

Análisis de correspondencias

Francesc Carmona

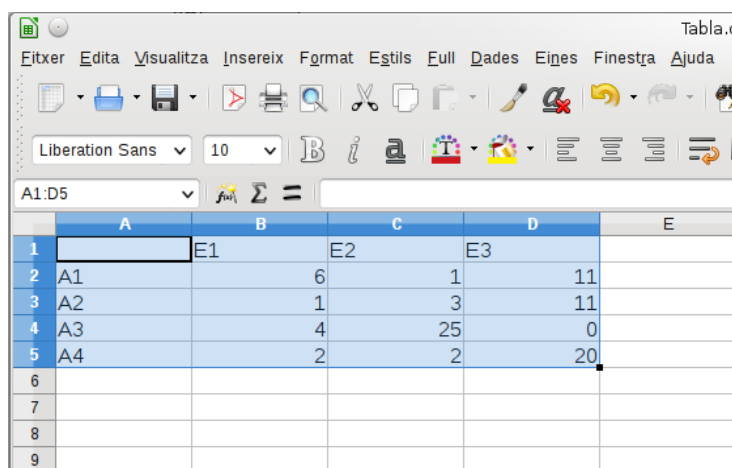
2 de marzo de 2021

1. Sea $\mathbf{N} = (n_{ij})$ la matriz $I \times J$ de frecuencias absolutas observadas de la siguiente tabla

	E1	E2	E3
A1	6	1	11
A2	1	3	11
A3	4	25	0
A4	2	2	20

donde se han contabilizado unos animales en función de su especie E1, E2 y E3 y el área o región A1, A2, A3 y A4, de modo que $I = 4$ y $J = 3$ en este caso.

Escribir la tabla como el objeto `tabla.N` de **R** con sus nombres de filas y de columnas.



	A	B	C	D	E
1		E1	E2	E3	
2	A1	6	1	11	
3	A2	1	3	11	
4	A3	4	25	0	
5	A4	2	2	20	
6					
7					
8					
9					

Figura 1: Tabla en LibreOffice Calc con las frecuencias absolutas.

Para ello podemos entrar los datos en una hoja de cálculo como se puede ver en la figura 1. A continuación seleccionamos las celdas, incluyendo las que contienen las etiquetas, y las copiamos al portapapeles. Entonces con el siguiente código¹ los cargamos en **R**:

```
tabla <- read.table("clipboard")
tabla.N <- as.table(as.matrix(tabla))
```

2. Calcular la matriz de distancias ji-cuadrado entre los perfiles de las filas de la tabla 1 y su inercia total².

Observar que la distancia ji-cuadrado entre perfiles equivale a la distancia euclídea entre los vectores transformados $\mathbf{y}_i = \mathbf{D}_c^{-1/2} \mathbf{p}_i$, es decir, entre las filas de la matriz de datos

$$\mathbf{Y} = \mathbf{P} \mathbf{D}_c^{-1/2} = \mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2}$$

¹En OSX la instrucción es `read.table(pipe("pbpaste"))`

²Para calcular la inercia total podemos utilizar la función `chisq.test()` sobre la tabla de contingencia.

donde \mathbf{P} es la matriz de perfiles de las filas, \mathbf{F} es la matriz de frecuencias relativas, \mathbf{D}_f es la matriz diagonal con los elementos del vector \mathbf{f} con las masas de las filas y \mathbf{D}_c es la matriz diagonal con los elementos del vector \mathbf{c} con las masas de las columnas.

3. Escalado multidimensional

Realizar un MDS con la matriz de distancias ji-cuadrado entre los perfiles de las filas del ejercicio anterior.

Hallar las coordenadas principales para las filas de la matriz de correspondencias del ejercicio 1 que se deducen del MDS.

Repetir el procedimiento con las columnas y realizar la representación simultánea.

