

# PARTE III. MÉTODOS DE REDUCCIÓN DE LA DIMENSIÓN

## ***TEMA 7. Análisis de Correspondencias***

1. Objetivos del Análisis Factorial de Correspondencias. Aplicaciones
2. Estudio de perfiles fila/columna y concepto de inercia
3. Extracción del espacio factorial
4. Interpretación de los resultados del análisis

## Objetivos del Análisis de Correspondencias. Aplicaciones

- El análisis de correspondencias es una técnica descriptiva para representar **tablas de contingencia**
- Las tablas de contingencia son tablas donde se recogen las frecuencias de aparición de las categorías de dos variables cualitativas en un conjunto de elementos. Las frecuencias son la información de partida
- Si el número de variables involucradas es 2 se llama Análisis de Correspondencias Simple, si son más Múltiple
- Constituye el equivalente del Análisis Factorial para variables cualitativas

**Ejemplo.** Se dispone de datos sobre el consumo de cuatro marcas en tres segmentos de consumidores. *“Técnicas de Análisis de Datos en Investigación de Mercados” Teodoro Luque Martínez. Pirámide 2000*

		segmento			Total
		1	2	3	
marca	A	30	30	155	215
	B	30	130	30	190
	C	80	30	30	140
	D	80	30	5	115
Total		220	220	220	660

- ¿Cuáles son las marcas más parecidas?
- ¿Cómo se consumen las marcas entre los distintos segmentos?
- ¿Hay alguna asociación entre las marcas y los segmentos de consumidores que las compran?

En este caso la mera inspección de las frecuencias de la tabla nos permite afirmar que la marca A es mayoritariamente comprada por los individuos del segmento 3, la marca B por los del segmento 2 y las marcas C y D por los pertenecientes al segmento 1.

No siempre la asociación entre las distintas categorías de filas y columnas resultará tan evidente y además puede resultar complicado si el número de categorías de filas y columnas es muy elevado.

## Objetivos

- Representar las variables en un espacio de dimensión menor (generalmente de dimensión 2), proyectando las filas o columnas de la tabla de forma que estén próximas las que tengan una estructura similar y alejadas las que tengan estructura diferente
- Se pueden proyectar sólo las filas (columnas) en caso de estar interesados en las diferencias entre las categorías de la variable fila (variable columna), o filas y columnas simultáneamente cuando estamos interesados además en la relación entre categorías de filas y columnas
- Asignar valores numéricos (puntuaciones) a las categorías de las variables cualitativas.

## Contraste chi-cuadrado

Para que el ACS tenga sentido, es necesario que las variables tengan algún grado de asociación, es decir, que no sean independientes.

Para determinar si hay o no relación entre las variables se realiza el contraste chi- cuadrado:

$H_0$  : Independencia entre marca consumida y segmento al que se pertenece

$H_1$  : No independencia

**Resumen**

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,584	,340			,620	,620	,032	,155
2	,457	,209			,380	1,000	,038	
Total		,549	362,413	,000 <sup>a</sup>	1,000	1,000		

a. 6 grados de libertad

Rechazamos la hipótesis nula. Hay asociación entre las dos variables: marca y segmento

## Estudio de perfiles fila/columna y concepto de inercia

Perfiles de fila Son los porcentajes de fila y se obtienen dividiendo la frecuencia conjunta de cada celda por la marginal de la fila correspondiente.

*En el ejemplo hay 4 perfiles de fila, uno por marca. Cada perfil de fila puede representarse como un punto en el espacio de dimensión "3" que son las categorías de la variable segmento*

Perfiles de columna Son los porcentajes de columna y se obtienen dividiendo la frecuencia conjunta de cada celda por la marginal de la columna correspondiente.

*En el ejemplo hay 3 perfiles de columna, uno por segmento. Cada uno puede representarse como un punto en el espacio de dimensión "4" que son las categorías de la variable marca*

Tabla de correspondencias

marca	segmento			Margen activo
	1	2	3	
A	30	30	155	215
B	30	130	30	190
C	80	30	30	140
D	80	30	5	115
Margen activo	220	220	220	660

Perfiles de fila

marca	segmento			Margen activo
	1	2	3	
A	,140	,140	,721	1,000
B	,158	,684	,158	1,000
C	,571	,214	,214	1,000
D	,696	,261	,214	1,000
Masa	,333	,333	,333	

$$\frac{155}{215} = 0.721$$

$$\frac{30}{115} = 0.261$$

Perfil fila "medio" (frecuencia marginal de las columnas)  $\frac{220}{660} = 0.333$

