Lección

Estadística

ANÁLISIS DE CORRESPONDENCIAS

© Citar como: Salvador Figueras, M (2003): "Análisis de Correspondencias", [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/correspondencias> [y añadir fecha consulta].

Presentación:

El Análisis de Correspondencias es una técnica estadística que se aplica al análisis de tablas de contingencia y construye un diagrama cartesiano basado en la asociación entre las variables analizadas. En dicho gráfico se representan conjuntamente las distintas modalidades de la tabla de contingencia, de forma que la proximidad entre los puntos representados está relacionada con el nivel de asociación entre dichas modalidades.

En esta lección se va a dar una breve visión general de dicha técnica ilustrada con ejemplos.

Introducción

¿Existe algún tipo de relación entre el sexo, la religión y la nacionalidad de una persona?

¿Hay alguna relación entre el sexo, el nivel de estudios y la provincia en la que viven de los parados de un país?

¿Es cierto que las personas con los ojos claros tienden a ser rubias y los que tienen los ojos oscuros tienden a tener el pelo de color castaño o negro?

¿Existe alguna relación entre el sector en el que se encuentra encuadrada una empresa y la posibilidad de que quiebre?

¿Existe alguna relación entre el estado marital de una persona que solicite un crédito y la posibilidad de que sea un moroso?

¿QUÉ TIENEN EN COMÚN TODOS ESTOS PROBLEMAS? ¿CÓMO RESOLVERLOS?

En esta lección trataremos de responder a estas cuestiones.

Objetivos

1) Plantear el problema a resolver por un Análisis de Correspondencias
2) Calcular los perfiles marginales, fila y columna de una tabla de contingencia bidimensional e interpretarlos
3) Analizar la dependencia/independencia de las filas y columnas de una tabla de contingencia bidimensional
4) Calcular e interpretar los resultados de un Análisis de Correspondencias tanto desde un punto de vista gráfico como numérico
5) Calcular e interpretar los resultados de un Análisis de Correspondencias Múltiples

Apartados

Contenidos

1.- PLANTEAMIENTO DEL PROBLEMA

El Análisis de Correspondencias es una técnica estadística que se utiliza para analizar, desde un punto de vista gráfico, las relaciones de dependencia e independencia de un conjunto de variables categóricas a partir de los datos de una tabla de contingencia.

Para ello asocia a cada una de las modalidades de la tabla, un punto en el espacio $\mathbf{R}^n$ (habitualmente n=2) de forma que las relaciones de cercanía/lejanía entre los puntos calculados reflejen las relaciones de dependencia y semejanza existentes entre ellas.

En esta lección comenzaremos analizando el problema bidimensional que es el que analiza el **Análisis de Correspondencias** propiamente dicho. Posteriormente consideraremos, brevemente, el problema n-dimensional con $n \geq 3$ que es el problema que analiza el **Análisis de Correspondencias Múltiples**.

1.1 Tabla de Correspondencias

Sea X e Y dos variables categóricas con valores $\{x_1,...,x_r\}$ y $\{y_1,...,y_c\}$, respectivamente.

Se observan dichas variables en n elementos de una población obteniéndose los siguientes resultados:

X/Y	y_1	...	y_j	...	y_c	Marginal Y
x_1	n_{11}	...		...	n_{1c}	$n_{1.}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	..	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...
x_r	n_{r1}	...	n_{rj}	...	n_{rc}	$n_{r.}$
Marginal X	$n_{.1}$	...	$n_{.j}$	...	$n_{.c}$	$n_{..}$

donde n_{ij} = número de elementos de la muestra con $X=x_i$, $Y=y_j$.

La tabla de frecuencias cruzadas anterior recibe el nombre de **Tabla de Correspondencias.**

La frecuencia $n_{i.} = \sum_{j=1}^{c} n_{ij}$ es el número de casos con $X=x_i$ y recibe el nombre de **Frecuencia Marginal de X = x_i.**

La frecuencia $n_{.j} = \sum_{i=1}^{r} n_{ij}$ es el número de casos con $Y=y_j$ y recibe el nombre de **Frecuencia Marginal de Y = y_j.**

Ejemplo (Parados de Aragón):

Los siguientes datos corresponden a la distribución del número de parados de Aragón (España) en el año 1996 clasificados por Sexo, Provincia y Nivel de Estudios

Tabla 1
Tabla de correspondencias del paro en Aragón en 1996

Tabla de correspondencias

	Nivel de Estudios							
Sexo y Provincia	Est_Pri	Cf_Esc	Gra_Esc	BUP	FP	Diplomado	Universitario	Margen activo
H_Huesca	147	1120	908	268	149	127	94	2813
H_Teruel	182	751	564	108	138	50	58	1851
H_Zaragoza	415	6545	5690	1997	1415	670	877	17609
M_Huesca	72	902	1646	561	417	461	236	4295
M_Teruel	57	534	1127	288	331	260	127	2724
M_Zaragoza	204	5931	9434	3250	2872	2196	1890	25777
Margen activo	1077	15783	19369	6472	5322	3764	3282	55069

En este caso X = Sexo*Provincia y toma r=6 valores correspondientes a todas las combinaciones de Sexo (Hombre, Mujer) y Provincia (Huesca, Teruel y Zaragoza) e Y = Nivel de estudios y toma c=7 valores (Estudios Primarios, Certificado Escolar, Graduado Escolar, BUP, FP, Diplomado y Universitario)

El número total de casos es 55069 y $n_{3.}$ = 17609 es la frecuencia marginal de parados varones de Zaragoza y $n_{.4}$ = 6474 es la frecuencia marginal de parados cuyo nivel de estudios alcanza hasta BUP

2. PERFILES MARGINALES Y CONDICIONALES

Los **perfiles marginales** describen la distribución marginal de las variables X e Y. Vienen dados por las siguientes tablas:

Perfil marginal de X

X	x_1	...	x_i	...	x_r	Total
Frecuencias Marginales	$100\frac{n_{1.}}{n_{..}}$	...	$100\frac{n_{i.}}{n_{..}}$	...	$100\frac{n_{r.}}{n_{..}}$	100

Perfil marginal de Y

Y	y_1	...	y_j	...	y_c	Total
Frecuencias Marginales	$100\frac{n_{.1}}{n_{..}}$	...	$100\frac{n_{.j}}{n_{..}}$	...	$100\frac{n_{.c}}{n_{..}}$	100

Los **perfiles condicionales** describen las distribuciones condicionadas asociadas a la Tabla de Correspondencias.