4. Análisis de componentes principales

Realizar un PCA con los datos de los ejercicios 1 y 2, mediante la descomposición en valores singulares de la matriz

$$\mathbf{Z} = \mathbf{D}_f^{-1/2}(\mathbf{F} - \mathbf{f}\mathbf{c}')\mathbf{D}_c^{-1/2}$$

Calcular las coordenadas principales y estándares de \mathbf{Z} . Calcular también las inercias principales.

Repetir el análisis con la función `ca()` del paquete `ca` de **R**.

5. La tabla 1 muestra los colores de pelo y de ojos de un gran número de personas.

Tabla 1: Color de pelo y de ojos de una muestra de individuos

Eye color	Hair color				
	Fair	Red	Medium	Dark	Black
Light	688	116	584	188	4
Blue	326	38	241	110	3
Medium	343	84	909	412	26
Dark	98	48	403	681	81

Hallar la solución bidimensional del análisis de correspondencias

- Como un escalado multidimensional de filas y de columnas con la distancia ji-cuadrado.
- Como un análisis de componentes principales sobre la matriz \mathbf{Z} estandarizada.
- Con la función `ca()` del paquete `ca` de **R**.

Un `plot()` del resultado proporciona la representación en dos dimensiones.

6. La tabla `smoke` del paquete `ca` contiene la clasificación de los empleados de una empresa según su nivel profesional (cinco grupos) y sus hábitos fumadores (cuatro grupos).

- Dibujar un mapa óptimo del AC bidimensional y asimétrico, con las filas en coordenadas principales (proyecciones de los perfiles) y las columnas en coordenadas estándares (proyecciones de los vértices).

El mapa asimétrico se puede conseguir con la opción `map="rowprincipal"` de la función `plot()` para un `ca`.

- Dibujar un mapa asimétrico, con las columnas en coordenadas principales y las filas en coordenadas estándares.
- Dibujar un mapa simétrico³ de los datos sobre los hábitos de los fumadores, es decir, representar tanto las filas como las columnas en coordenadas principales.

³Cuando interpretemos mapas simétricos, debemos tener siempre bien presente que un mapa simétrico no es más que el “solapamiento de dos mapas distintos”. Las distancias entre filas y las distancias entre columnas son aproximadamente distancias χ^2 de sus respectivos perfiles. En un mapa simétrico no existe una interpretación específica para las distancias entre las filas y las columnas

7. El AC se utiliza ampliamente para analizar datos en ecología. Los datos del archivo `benthos.xls` que se pueden hallar en la web www.carme-n.org corresponden a los recuentos de 92 especies marinas identificadas en 13 muestras del fondo marino del mar del Norte. La mayor parte de las muestras se obtuvieron cerca de una plataforma petrolífera que producía una cierta contaminación del fondo marino. Existen dos muestras, utilizadas como referencia, supuestamente no contaminadas, que se obtuvieron lejos de la zona de influencia de la plataforma petrolífera.
- a) Calcular la inercia total.
 - b) Representar los datos en un mapa asimétrico con las estaciones de muestreo en coordenadas principales y las especies en coordenadas estándares, es decir, el mapa asimétrico de los perfiles de las muestras (columnas) y de los vértices de las especies (filas).
 - c) Identificar en el mapa las 7 especies más abundantes e interpretar los resultados en cuanto a las muestras y la contaminación.
8. Recordemos los datos de los 24 meses observados por Florence Nightingale que pueden obtenerse en la página

<http://understandinguncertainty.org/node/214>

donde los 12 primeros son antes de aplicar sus nuevos métodos de cuidado en los hospitales militares.

Consideremos las frecuencias de muertes por tres causas: *Zymotic diseases*, *Wounds & injuries* y *All other causes*, junto con la cuarta categoría de soldados en activo que se obtiene al restar los soldados muertos por alguna causa del total.

Con esa tabla de contingencia realizar un análisis de correspondencias completo y valorar e interpretar el resultado.