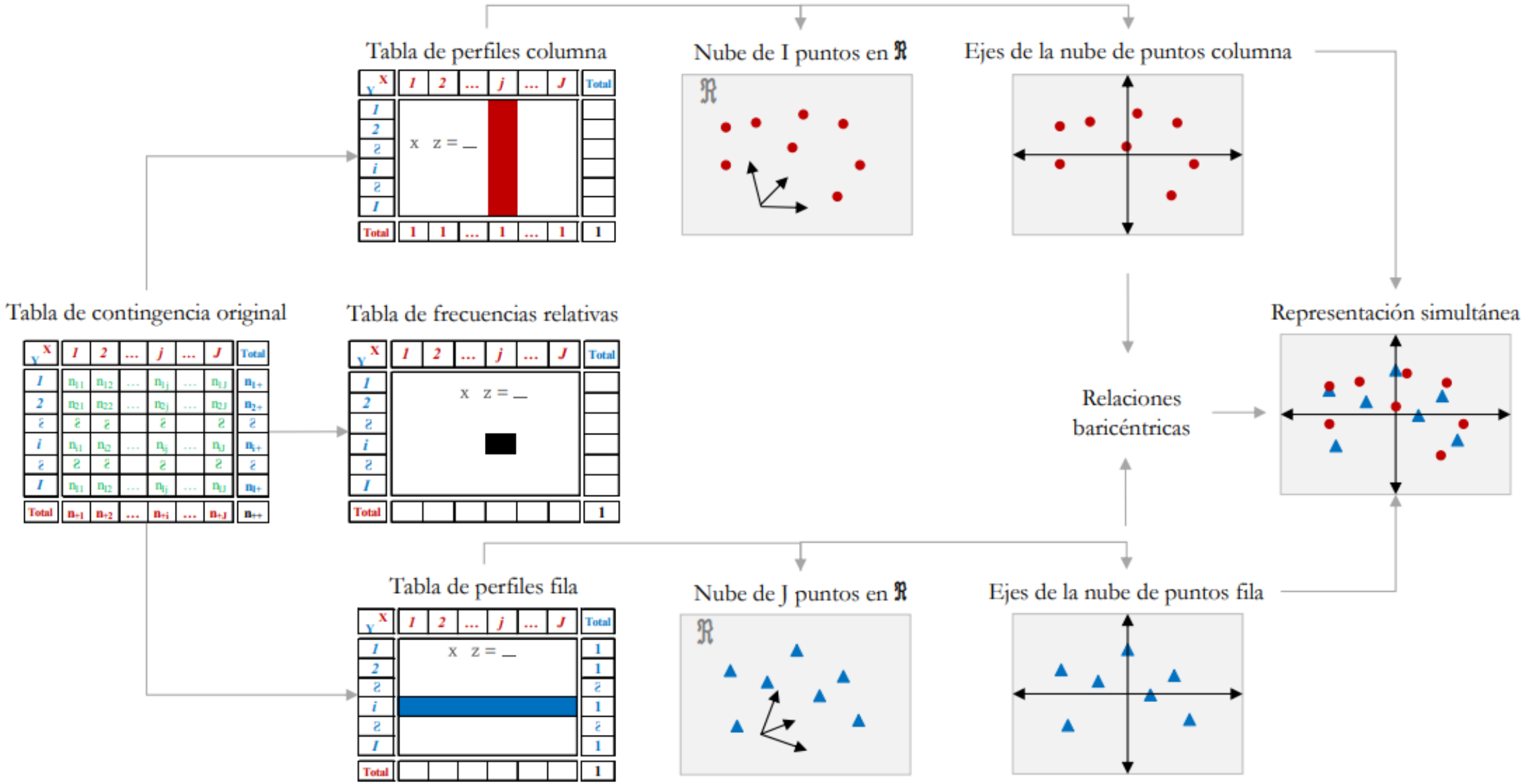
ANÁLISIS DE CORRESPONDENCIA

Dr. Misael Erikson Maguiña Palma

Análisis de Correspondencias Simples

- El **ACS** trata de **analizar, describir y representar** gráficamente la información contenida en una tabla de distribución conjunta de datos dispuestos en filas y columnas: **sus correspondencias** (asociaciones)
- Es una técnica destinada al análisis de la relación de dos variables cualitativas, tratadas como **nominales**
- En general se trata de una tabla de doble entrada de **números positivos:**
  - Tabla de contingencia (conocimiento de la lengua y edad)
  - Casos por variables (ubicación y ocupación por sectores)
  - Matriz de distancias (distancias entre objetos, "municipios")
  - Matrices de transición o tabla de movilidad (origen y destino)
- En ACS, en general, la mayor parte de la información de la tabla se suele expresar en términos de **2 factores**
- En la **representación gráfica** cada categoría o valor de la variable se representa como un punto en el espacio: puntos-fila y puntos-columna Las proximidades geométricas entre puntos-fila y puntos-columna traducen las asociaciones estadísticas entre filas y columnas

Esquema del ACS. Transformación de la tabla de contingencia



Análisis de Correspondencias Simples