Los **perfiles fila** describen las distribuciones condicionadas de la variable Y por los distintas modalidades de la variable X. Se obtienen a partir de la Tabla de Correspondencias y el perfil marginal de X mediante las siguientes expresiones:

Y	y_1	...	y_j	...	y_c	Totales
$f(y/X=x_1)$	$100\frac{n_{11}}{n_{1.}}$	...	$100\frac{n_{1j}}{n_{1.}}$	...	$100\frac{n_{1c}}{n_{1.}}$	100
...	...	...	...	...	...	...
$f(y/X=x_i)$	$100\frac{n_{i1}}{n_{i.}}$	...	$100\frac{n_{ij}}{n_{i.}}$	...	$100\frac{n_{ic}}{n_{i.}}$	100
...	...	...	...	...	...	...
$f(y/X=x_r)$	$100\frac{n_{r1}}{n_{r.}}$	...	$100\frac{n_{rj}}{n_{r.}}$	...	$100\frac{n_{rc}}{n_{r.}}$	100

Los **perfiles columna** describen las distribuciones condicionadas de la variable X por los distintas modalidades de la variable Y. Se obtienen a partir de la tabla de correspondencias y el perfil marginal de X mediante las siguientes expresiones:

Y	y_1	...	y_j	...	y_c	Totales
$f(y/X=x_1)$	$100\frac{n_{11}}{n_{1.}}$	...	$100\frac{n_{1j}}{n_{1.}}$	...	$100\frac{n_{1c}}{n_{1.}}$	100
...	...	...	...	...	...	...
$f(y/X=x_i)$	$100\frac{n_{i1}}{n_{i.}}$	...	$100\frac{n_{ij}}{n_{i.}}$	...	$100\frac{n_{ic}}{n_{i.}}$	100
...	...	...	...	...	...	...
$f(y/X=x_r)$	$100\frac{n_{r1}}{n_{r.}}$	...	$100\frac{n_{rj}}{n_{r.}}$	...	$100\frac{n_{rc}}{n_{r.}}$	100

Ejemplo (Parados de Aragón)(continuación)

En la Tabla 2 se muestran los perfiles fila así como el perfil marginal de la variable Sexo*Provincia. Así mismo, en la Figura 1 se representan, en forma de diagrama de líneas, los perfiles fila.

Así, por ejemplo, se observa que un 9.8% de los Hombres de Teruel parados tienen un nivel de estudios primario cifra mucho más elevada que la correspondiente a la distribución marginal en la que únicamente un 2% de los parados poseen dicho nivel de estudios.

Se aprecia (ver Figura 1) una clara distinción por Sexos. Así entre los parados que son hombres hay una mayor tendencia a tener niveles de estudios bajos (Estudios Primarios y Certificado Escolar) mientras que las mujeres hay una mayor tendencia a tener niveles superiores (Graduado Escolar, FP, BUP, Diplomado y Universitario)

Tabla 2
Perfiles fila

% de Sexo y Provincia

		Nivel de Estudios							
		Est_Pri	Cf_Esc	Gra_Esc	BUP	FP	Diplomado	Universitario	Total
Sexo y Provincia	H_Huesca	5.2%	39.8%	32.3%	9.5%	5.3%	4.5%	3.3%	100.0%
	H_Teruel	9.8%	40.6%	30.5%	5.8%	7.5%	2.7%	3.1%	100.0%
	H_Zaragoza	2.4%	37.2%	32.3%	11.3%	8.0%	3.8%	5.0%	100.0%
	M_Huesca	1.7%	21.0%	38.3%	13.1%	9.7%	10.7%	5.5%	100.0%
	M_Teruel	2.1%	19.6%	41.4%	10.6%	12.2%	9.5%	4.7%	100.0%
	M_Zaragoza	.8%	23.0%	36.6%	12.6%	11.1%	8.5%	7.3%	100.0%
Marginal		2.0%	28.7%	35.2%	11.8%	9.7%	6.8%	6.0%	100.0%

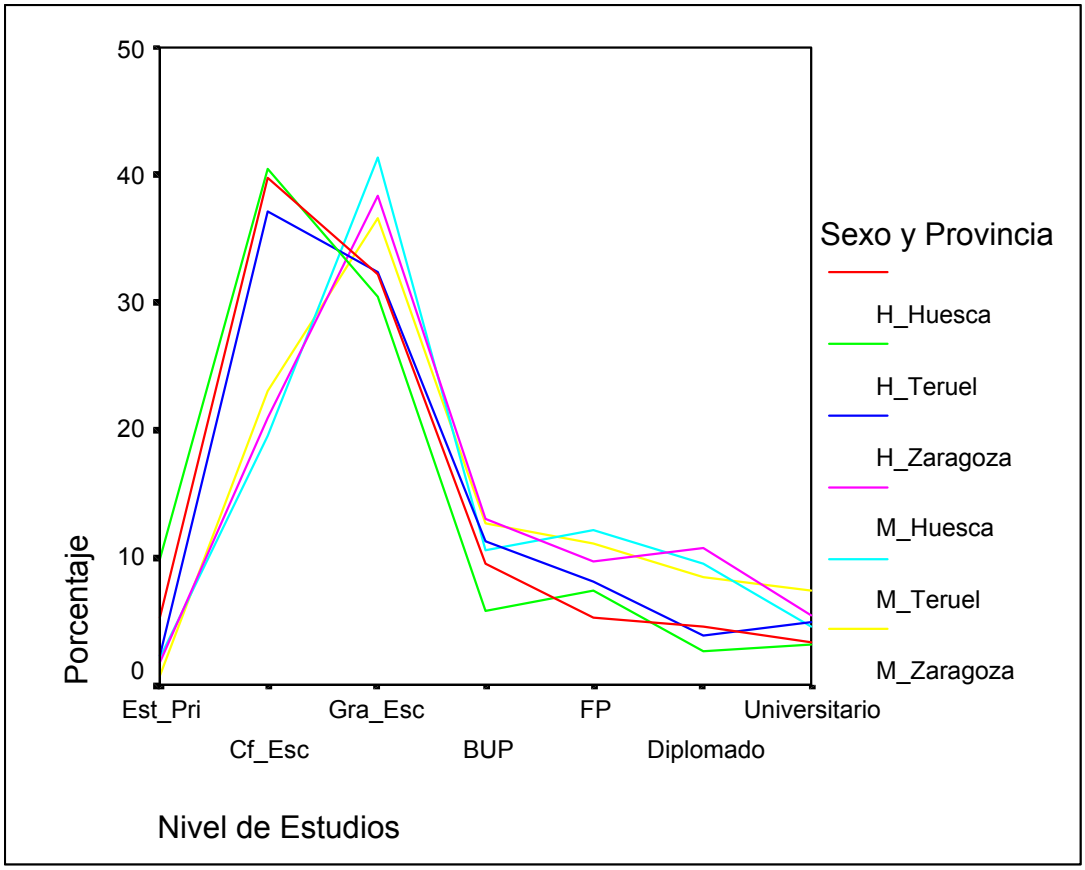


Figura 1: Diagramas de línea correspondientes a los perfiles fila

En la Tabla 3 se muestran los perfiles columna así como el perfil marginal de la variable Nivel de Estudios. Los perfiles columna aparecen, además, representados en forma de diagrama de líneas en la Figura 2