### Tabla de correspondencias

marca	segmento			Margen activo
	1	2	3	
A	30	30	155	215
B	30	130	30	190
C	80	30	30	140
D	80	30	5	115
Margen activo	220	220	220	660

### Perfiles de columna

marca	segmento			Masa
	1	2	3	
A	,136	,136	,705	,326
B	,136	,591	,136	,288
C	,364	,136	,136	,212
D	,364	,136	,023	,174
Margen activo	1,000	1,000	1,000	

$$\frac{30}{220} = 0.136$$

$$\frac{5}{220} = 0.023$$

Perfil columna  
 "medio" (frecuencia marginal de las filas)

$$\frac{215}{660} = 0.326$$

$$\frac{190}{660} = 0.288$$

En el caso de ACS las variables son cualitativas y no podemos definir correlaciones ni hablar de varianza total en el mismo sentido que lo hacemos cuando trabajamos con variables cuantitativas. Como concepto análogo al de varianza se define la **INERCIA**: promedio de las distancias de los distintos perfiles al perfil medio o centro de gravedad, ponderado por la frecuencia o masa de cada perfil. La inercia total de la tabla es la misma calculada sobre los perfiles fila que sobre los perfiles columna.

- Si el interés está en la diferencia entre las categorías de la variable fila, se hace el análisis por filas y se representan las filas
- Si el interés está en la diferencia entre las categorías de la variable columna, se hace el análisis por columnas y se representan las columnas
- Si el interés está en la relación entre las dos variables, el método permite representar en el mismo espacio (2 dimensiones), las categorías de filas y de columnas simultáneamente, de forma que las categorías relacionadas aparecerán próximas.

## Cálculo de la inercia de la tabla. Análisis por filas

### Perfiles de fila

marca	segmento			Margen activo
	1	2	3	
A	,140	,140	,721	1,000
B	,158	,684	,158	1,000
C	,571	,214	,214	1,000
D	,696	,261	,043	1,000
Masa	,333	,333	,333	

$$Inercia_A = 0.326 * \left[ \frac{(0.140 - 0.333)^2}{0.333} + \frac{(0.140 - 0.333)^2}{0.333} + \frac{(0.721 - 0.333)^2}{0.333} \right] = 0.220$$

$$Inercia_B = 0.288 * \left[ \frac{(0.158 - 0.333)^2}{0.333} + \frac{(0.684 - 0.333)^2}{0.333} + \frac{(0.158 - 0.333)^2}{0.333} \right] = 0.159$$

$$Inercia_C = 0.212 * \left[ \frac{(0.571 - 0.333)^2}{0.333} + \frac{(0.214 - 0.333)^2}{0.333} + \frac{(0.214 - 0.333)^2}{0.333} \right] = 0.054$$

$$Inercia_D = 0.174 * \left[ \frac{(0.696 - 0.333)^2}{0.333} + \frac{(0.261 - 0.333)^2}{0.333} + \frac{(0.043 - 0.333)^2}{0.333} \right] = 0.115$$

**INERCIA FILAS = 0.548**

### Puntos de fila generales<sup>a</sup>

marca	Masa	Puntuación en dimensión			Inercia	Contribución				
		1	2	Del punto en la inercia de dimensión		De la dimensión en la inercia del punto				
				1		2	1	2	Total	
A	,326	-1,072	,106	,220	,642	,008	,992	,008	1,000	
B	,288	,412	-,998	,159	,084	,628	,179	,821	1,000	
C	,212	,379	,612	,054	,052	,174	,329	,671	1,000	
D	,174	,862	,706	,115	,222	,190	,656	,344	1,000	
Total activo	1,000			,549	1,000	1,000				

a. Normalización simétrica

La inercia total es 0.549 que es nuestro TOTAL A EXPLICAR

## Extracción del espacio factorial

Las dimensiones que definen el nuevo espacio son los autovectores asociados a los mayores autovalores de la matriz de inercia

El número de autovalores no nulos es igual a  $\min(r,s)-1$ , siendo  $r$  el número de categorías de la variable de fila y  $s$  el número de categorías de columna.

La inercia atribuible a cada dimensión vendrá dada por el cuadrado del autovalor correspondiente.

Si consideramos el máximo número de dimensiones conseguiremos explicar el 100% de la inercia. No obstante la consideración de todas las dimensiones nos llevaría a no reducir la dimensión del problema.