- Objetivo del análisis: comparar las filas y las columnas para determinar las correspondencias que se dan entre la diferentes categorías o modalidades

- Procedimiento técnico:

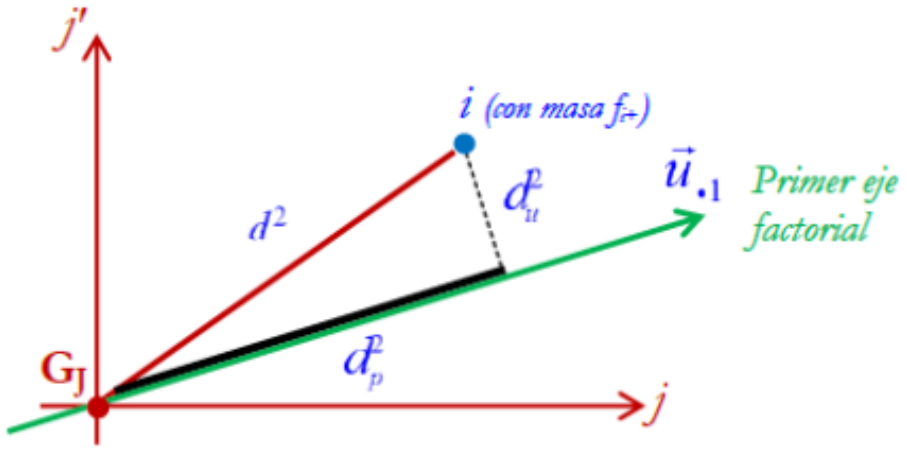
1) Métrica para determinar la proximidad: medida de distancia χ^2

$$d^2(i,i') = \sum_{j=1}^{J} \frac{1}{f_{+j}} \cdot \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

2) La suma de la distancias de cada punto al centro de gravedad es la inercia. La inercia total es $I_G = \sum_{k=1}^{K} \lambda_k$

3) La distancia χ^2 se transforma en euclidiana y se obtiene la Matriz de Inercia (o de Varianzas y Covarianzas)

4) Como en ACP se procede a la Diagonalización: a la obtención de los vectores propios (factores) y valores propios (inercia explicada por los factores)



Análisis de Correspondencias Simples

- Resultados e interpretación

– Vectores propios: son los factores, se extraen un total de $\min\{I, J\} - 1$

– Valores propios: expresan la inercia relativa (la varianza explicada) de cada eje

– Criterios del número de factores a retener

1. Considerar el número de ejes que acumulan en torno al **70%** de la inercia total
2. Representar gráficamente los factores y los valores: Gráfico de sedimentación

   "**Scree test**" (Catell, 1966)
3. Interpretabilidad y pertinencia conceptual de los ejes obtenidos