Tabla 3
Perfiles columna

% de Nivel de Estudios

		Nivel de Estudios							
		Est_Pri	Cf_Esc	Gra_Esc	BUP	FP	Diplomado	Universitario	Marginal
Sexo y Provincia	H_Huesca	13.6%	7.1%	4.7%	4.1%	2.8%	3.4%	2.9%	5.1%
	H_Teruel	16.9%	4.8%	2.9%	1.7%	2.6%	1.3%	1.8%	3.4%
	H_Zaragoza	38.5%	41.5%	29.4%	30.9%	26.6%	17.8%	26.7%	32.0%
	M_Huesca	6.7%	5.7%	8.5%	8.7%	7.8%	12.2%	7.2%	7.8%
	M_Teruel	5.3%	3.4%	5.8%	4.4%	6.2%	6.9%	3.9%	4.9%
	M_Zaragoza	18.9%	37.6%	48.7%	50.2%	54.0%	58.3%	57.6%	46.8%
Total		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Así, por ejemplo, se observa que un 16.9% de los parados con un nivel de estudios primario son hombres de Teruel cifra mucho más elevada que la correspondiente a la distribución marginal en la que tan sólo un 3.4% son hombres de Teruel.

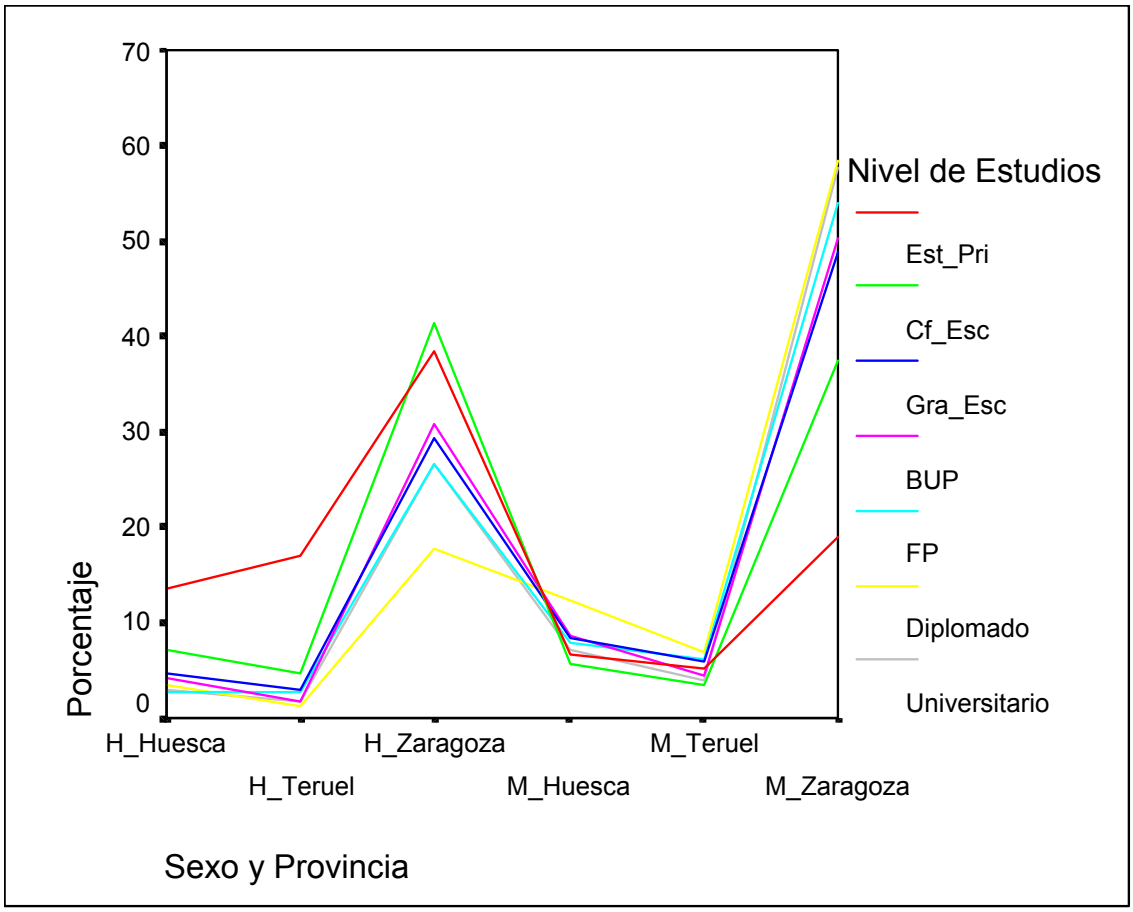


Figura 2: Diagramas de línea de los perfiles columna

Se vuelven a apreciar las diferencias por sexos comentadas anteriormente siendo éstas especialmente agudas en los hombres de las provincias de Huesca y Teruel en las que se observa una especial incidencia del paro en los niveles de estudios más bajos (sin estudios y certificado escolar) y en las mujeres de Zaragoza en los niveles de estudio más altos (diplomados y licenciados) . Así mismo se observa que hay una mayor incidencia del

paro en las diplomadas de Huesca y una menor en los graduados escolares varones que viven en Zaragoza.

3.- DEPENDENCIA E INDEPENDENCIA EN TABLAS DE CORRESPONDENCIAS

La existencia o no de algún tipo de relación entre las variables X e Y se analiza mediante contrastes de hipótesis sobre la independencia de dichas variables. El test de hipótesis habitualmente utilizado es el de la χ^2 de Pearson.