## autovalores

Dimensión	Valor	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
	singular				Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	,584	,340			,620	,620	,032	,155
2	,457	,209			,380	1,000	,038	
Total		,549	362,413	,000 <sup>a</sup>	1,000	1,000		

a. 6 grados de libertad

Al ser  $r = 4$  y  $s = 3$ , el número de autovalores no nulos será 2

La primera dimensión asociada al primer autovalor tiene una inercia de 0.34 =  $(0.584)^2$ , que supone un 62% de la inercia total

$$\frac{0.34}{0.549} = 0.62$$

La segunda dimensión tiene una inercia de 0.209, que supone un 38% de la inercia total

Retendremos el número de dimensiones necesarios para acumular un porcentaje considerable de inercia explicada. Habitualmente tendremos más de dos dimensiones y tendremos que decidir con cuantas nos quedamos. Si el número es pequeño la solución será fácil de analizar pero perderemos en calidad de representación (perderemos información). Por el contrario si el número de dimensiones retenidas es grande la calidad de la representación será buena aunque será difícil analizar las similitudes entre las categorías de las variables. Lo ideal sería que la parte de inercia atribuible a las  $k$  primeras dimensiones sea grande con  $k$  pequeño.

SPSS considera por defecto 2 dimensiones.

## Interpretación de los resultados del análisis

- Peso de las categorías en la dimensión**

Para interpretar una dimensión, es necesario estudiar las contribuciones absolutas de las categorías a la dimensión (en la misma línea que las cargas factoriales en el AF). El porcentaje de inercia de una dimensión o factor imputable a cada categoría se denomina contribución absoluta de la categoría a la dimensión. El examen de las mismas nos permite establecer qué categorías tienen más peso en una dimensión. Esta información aparece en SPSS bajo la columna “contribución de puntos para la inercia de la dimensión”.

Puntos de fila generales <sup>a</sup>									
marca	Masa	Puntuación en dimensión		Inercia	Del punto en la inercia de dimensión		Contribución		
		1	2		1	2	De la dimensión en la inercia del punto		Total
							1	2	
A	,326	-1,072	,106	,220	,642	,008	,992	,008	1,000
B	,288	,412	-,998	,159	,084	,628	,179	,821	1,000
C	,212	,379	,612	,054	,052	,174	,329	,671	1,000
D	,174	,862	,706	,115	,222	,190	,656	,344	1,000
Total activo	1,000			,549	1,000	1,000			

a. Normalización simétrica

- **Calidad de representación de cada categoría en el nuevo espacio**

Las contribuciones relativas de las dimensiones a una categoría tienen una interpretación análoga a las comunalidades en el AF e indican la proporción de inercia de cada categoría explicada por la dimensión. La suma de las contribuciones relativas de las dimensiones a cada categoría sería entonces la comunalidad, que interpretamos como una medida de la calidad de la representación de la categoría en los nuevos ejes. Si retuviésemos todas las dimensiones la suma de las contribuciones relativas sería la unidad.

Las contribuciones relativas aparecen bajo “contribución de la dimensión a la inercia del punto”, (cómo cada dimensión explica cada categoría).

marca	Masa	Puntos de fila generales <sup>a</sup>					Contribución		
		Puntuación en dimensión		Inercia	Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
A	,326	-1,072	,106	,220	,642	,008	,992	,008	1,000
B	,288	,412	-,998	,159	,084	,628	,179	,821	1,000
C	,212	,379	,612	,054	,052	,174	,329	,671	1,000
D	,174	,862	,706	,115	,222	,190	,656	,344	1,000
Total activo	1,000			,549	1,000	1,000			

a. Normalización simétrica

## Puntuaciones en la dimensión

Se llama puntuación en la dimensión de cada categoría, a la proyección de cada categoría sobre la dimensión correspondiente.

Aparecen bajo “puntuación en la dimensión”.

