac{(l'\mu_1 - l'\mu_2)^2}{\sigma_Y^2}$$

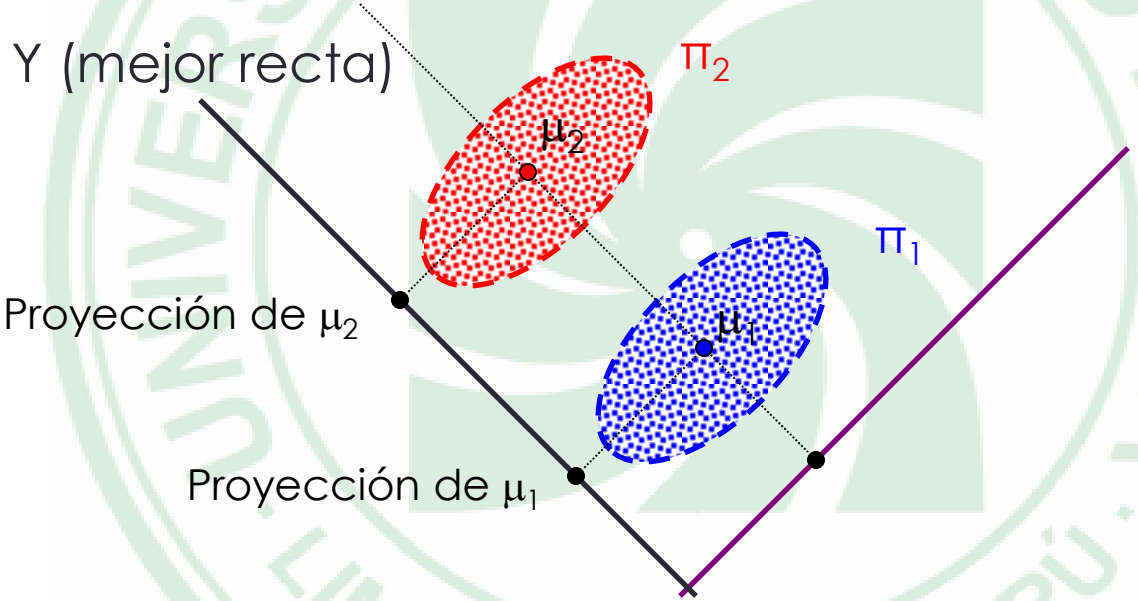
La solución que se obtiene es:

$$Y = (\mu_1 - \mu_2)'\Sigma^{-1} X$$

Función discriminante lineal de Fisher

REGLA DISCRIMINANTE LINEAL DE FISHER

En el caso en que $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ se tiene:



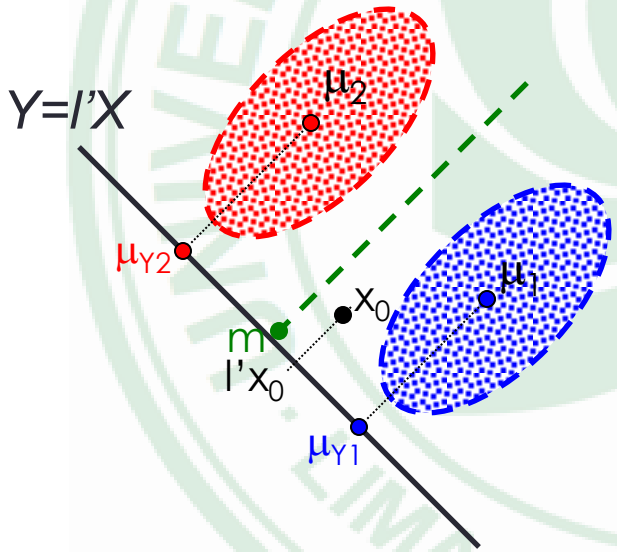
$$Y = l'X = l_1X_1 + l_2X_2$$

l_1 y l_2 determinan la recta

REGLA DISCRIMINANTE LINEAL DE FISHER

El *punto medio* es:

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



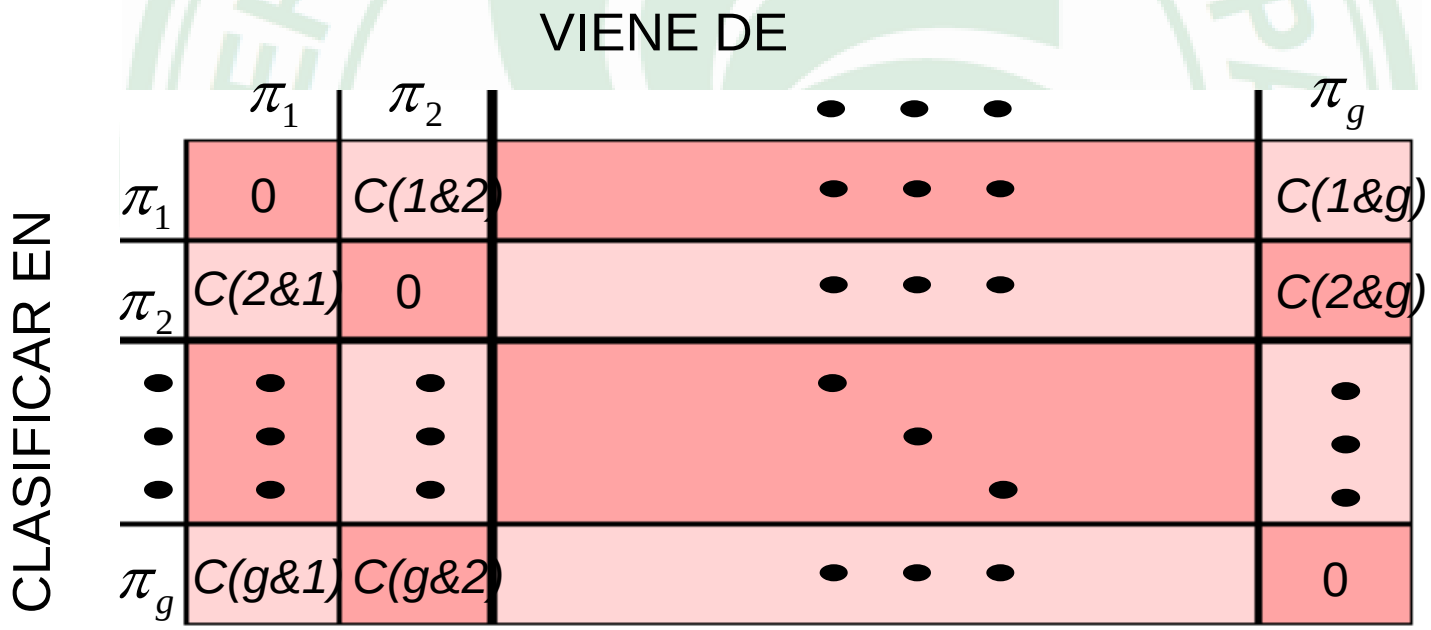
Dada una nueva observación x_0:

- Asignar x_0 a π_1 si $(\mu_1 - \mu_2)'\Sigma^{-1} x_0 - m \geq 0$
- Asignar x_0 a π_2 si

$$(\mu_1 - \mu_2)'\Sigma^{-1} x_0 - m < 0$$

CLASIFICACIÓN GENERAL PARA G POBLACIONES

- El objetivo es encontrar la mejor regla de clasificación, que dará lugar a las regiones que hacen mínimo el *coste por mala clasificación*.



CLASIFICACIÓN GENERAL PARA G POBLACIONES

El coste esperado por mala clasificación dado que la observación viene de es: π_i

$$CEMC(i) = \sum_{\substack{k=1 \\ k \neq i}}^{g} C(k \,\&\, i)\, p(k \mid i)$$

En general, el coste esperado por mala clasificación es:

$$CEMC = \sum_{i=1}^{g} \sum_{\substack{k=1 \\ k \neq i}}^{g} C(k \,\&\, i)\, p_i \int_{R_k} f_i(x)\, dx$$

MÉTODOS ITERATIVOS DE SELECCIÓN DE VARIABLES

TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”

Métodos iterativos de selección de variables

- **Método Forward**

  En cada paso se selecciona la variable que más contribuye a la separación de los grupos. El proceso se detiene si ninguna variable separa los grupos significativamente más de lo que ya estaban.

- **Método Backward**

  Se incluyen todas las variables y en cada paso se elimina la que menos contribuye a la separación de los grupos. El proceso se detiene cuando la exclusión de cualquiera de las variables hace disminuir significativamente la separación entre los grupos.

- **Método Stepwise**

MÉTODOS ITERATIVOS DE SELECCIÓN DE VARIABLES : STEPWISE

Con el landa de Wilks se calcula un estadístico F. Cuanto mayor sea F, más significativa será la variable para la que se calcula. Hay que fijar:

- **F mínimo para entrar**
- **F máximo para salir**

Donde (F to enter > F de salida).

- **Nivel de tolerancia**: Medida del grado de asociación lineal entre las variables clasificadoras.

Si la tolerancia de la variable $1-r_i^2$ es muy pequeña, significa que dicha variable está muy correlacionada con el resto, lo que puede provocar problemas en la estimación. Generalmente, se fija un nivel mínimo de tolerancia.

¡Gracias!