En dicho test la hipótesis nula es H_0: X e Y son independientes y la alternativa es H_1: X e Y son dependientes

El test se basa en comparar los perfiles fila y columna con los perfiles marginales correspondientes, teniendo en cuenta que si H_0 es cierta todos los perfiles fila (resp. columna) son iguales entre sí e iguales al perfil marginal de X (resp. de Y).

El estadístico del test viene dado por la expresión:

$$G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{i.}\left(\frac{n_{ij}}{n_{i.}}-\frac{n_{.j}}{n_{..}}\right)^2}{\frac{n_{.j}}{n_{..}}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{.j}\left(\frac{n_{ij}}{n_{.j}}-\frac{n_{i.}}{n_{..}}\right)^2}{\frac{n_{i.}}{n_{..}}}$$

donde $e_{ij} = E[n_{ij}|H_0 \text{ cierta}] = \frac{n_{i.}n_{.j}}{n_{..}}$. Intuitivamente, valores pequeños de G^2 significan que los valores de n_{ij} y e_{ij} son cercanos y, por lo tanto, que H_0 es cierta y, por el contrario, valores grandes de G^2 darían evidencia de que H_0 es falsa.

Bajo la hipótesis nula G^2 se distribuye, asintóticamente, según una $\chi^2_{(r-1)(c-1)}$ y el p-valor del test viene dado por:

$$P[\chi^2_{(r-1)(c-1)} \geq G^2_{obs}]$$

donde G^2_{obs} es el valor observado en la muestra del estadístico G^2. Para un nivel de significación $0<\alpha<1$ la hipótesis H_0 se rechaza si dicho p-valor es menor o igual que α.

Si la hipótesis nula se rechaza, las variables X e Y son dependientes. En este caso conviene analizar los perfiles condicionales fila y columna así como los residuos del modelo para estudiar qué tipo de dependencia existe entre ellas. Los residuos más

utilizados son los llamados **residuos tipificados corregidos** que vienen dados por la expresión:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}\sqrt{\left(1 - \frac{n_{i.}}{n_{..}}\right)\left(1 - \frac{n_{.j}}{n_{..}}\right)}}$$

y si toman valores grandes en valor absoluto será debido a que en la celda $X=x_i$, $Y=y_j$ de la Tabla de Correspondencias los valores de n_{ij} y e_{ij} son muy diferentes y que, por lo tanto, existe un número anormalmente alto (si $r_{ij}>0$) o bajo (si $r_{ij}<0$) de casos.

Los residuos se distribuyen asintóticamente como una N(0,1) la hipótesis H_0 y, a un nivel del 95.5% de confianza, residuos con un valor absoluto mayor que dos se consideran como valores anormalmente altos.

Ejemplo (Paro en Aragón) (continuación)

En este caso se tiene que $G^2_{obs} = 3160.768$ y, por lo tanto, el p-valor es igual a $P[\chi^2_{30} \geq G^2_{obs}] = 0$ por lo que se rechaza H_0.

En la Tabla 4 se muestran los residuos tipificados corregidos.

Tabla 4
Residuos tipificados corregidos

Residuos corregidos

		Nivel de Estudios						
		Est_Pri	Cf_Esc	Gra_Esc	BUP	FP	Diplomado	Universitario
Sexo y Provincia	H_Huesca	12.9	13.4	-3.3	-3.8	-8.0	-5.0	-6.0
	H_Teruel	24.9	11.5	-4.3	-8.0	-3.3	-7.2	-5.2
	H_Zaragoza	4.7	30.3	-9.6	-2.1	-8.9	-19.3	-6.7
	M_Huesca	-1.4	-11.6	4.5	2.8	.1	10.5	-1.3
	M_Teruel	.5	-10.7	7.0	-2.0	4.5	5.7	-2.9
	M_Zaragoza	-18.5	-27.5	6.6	5.8	11.0	14.7	12.8

La mayor parte de los residuos son mayores, en valor absoluto, que 2. Observando, además, el patrón de los signos se observa que los residuos positivos tienden a situarse en los niveles de estudio inferiores (Estudios Primarios y Certificado Escolar) para los hombres y en los superiores (Graduado Escolar, FP y Diplomados en todas las provincias; BUP en Huesca y Zaragoza y Universitario en Zaragoza) para las mujeres corroborando los comentarios hechos anteriormente al analizar los perfiles fila y columna.

4.- ANÁLISIS DE CORRESPONDENCIAS CLASICO

El examen de las razones específicas de las desviaciones de la hipótesis de independencia es la razón de ser del Análisis de Correspondencias. El método consiste, esencialmente, en encontrar la descomposición en valores singulares de la matriz:

$$\mathbf{C} = (c_{ij}) \text{ con } c_{ij} = \frac{n_{ij..} - e_{ij..}}{\sqrt{e_{ij..}}}$$

para, a partir de ella, construir un sistema de coordenadas (habitualmente bidimensional) asociado a las filas y columnas de la tabla de contingencia, que refleje las relaciones existentes entre dichas filas y columnas.

En dicha representación juegan un papel importante las llamadas "distancias χ^2" entre perfiles que son las que el Análisis de Correspondencias intenta reproducir en sus representaciones gráficas. Dichas distancias son distancias pitagóricas ponderadas entre perfiles que vienen dadas por las siguientes expresiones:

$$\text{Distancia entre perfiles filas } d_{ij} = \sum_{k=1}^{c} \frac{1}{n_{.k}} \left(\frac{n_{ik}}{n_{i.}} - \frac{n_{jk}}{n_{j.}} \right)^2$$

$$\text{Distancia entre perfiles columnas } d_{i_{ij}} = \sum_{k=1}^{r} \frac{1}{n_{k.}} \left(\frac{n_{ki}}{n_{.i}} - \frac{n_{kj}}{n_{.j}} \right)^2$$

Estas distancias tienen la llamada propiedad de equivalencia distribucional la cual afirma que si dos filas (columnas) de **N** son proporcionales entonces su amalgamiento no afectará a las distancias entre columnas (filas). Por lo tanto las distancias χ^2 son invariantes a variaciones en la codificación de las categorías con comportamiento similar en cuanto a sus perfiles condicionales.