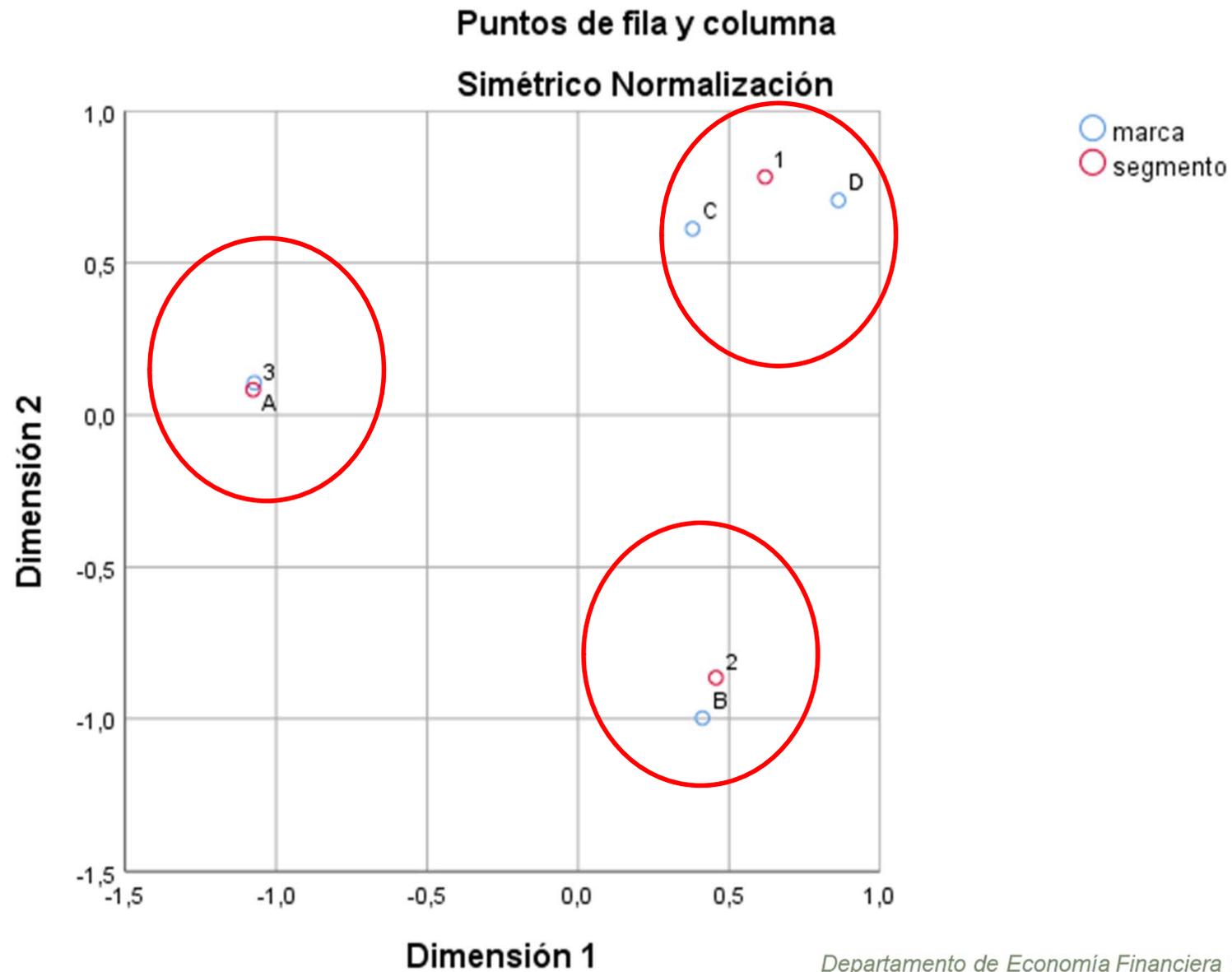
Así la marca A puntúa -1.072 en la dimensión 1 y 0.106 en la dimensión 2.

En principio dos categorías de una variable son similares si puntúan de manera similar en todas las dimensiones (cuidado, siempre que las categorías estén bien representadas).

Puntos de fila generales <sup>a</sup>									
marca	Masa	Puntuación en dimensión		Inercia	Del punto en la inercia de dimensión		Contribución		Total
		1	2		1	2	1	2	
		De la dimensión en la inercia del punto							
A	,326	-1,072	,106	,220	,642	,008	,992	,008	1,000
B	,288	,412	-,998	,159	,084	,628	,179	,821	1,000
C	,212	,379	,612	,054	,052	,174	,329	,671	1,000
D	,174	,862	,706	,115	,222	,190	,656	,344	1,000
Total activo	1,000			,549	1,000	1,000			

a. Normalización simétrica

## Mapa



## Componentes Principales/AF

Pasar de un espacio de  $p$  variables a uno de menor dimensión conservando el máximo posible de la información contenida en los datos

La información contenida en los datos se maneja a través de la matriz de covarianzas o de correlaciones