– La contribución absoluta de cada punto a la inercia explicada por el eje factorial

$$CTA_{ik} = \frac{f_{i+} \cdot y_{ik}^2}{\lambda_k}$$

– La contribución relativa, la correlación entre puntos-fila y ejes factoriales, mide la contribución relativa del factor o eje en la posición de una modalidad, la calidad de su representación

$$CTR_{ik} = \frac{y_{ik}^2}{d^2(i, G_j)} = \cos^2(i, k)$$

– Valores test de significación

– Representación gráfica

Análisis de Correspondencias Simples

- Resultados e interpretación

– Representación gráfica
  - Buscar las categorías con mayor contribución absoluta
  - De estos se distinguen entre los positivos y los negativos para definir las polaridades del eje
  - Se estudia la calidad de la representación de los puntos, las valores más altos de contribución relativa
  - Interrelacionan los ejes para dar cuenta de la estructura de relaciones teniendo en cuenta el orden jerárquico de cada eje
  - Una categoría que coincide con el perfil medio se ubicará en el centro del espacio cercano al origen ("tipo ideal promedio"). Si se aleja difiere de este promedio.
  - Si dos filas (o columnas) tienen perfiles similares se situarán próximos en el espacio.
  - Equivalencia distribucional: las distancias entre dos modalidades no se alteran si se juntan. Criterio de recodificación.
  - Modalidades suplementarias (ilustrativas)

Análisis de Correspondencias Múltiples

- El Análisis de Correspondencias Múltiples (ACM) es la aplicación de la ACS al estudio de tablas lógicas donde se considera un nº cualquiera de variables cualitativas
- Pero con procedimientos de cálculo y reglas de interpretación específicas
- Notación. Consideremos la matriz ***X***:

  n individuos (***i=1...n***)

  p variables cualitativas (***j=1...p***)

  Cada variable x_{+j} tiene c categorías (diferentes según la variable)
  que permiten descomponer la variable en tantas modalidades o categorías

- Codificación disyuntiva completa:

  Si un individuo i tiene en la variable j la categoría $c = c_o$, entonces tendrá:

  - El valor 1 para esta categoría, x_{ijk}, y
  - 0 para el resto de las categorías de la variable, $x_{ijk} = 0$ si $c \neq c_o$

  Se obtiene así la Matriz o Tabla Disyuntiva

Análisis de Correspondencias Múltiples

n: individuos
p: variables cualitativas

	Sexo	Prov	Tipo de salario
1	M	Alajuela	Alto
2	F	Heredia	Bajo
3	F	San Jose	Muy Alto
4	M	Cartago	Bajo
⋮	⋮	⋮	⋮

Tabla **disyuntiva completa**:

	M	F	SJ	A	C	H	MB	B	M	A	MA	...
1	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	
2	**0**	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	
3	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	
4	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	
⋮												

Análisis de Correspondencias Múltiples

La tabla de datos:

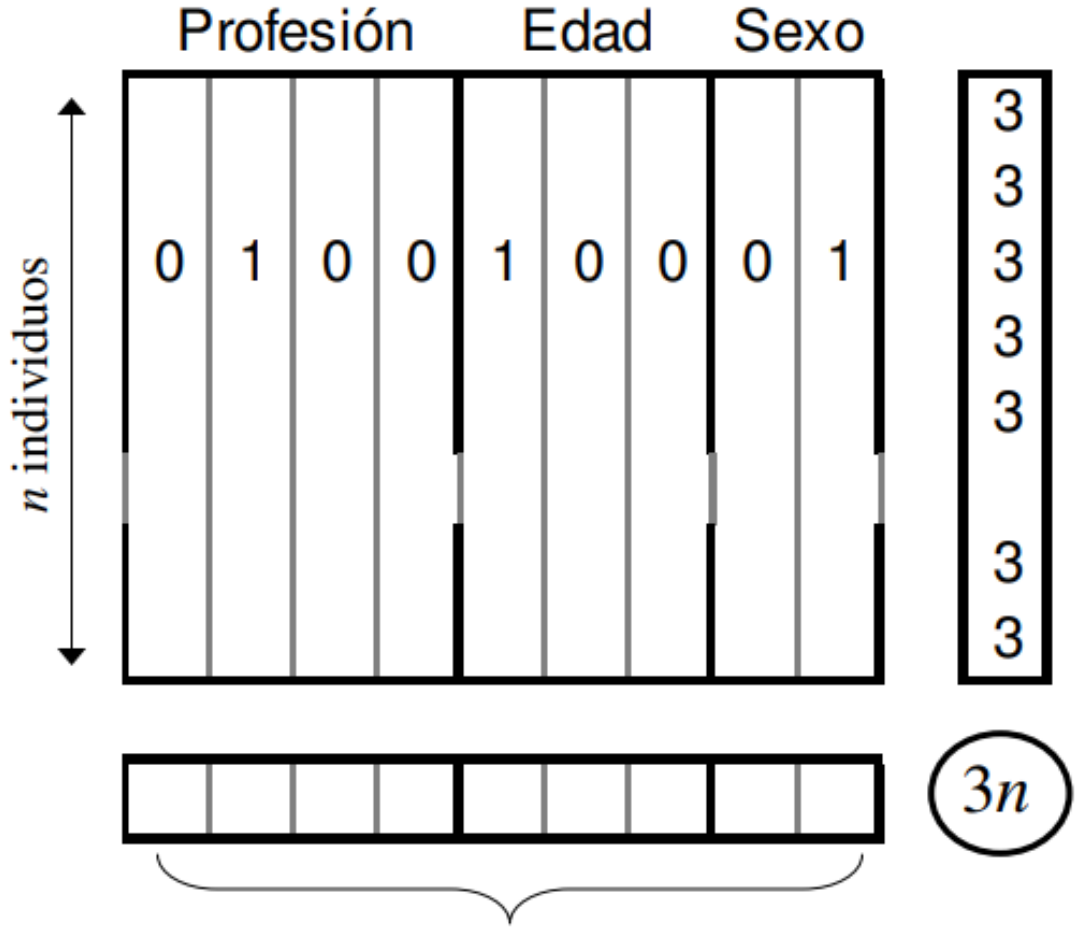
ind	SEXO	EDAD	INGRESO
1	F	5	Medio
2	F	3	Alto
3	M	4	Bajo
4	F	1	Bajo
5	F	2	Medio
6	M	5	Alto
7	F	2	Medio
8	M	3	Bajo
9	M	1	Alto
10	F	4	Medio

Tabla de Burt

		Sexo		Edad					Ingresos		
		F	M	1	2	3	4	5	B	M	A
Sexo	F	6		1	2	1	1	1	1	4	1
	M		4	1	0	1	1	1	2	0	2
Edad	1	1	1	2					1	0	1
	2	2	0		2				0	2	0
	3	1	1			2			1	0	1
	4	1	1				2		1	1	0
	5	1	1					2	0	1	1
Ingresos	B	1	2						3		
	M	4	0							4	
	A	1	2								3
		18	12								

Análisis de Correspondencias Múltiples

ACM $\Leftrightarrow$ Análisis de Correspondencias de una tabla disyuntiva completa



Estructura particular de la tabla

$\Downarrow$

Propiedades particulares del análisis

Análisis de Correspondencias Múltiples

- Matriz o Tabla Disyuntiva D (matriz lógica o binaria) asociada a la matriz de datos original:

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 2 & 1 \\ 1 & 3 & 2 \\ 2 & 1 & 2 \\ 1 & 2 & 1 \end{pmatrix} \qquad D = \begin{pmatrix} 10 & 100 & 01 \\ 01 & 010 & 10 \\ 10 & 001 & 01 \\ 01 & 100 & 01 \\ 10 & 010 & 10 \end{pmatrix}$$

$$B = D'D = \begin{pmatrix} 3 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 2 & 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & 2 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 2 & 0 & 2 & 0 & 2 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 & 3 \end{pmatrix}$$