4.1 Objetivos del Análisis de Correspondencias

El Análisis de Correspondencias busca encontrar 2 matrices de coordenadas cartesianas

$$\mathbf{A} = \begin{pmatrix} a_1' \\ \dots \\ a_r' \end{pmatrix}$$ que represente a los puntos fila con $a_i = (a_{i1},\dots,a_{ik})'$

$\mathbf{B} = \begin{pmatrix} b'_1 \\ \dots \\ b'_c \end{pmatrix}$ que represente a los puntos columna con $b_j = (b_{j1},\dots,b_{jk})'$

(habitualmente k=2) tales que se verifique que:

1) La distancia euclídea al cuadrado entre a_i y $a_{i'}$ sea la distancia χ^2 entre las filas i y i'

2) La distancia euclídea al cuadrado entre b_i y $b_{j'}$ sea la distancia χ^2 entre las columnas j y j'

3) El producto escalar entre a_i y b_j sea proporcional a los residuos tipificados c_{ij}.

4.2 Cálculo de las coordenadas

Existen diversas formas de calcular las matrices A y B anteriores. Dichas formas reciben el nombre de **normalizaciones** y se diferencian en a cuál de los objetivos 1) a 3) dan más prioridad. Una de las más utilizadas es la llamada **normalización simétrica o canónica** que busca satisfacer el objetivo 3 anterior.

Para ello descompone la matriz **C** anterior en valores singulares calculando matrices **U, D** y **V** tales que

$$\mathbf{C} = = \mathbf{UDV'}$$

$$\mathbf{U'U= V'V = I} \quad \mathbf{U}\ \text{rxK}, \mathbf{V}\ \text{cxK},\ K = \min\{r-1,c-1\}$$

$$\mathbf{D}= \text{diag}(\mu_1,\dots,\mu_K)$$

μ_i reciben el nombre de **valores singulares** i=1,...,K

Las matrices **A** y **B** se calculan a partir de las expresiones:

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{UD} \text{ y } \mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{VD}$$

donde $\mathbf{D}_r = \text{diag}(n_{1.},\dots,n_{r.})$ y $\mathbf{D}_c = \text{diag}(n_{.1},\dots,n_{.c})$.

4.3 Interpretación baricéntrica

Se verifica que:

$$\sum_{k=1}^{r} \frac{n_{k.}}{n_{..}} a_{kj} = 0\,;\ j=1,\dots,K$$

$$\sum_{k=1}^{c} \frac{n_{.k}}{n_{..}} b_{kj} = 0\,;\ j=1,\dots,K$$

por lo que los puntos $\mathbf{a}_i$ i=1,...,r y $\mathbf{b}_j$ j=1,...,c tendrá una media baricéntrica igual al origen. Además:

$$\mu_j a_{ij} = \sum_{k=1}^{c} \frac{n_{ik}}{n_{i.}} b_{kj} \quad i=1,\ldots,r;\ j=1,\ldots,K$$

$$\mu_j b_{ij} = \sum_{k=1}^{r} \frac{n_{ki}}{n_{.i}} a_{kj} \quad i=1,\ldots,c;\ j=1,\ldots,K$$

por lo que las coordenadas de los puntos fila (columna) son medias ponderadas de las coordenadas de los puntos columna (fila) salvo un factor dado por los valores singulares, es decir los puntos fila (columna) son, salvo un factor de dilatación $1/\mu_j$, el baricentro de los puntos columna (fila).

5. REGLAS DE INTERPRETACIÓN DE LOS RESULTADOS

Además de las representaciones gráficas de los puntos $\{\mathbf{a}_i;\ i=1,...,r\}$ y $\{\mathbf{b}_j;\ j=1,...,c\}$ las siguientes medidas numéricas ayudan a intepretar mejor los resultados obtenidos.

Inercia Total

Es una medida similar a la variación total en el caso de las componentes principales y mide el grado total de dependencia existente entre las variables X e Y. Viene dada por

$$IT = \frac{G^2}{n}$$

y se tiene que

$$IT = \sum_{k=1}^{K} \mu_k^2 = \sum_{k=1}^{K} \sum_{i=1}^{r} n_{i.} a_{ik}^2 = \sum_{k=1}^{K} \sum_{j=1}^{c} n_{.j} b_{jk}^2$$

A partir de ella se calculan las **proporciones de inercia explicada** por cada una de las dimensiones $\left\{ \frac{\mu_i^2}{IT}; i = 1,..., K \right\}$ que ayudan a calibrar la importancia de cada una de las dimensiones a la hora de explicar las dependencias observadas así como las **proporciones de inercia acumulada** explicada por las i primeras dimensiones $\left\{ \sum_{k=1}^{i} \frac{\mu_k^2}{IT}; i = 1,..., K \right\}$ que ayudan a decidir el número mínimo de dimensiones necesario para explicar dichas dependencias.

Contribuciones totales

Miden la importancia de cada una de las modalidades de las variables analizadas en la construcción de los ejes factoriales construidos por el Análisis de Correspondencias. Vienen dadas por:

Contribución i-ésima fila: $Ct_k(i)= \frac{n_{i.} a_{ik}^2}{\sum_{j=1}^{r} n_{j.} a_{jk}^2} = \frac{n_{i.} a_{ik}^2}{\mu_k^2}$

Contribución j-ésima columna: $Ct_k(j)= \frac{n_{.j} b_{jk}^2}{\sum_{i=1}^{c} n_{.i} b_{ik}^2} = \frac{n_{.j} b_{jk}^2}{\mu_k^2}$

Se verifica que $\sum_{i=1}^{r} Ct_k(i) = \sum_{j=1}^{c} Ct_k(j) = 1$

Se utilizan para interpretar el significado de los ejes utilizando, para cada uno de ellos, las modalidades con contribuciones más fuertes

Contribuciones relativas

Miden la importancia de cada factor para explicar la posición, en el diagrama cartesiano, de cada una de las modalidades de las variables analizadas, representando la parte de la distancia al origen de coordenadas, explicada por dicho factor. Vienen dadas por:

$$Cr_k(i) = \frac{a_{ik}^2}{\sum_{l=1}^{K} a_{il}^2} \qquad Cr_k(j) = \frac{b_{jk}^2}{\sum_{l=1}^{K} b_{jl}^2}$$

y son los cuadrados de los cosenos de los ángulos entre la dimensión k-ésima y el punto representando el perfil de la fila i-ésima o la columna j-ésima.

Se verifica que:

$$\sum_{i=1}^{r} Cr_k(i) = \sum_{j=1}^{c} Cr_k(j) = 1$$

Se utilizan para analizar las proximidades entre los puntos haciendo más hincapié en aquellos factores cuyas contribuciones sean más elevadas a la hora de explicar dichas proximidades.