Las componentes principales se obtienen a partir de los mayores valores propios de la matriz de correlaciones (son las direcciones que retienen mayor cantidad de varianza)

Se calculan los valores de los individuos en las componentes principales (puntuaciones factoriales) y se representan en el nuevo espacio



## Análisis de Correspondencias (filas)

Pasar de un espacio de  $s$  dimensiones (columnas) a uno de menor dimensión (2) conservando el máximo posible de la información contenida en los datos



La información contenida en los datos se maneja a través de la matriz de inercia



Las nuevas dimensiones se obtienen a partir de los mayores valores propios de la matriz de inercia



Se calculan las puntuaciones de cada categoría en las nuevas dimensiones y se representan en el nuevo espacio

## Componentes Principales/AF

Cargas factoriales



## Análisis de Correspondencias (filas)

Contribución absoluta de la categoría a la inercia de la dimensión. (Cómo cada categoría de las filas contribuye a la inercia de la dimensión, o peso de cada categoría en cada dimensión)

*Del punto en la inercia de dimensión*

Comunalidad



Contribución relativa de las dimensiones a una categoría. (Cuanto explica cada dimensión de cada categoría de fila)

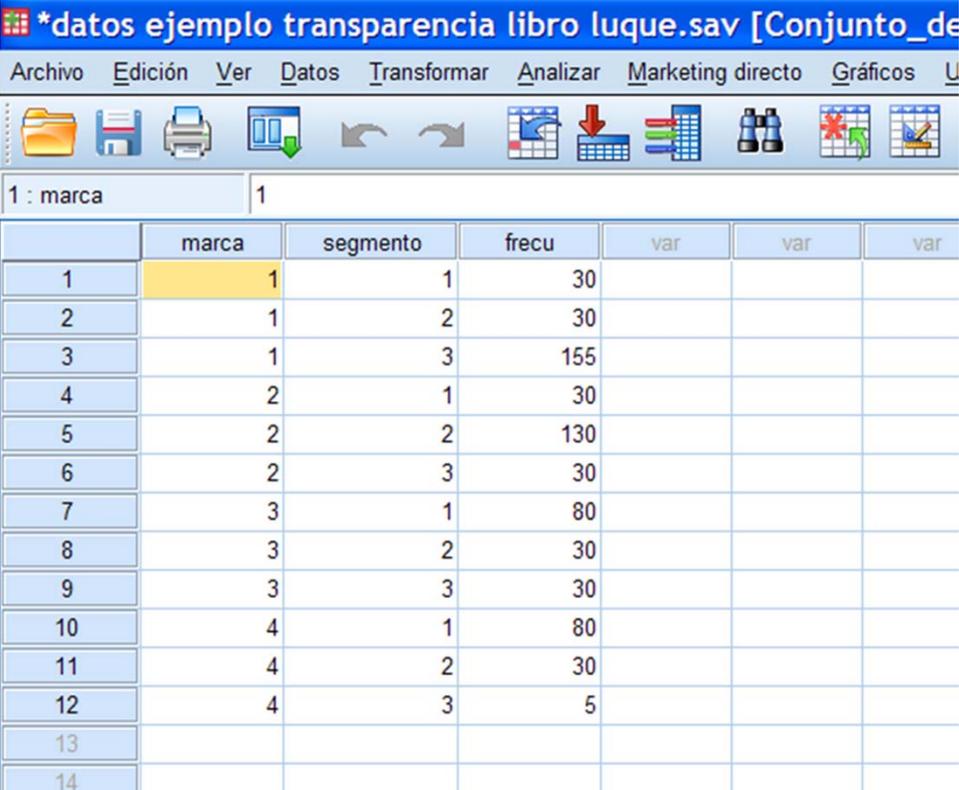
*De la dimensión en la inercia del punto*

Se haría un análisis análogo para las columnas

## GUIÓN ANÁLISIS DE CORRESPONDENCIAS

**Importante: como se introducen los datos en SPSS:**

Las dos variables ocupan las dos primeras columnas y en la tercera deben aparecer las frecuencias absolutas de cada combinación de categorías