- La Matriz o Tabla de Contingencia de Burt **B**, $B=D'D$, es la que resulta de todas las posibles tablas de contingencia las ***p*** variables.
- Propiedad que se cumple para la extensión del ACS en ACM:
  - es equivalente un ACS de la tabla de contingencia entre ***Y*** y ***X***
  - que analizar la tabla disyuntiva ***D*** (de ***n*** filas e ***I+J*** columnas)
  - o analizar la tabla de Burt de ***I+J*** filas y ***I+J*** columnas
- A partir de la tabla de Burt, se obtienen los vectores y valores propios diagonalizando la matriz:

$$V = \frac{1}{p} D^{-1} B$$

Análisis de Correspondencias Múltiples

- Propiedades particulares y reglas de interpretación

– Cada categoría es el punto medio de los individuos que la componen, ponderado por el coeficiente$_{ü}$

– La proporción de inercia explicada por los ejes factoriales es débil (pesimista). Es necesaria una fórmula de cálculo de transformación y obtener así los valores propios corregidos:

a) Benzécri (1979) propuso la fórmula:

1) Calcular la inversa del número de variables: $1/p$
2) Seleccionar los valores propios superiores a: $1/p$
3) Calcular los valores propios corregidos con:

$$\lambda_j^C = \left(\frac{p}{p-1}\right)^2 \left(\lambda_j - \frac{1}{p}\right)^2$$

4) Calcular de nuevo la proporción de varianza explicada

b) Greenacre (2008: 187-191, 198-201, 274) añade una propuesta de mejora a partir de eliminar la diagonal de la matriz de Burt, y recalculando la inercia total como:

$$I_T^C = \frac{p}{p-1} \times \left(I(B) - \frac{m-p}{p^2}\right)$$

Análisis de Correspondencias Múltiples

- Propiedades particulares y reglas de interpretación

– La inercia explicada por una categoría es mayor cuanto menos frecuente. En este sentido considerar:
  - Como mínimo el error muestral. En general un mínimo del 5%
  - En SPAD, procedimiento CORMU, permite "ventilar" (de hecho "imputar" el valor medio) las categorías con una frecuencia inferior al 2% (ajustable)
  - En SPAD es posible la selección de modalidades en COREMA (ACM con selección de categorías), se eliminan pero se visualizan como ilustrativas

– La inercia explicada por una variable es mayor cuantas más categorías tenga

– El número de factores o ejes en ACM es: $m-p$

m modalidades o categorías menos p variables

– La suma de los valores propios (la inercia total) es: $\sum_{j=1}^{p} \lambda_j = \frac{m-p}{p}$

– Categorías suplementarias o ilustrativas (papel de "VI", los factores "VD")

– Gráficos factoriales: categorías activas, ilustrativas e individuos

Análisis de Correspondencias Múltiples

ACM: INTERPRETACIÓN

- Proximidad entre individuos en términos de parecido:
  - ✓ Dos individuos se parecen si tienen casi las mismas modalidades:
- Proximidad entre modalidades de variables diferentes en términos de asociación:
  - ✓ Son cercanos puesto que globalmente están presentes en los mismos individuos
- Proximidad entre modalidades de una misma variable en términos de parecido:
  - ✓ Son excluyentes por construcción
  - ✓ Si son cercanas es porque los individuos que las poseen presentan casi el mismo comportamiento en las otras variables

Análisis de Correspondencias Múltiples

Proceso de análisis de un ACM

1. Elección del conjunto de variable original
   - Según un modelo de análisis mas o menos elaborado
   - Comprobación de las condiciones de aplicación.
   - Análisis de tablas de contingencia entre las variables

2. Extracción de los factores
   - Vectores y valores propios (inercia de los ejes)
   - Calculo de los valores propias corregidos
   - Elección del numero de factores a retener

3. Interpretación de los factores
   - Obtención de las coordinadas, contribuciones absolutas, contribuciones relativas y valores test
   - Interpretación de los factores con la ayuda de los gráficos factoriales. Variables activas y suplementarias
   - Evaluación y validación del modelo factorial.

¿Valido?

No

Si

4. Puntuaciones factoriales
   - Valores de los individuos en las variables factoriales
   - Análisis complementarios


Gracias!

- https://emilopezcano.github.io/seminario_urjc_2018/readme.html