Elementos suplementarios

Son filas o columnas de la tabla de contingencia no utilizadas en el cálculo de los ejes factoriales pero que, una vez calculados éstos, se sitúan en el diagrama cartesiano con el fin de ayudar en la interpretación de los resultados obtenidos. Sus coordenadas se calculan utilizando las relaciones baricéntricas existentes entre los puntos fila y columna.

No todos los paquetes estadísticos proporcionan, explícitamente, esta utilidad por lo que se aconseja estudiar los manuales de ayuda en cada caso.

Ejemplo (Paro en Aragón) (continuación)

En las Tablas 5 a 7 y las Figuras 3 a 5 se muestran los resultados obtenidos al realizar un Análisis de Correspondencias con normalización simétrica a los datos de la Tabla 1 utilizando el programa Correspondence de SPSS 10.0.

En la Tabla 5 se muestran las contribuciones de cada una de las K = min{6-1,7-1} = 5 dimensiones calculadas por el programa, a la intercia total. Se observa que, solamente la primera dimensión contribuye un 82.5% a dicha inercia y que las dos primeras contribuyen un 96.9% por lo que se concluye que las dependencias observadas en la tabla vienen adecuadamente capturadas por las 2 primeras dimensiones

Tabla 5
Contribuciones a la inercia total de cada dimensión

Resumen

Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Proporción de inercia		Confianza para el Valor propio	
					Explicada	Acumulada	Desviación típica	Correlación 2
1	.218	.047			.825	.825	.004	.244
2	.091	.008			.144	.969	.006	
3	.035	.001			.021	.989		
4	.024	.001			.010	.999		
5	.006	.000			.001	1.000		
Total		.057	3160.768	.000[a]	1.000	1.000		

a. 30 grados de libertad

En la Tabla 6 y la Figura 3 se muestran las puntuaciones de los perfiles fila de la Tabla 1 así como las contribuciones totales de cada perfil fila a la inercia de cada dimensión y las contribuciones relativas de cada dimensión la inercia del punto.

Se observa (ver Figura 3) que la primera dimensión discrimina por Sexos. Además, (ver Tabla 6) los puntos fila que más contribuyen la inercia de la primera dimensión son las Mujeres de Zaragoza y los Hombres de las 3 provincias. Dicha dimensión es, a su vez, la que más contribuye a explicar la inercia de cada uno de dichos puntos.

La segunda dimensión (cuyo poder discriminante es menor, ver Tabla 5) discrimina por provincias separando, esencialmente a Teruel de Zaragoza (ver Figura 3). Los puntos que más contribuyen a su inercia son, consecuentemente, los puntos fila de Zaragoza y Teruel (ver Tabla 6). Además dicha dimensión tiene una contribución relativa no despreciable a la inercia de los puntos fila de Teruel, a las Mujeres de Huesca y a los Hombres de Zaragoza.

Tabla 6
Contribuciones totales y relativas de los perfiles fila

Examen de los puntos de fila[a]

Sexo y Provincia	Masa	Puntuación en la dimensión 1	Puntuación en la dimensión 2	Inercia	Contribución: De los puntos a la inercia de la dimensión 1	Contribución: De los puntos a la inercia de la dimensión 2	Contribución: De la dimensión a la inercia del punto 1	Contribución: De la dimensión a la inercia del punto 2	Contribución: Total
H_Huesca	.051	-.781	.253	.007	.143	.036	.924	.040	.964
H_Teruel	.034	-1.235	1.082	.015	.236	.433	.743	.238	.982
H_Zaragoza	.320	-.421	-.290	.015	.261	.295	.833	.164	.997
M_Huesca	.078	.376	.330	.004	.051	.094	.645	.207	.852
M_Teruel	.049	.360	.508	.003	.029	.141	.463	.386	.849
M_Zaragoza	.468	.361	-.016	.014	.280	.001	.983	.001	.983
Total activo	1.000			.057	1.000	1.000			

a. Normalización Simétrica

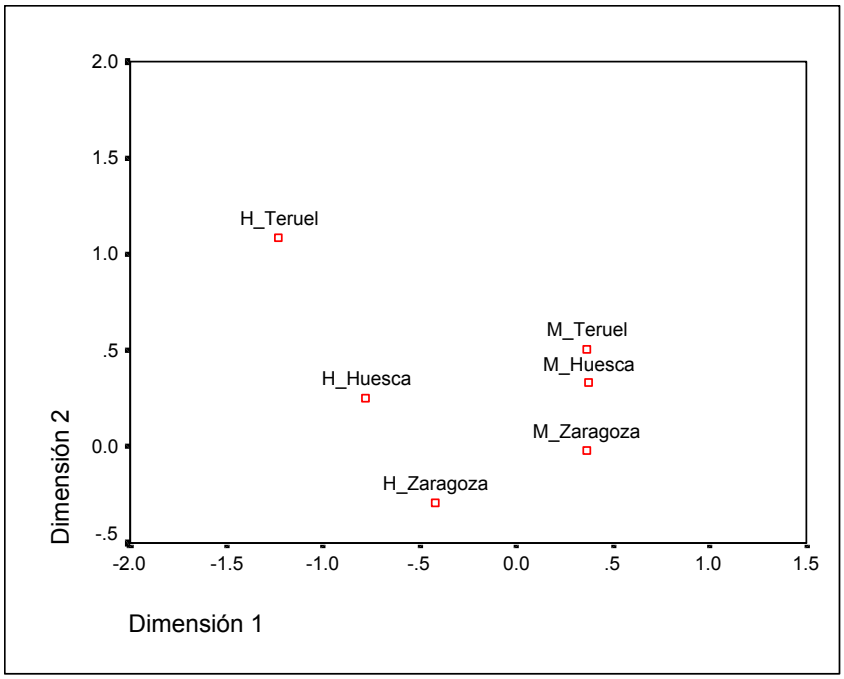


Figura 3: Gráfico de los puntos correspondientes los perfiles fila

En la Tabla 7 y la Figura 4 se muestran las puntuaciones de los perfiles columna de la Tabla 1 así como las contribuciones totales de cada perfil fila a la inercia de cada dimensión y las contribuciones relativas de cada dimensión la inercia del punto.