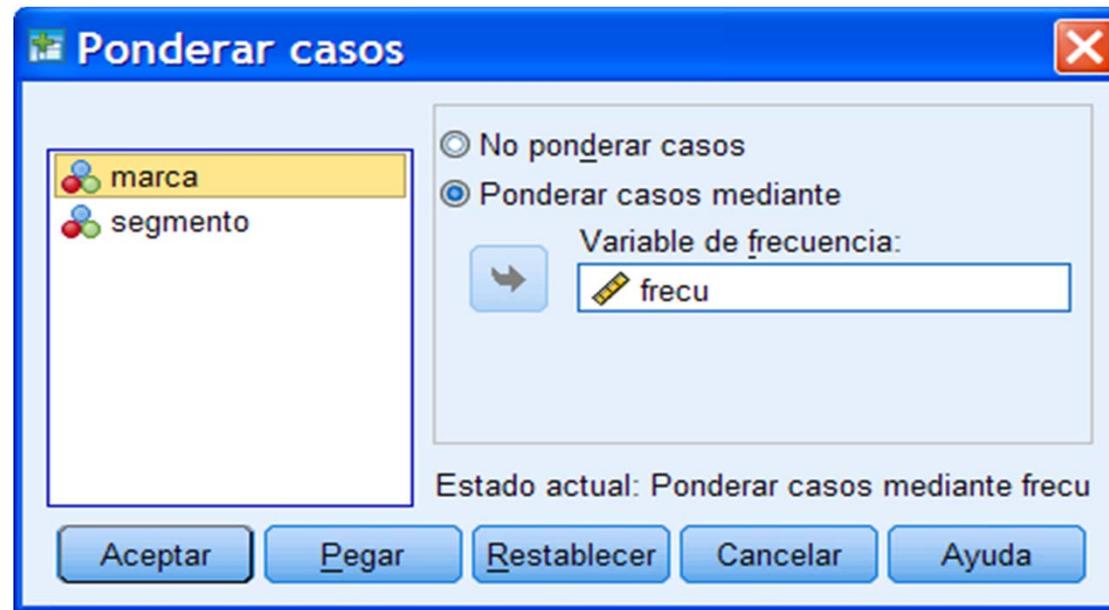


The screenshot shows the SPSS data editor window for a file named '\*datos ejemplo transparencia libro luque.sav'. The window title bar includes the menu options: Archivo, Edición, Ver, Datos, Transformar, Analizar, Marketing directo, Gráficos, and U. The data grid below shows a table with 14 rows and 7 columns. The first column is labeled '1 : marca' and the second column is labeled '1'. The third column is labeled 'frecu'. The data is as follows:

	marca	segmento	frecu	var	var	var
1	1	1	30			
2	1	2	30			
3	1	3	155			
4	2	1	30			
5	2	2	130			
6	2	3	30			
7	3	1	80			
8	3	2	30			
9	3	3	30			
10	4	1	80			
11	4	2	30			
12	4	3	5			
13						
14						

Hay que decirle al programa que los valores de las frecuencias son las apariciones de la pareja de categorías:

Datos > Ponderar casos: ponderar casos mediante variable de frecuencia



## Análisis de Correspondencias

Analizar > Reducción de dimensiones > Análisis de Correspondencias

Hay que especificar el rango de cada variable, SPSS no lo calcula por defecto

**Análisis de correspondencias**

Fila:

Columna:

**Análisis de correspondencias: Def...**

Rango de categorías para la variable de filas: marca

Valor mínimo:

Valor máximo:

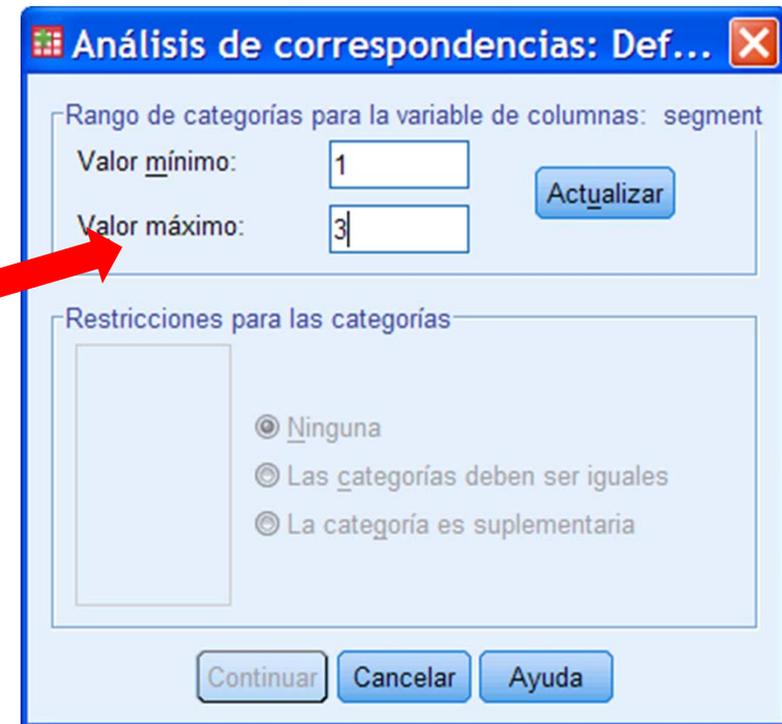
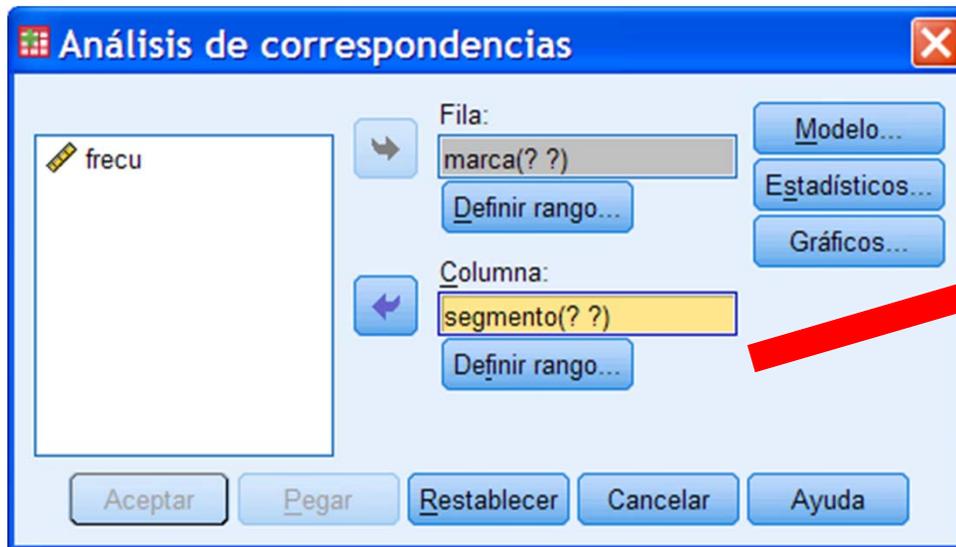
Restricciones para las categorías

Ninguna

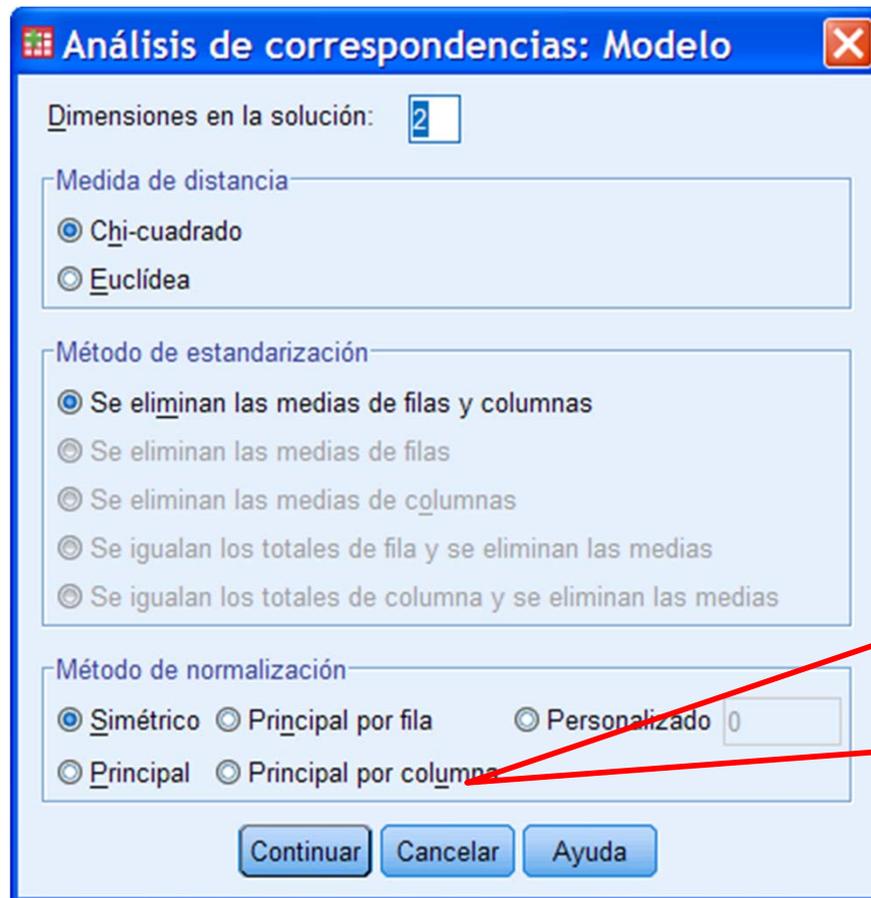
Las categorías deben ser iguales

La categoría es suplementaria

Pulsar actualizar.  
Ninguna restricción  
para las categorías



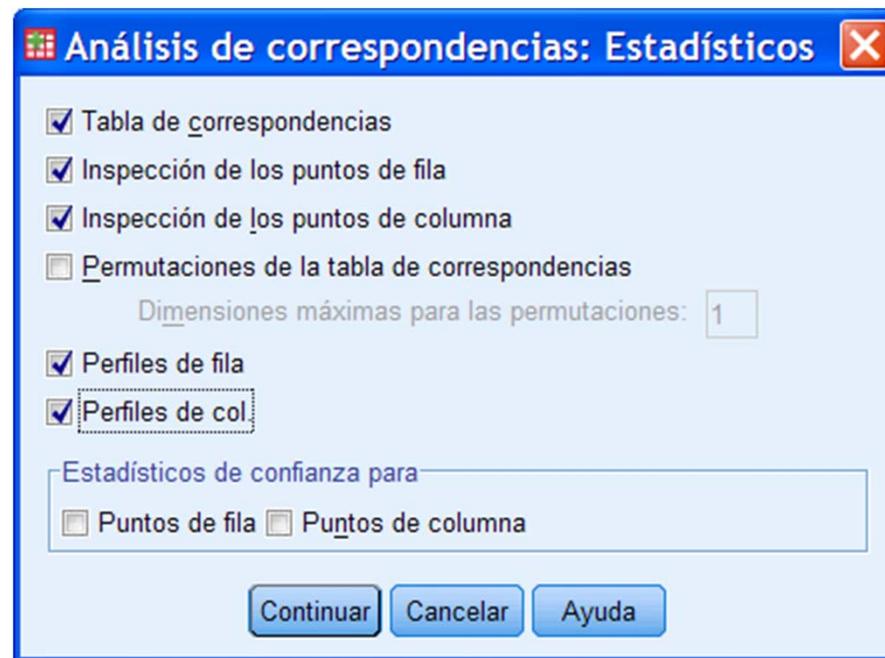
- Modelo (se mantienen las opciones por defecto)
  - Dimensiones en la solución : 2
  - Medida de distancia : Chi-cuadrado. SPSS proporciona el contraste de la hipótesis nula: las variables son independientes)
  - Método de normalización: Simétrico (análisis conjunto de filas y columnas)



Si sólo estuviéramos interesados en estudiar las diferencias por filas seleccionaríamos “principal por fila”. Si el interés son las diferencias por columnas “principal por columna”

- Estadísticos

- Tabla de Correspondencias: tabla de contingencia con las frecuencias absolutas observadas
- Inspección de los puntos de fila: puntuaciones y contribuciones para las filas
- Inspección de los puntos de columna: puntuaciones y contribuciones para las columnas
- Perfiles de fila: perfil de la fila en las columnas (frecuencia relativa de cada casilla en el total de la fila)
- Perfiles de columna: perfil de la columna en las filas (frecuencia relativa de cada casilla en el total de la columna)



- Gráficos

- Diagramas de dispersión

- Diagrama de dispersión biespacial (filas y columnas juntas)

- Puntos de fila (sólo filas)

- Puntos de columna (sólo columnas)

- Dimensiones del gráfico: Mostrar todas las dimensiones de la solución

Análisis de correspondencias: Gráficos

Diagramas de dispersión

Diagrama de dispersión biespacial

Puntos de fila

Puntos de columna

Ancho de la etiqueta de identificación para los diagramas de dispersión: 20

Gráfico de líneas

Categorías de fila transformadas

Categorías de columna transformadas

Ancho de la etiqueta de identificación para los gráficos de líneas: 20

Dimensiones del gráfico

Mostrar todas las dimensiones de la solución

Restringir el número de dimensiones

Dimensión menor:

Dimensión mayor:

Continuar Cancelar Ayuda