Se observa (ver Figura 4) que la primera dimensión discrimina los niveles de estudio más bajos (Estudios Primarios y Certificado Escolar) frente al resto siendo éstas modalidades junto con la de los Diplomados las que más contribuyen a su inercia (ver Tabla 7). Además, (ver Tabla 7) dicha dimensión es la que más contribuye a la inercia de todos los perfiles columna

La segunda dimensión separa al perfil correspondiente al nivel de Estudios Primario del resto de los niveles (ver Figura 3) siendo éte punto columna el que más contribuye a su inercia (ver Tabla 7) Además dicha dimensión tiene una contribución relativa no despreciable a la inercia de los que tienen un Certificado de Estudios Primario y los que tienen BUP.

Tabla 7
Contribuciones totales y relativas de los perfiles columna

Examen de los puntos columna[a]

		Puntuación en la dimensión			Contribución				
					De los puntos a la inercia de la dimensión		De la dimensión a la inercia del punto		
Nivel de Estudios	Masa	1	2	Inercia	1	2	1	2	Total
Est_Pri	.020	-1.678	1.669	.017	.253	.600	.705	.291	.996
Cf_Esc	.287	-.549	-.228	.020	.398	.164	.932	.067	.999
Gra_Esc	.352	.149	.088	.002	.036	.030	.805	.119	.924
BUP	.118	.216	-.195	.002	.025	.049	.690	.235	.925
FP	.097	.371	.076	.003	.061	.006	.845	.015	.860
Diplomado	.068	.753	.412	.010	.178	.128	.864	.108	.973
Universitario	.060	.423	-.186	.003	.049	.023	.726	.059	.785
Total activo	1.000			.057	1.000	1.000			

a. Normalización Simétrica

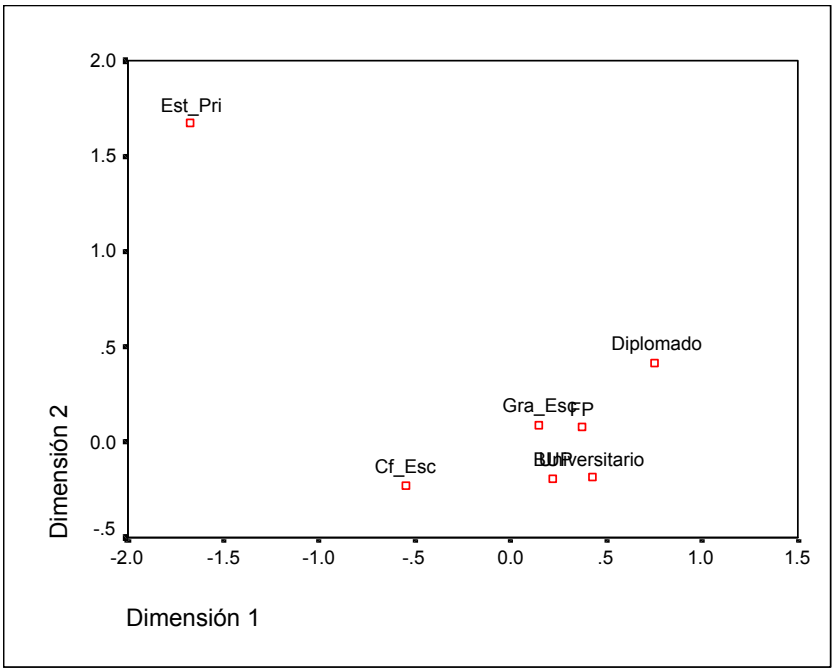


Figura 4
Gráfico de los puntos correspondientes los perfiles columna

Finalmente, en la Figura 5 se muestra el gráfico conjunto de los puntos correspondientes a los perfiles condicionales fila y columna. Dicho gráfico vuelve a poner de manifiesto las relaciones de dependencia existentes entre las dos variables ya comentadas al analizar los perfiles y los residuos tipificados corregidos.

Así se observa que la primera dimensión discrimina entre Sexos debido a la tendencia a haber más parados varones en los niveles de estudios más bajos (Estudios Primarios y Certificado Escolar) y más parados mujeres en el resto de los niveles. Este hecho de manifiesto analizando las relaciones de proximidad y alejamiento de los puntos fila y columna. Así, por ejemplo, la cercanía entre los puntos fila Hombres de Teruel y columna Estudios Primarios es debida a la tendencia en ambos perfiles a tener mayor número de parados de la modalidad representada por el otro perfil tal y como muestra el alto valor del residuo presentado en la Tabla 4.

Razonando de esta manera se observa que la segunda dimensión pone también de manifiesto la asociación positiva existente entre las categorías Mujer de Huesca y Teruel y el nivel de estudios Diplomado y las de mujer de Zaragoza y Universitario mostrando una especialización en el tipo de paro existente en las mujeres de Aragón. En Huesca y Teruel tiende a haber mayores niveles de paro relativo en las diplomadas mientras que en Zaragoza es en las licenciadas.

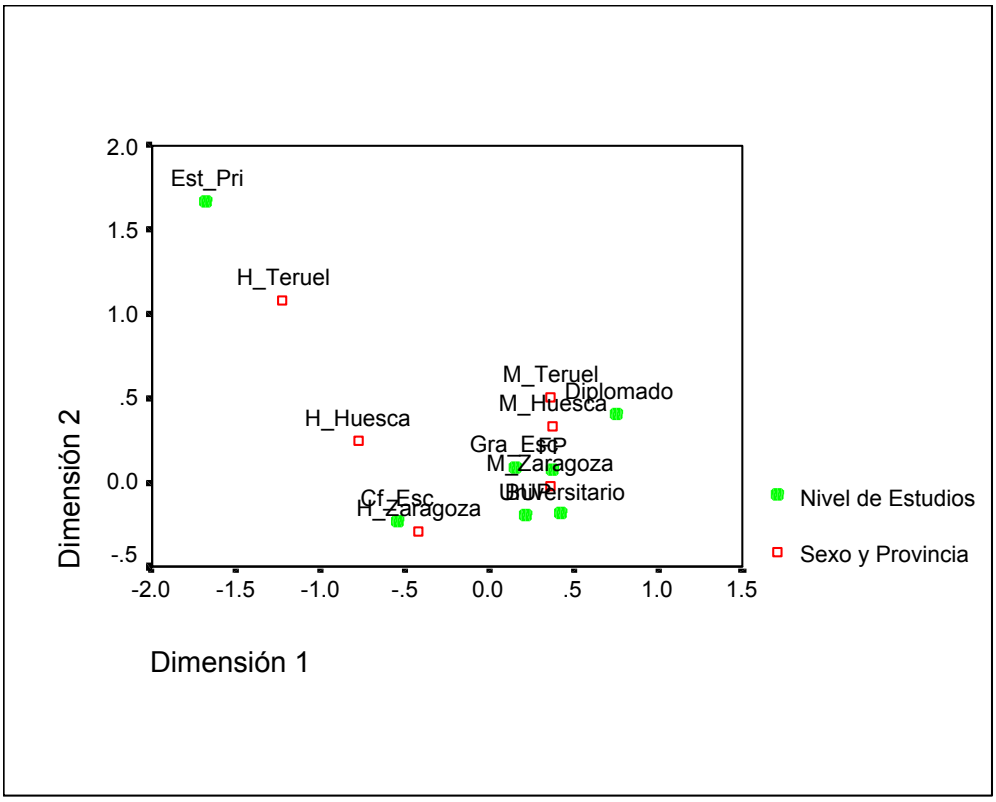


Figura 5
Gráfico conjunto de los puntos correspondientes a los perfiles condicionales fila y columna

6. - ANÁLISIS DE CORRESPONDENCIAS MÚLTIPLES

Se aplica a tablas de contingencias en las que por filas se tienen n individuos y por columnas s variables categóricas con p_i i=1,...,s categorías mutuamente excluyentes y exhaustivas.

La tabla de datos tiene, por lo tanto, la forma:

$$Z = [Z_1, Z_2, ..., Z_s]$$

con Z_i matriz nxp_i de forma que

z_{ij} = 1 si el individuo i-ésimo ha elegido la modalidad j

z_{ij} = 1 si el individuo i-ésimo no ha elegido la modalidad j

con i=1,...,n y j=1,..., $p=p_1 + p_2 + ... + p_s$

El Análisis de Correspondencias Múltiples se basa en realizar un Análisis de Correspondencias sobre la llamada matriz de Burt:

$$\mathbf{B = Z'Z}$$

Dicha matriz se construye por superposición de cajas. En los bloques diagonales aparecen matrices diagonales conteniendo las frecuencias marginales de cada una de las

variables analizadas. Fuera de la diagonal aparecen las tablas de frecuencias cruzadas correspondientes a todas las combinaciones 2 a 2 de las variables analizadas

Se toman como dimensiones aquellas cuya contribución a la inercia supera 1/p.

Distancias χ^2

En este caso vienen dadas por las expresiones

$$d^2(j,j') = \sum_{i=1}^{n} n\left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}}\right)^2 \text{ distancia entre modalidades}$$

$$d^2(i,i') = \frac{1}{s}\sum_{j=1}^{p} \frac{n}{z_{.j}}\left(z_{ij} - z_{i'j}\right)^2 \text{ distancia entre individuos}$$

Coordenadas baricéntricas

Se verifica, en este caso particular, que:

$$b_{jk} = \frac{1}{z_{.j}\mu_k}\sum_{i\in I(j)} a_{ik} \text{ donde } I(j) = \{i\in\{1,...,n\}: z_{ij} = 1\}$$

es decir, salvo un factor de dilatación, la modalidad j es el punto medio de la nube de individuos que la han elegido como respuesta

$$a_{ik} = \frac{1}{s\mu_k}\sum_{j\in p(i)} b_{jk} \text{ donde } p(i) = \{j\in\{1,...,p\}: z_{ij} = 1\}$$

es decir, salvo un factor de dilatación, el individuo i es el punto medio de la nube de modalidades que él ha elegido.

La representación obtenida es la mejor que satisface estas dos propiedades en el sentido de que el coeficiente de dilatación $1/\mu_k$ es el mínimo posible

Reglas de interpretación

1) Dos individuos están próximos si han elegido globalmente las mismas modalidades
2) Dos modalidades están próximas han sido elegidas globalmente por el mismo conjunto de individuos
3) La interpretación de los factores se hace teniendo en cuenta las contribuciones totales de cada variable que vienen dadas por

$$Ct_k(q) = \sum_{j\in q} Ct_k(j)$$

En su lugar algunos paquetes (como, por ejemplo, SPSS) calculan $\sum_{j=1}^{p_q} n_{.j} b_{jk}^2 = \mu_k^2 Ct_k(q)$ que es la varianza de las puntuaciones de las modalidades de cada variable. A esta medida la llama **medida de discriminación de la variable.**

4) $d^2(j,O) = \frac{n_{..}}{z_{.j}} - 1$ por lo que una modalidad estará más alejada del origen de coordenadas cuanto menor número de efectivos tenga

5) La inercia de una variable $I(q) = \sum_{j=1}^{p_q} I(j) = \frac{1}{s}(p_q - 1)$ es función creciente de su número de modalidades.

6) La inercia total vale $I = \frac{p}{s} - 1$ y no tiene ninguna significación estadística

Ejemplo (Análisis de los procedimientos y tipos de compra de los clientes de una empresa)

En este ejemplo analizamos los procedimientos y tipos de compra de los clientes de una empresa y su relación con el tamaño de dichos clientes así como con su pertenencia a un determinado sector. Los datos se han tomado de Hair et al. (1999) y corresponden a una encuesta realizada a una muestra de 100 clientes de una empresa que dichos autores denominan HATCO.

Las variables analizadas vienen detalladas en la Tabla 8

Tabla 8
Variables analizadas

Nombre	Significado	Valores
TAMAÑO	Tamaño de la empresa	Pequeña y Grande
ESPCOM	Especificación de compras	Al por mayor y al por menor
PROCOM	Procedimiento de compras	Centralizado y No centralizado
INDUSTRIA	Tipo de Industria	A y no A
SITUACOM	Situación de compra	Nueva, Modificada y Simple

En Tabla 9 se muestra la matriz de Burt correspondiente a dichas variables. Dicha matriz contiene en la diagonal principal las distribuciones marginales de cada una de las variables y por bloques las tablas de frecuencias cruzadas para cada posible par de ellas.

En la Tabla 10 y la Figura 6 se muestran algunos de los resultados obtenidos al aplicar un Análisis de Correspondencias Múltiples a los datos de la Tabla 8. El programa utilizado ha sido HOMALS de SPSS 10.0. Se han extraído 3 dimensiones con el fin de que todas las modalidades queden bien reflejadas en el gráfico tal y como lo demuestran las medidas de discriminación. La dimensión 1 tiene un valor singular más grande que las otras dos y es la que más discrimina entre las diversas modalidades. El poder discriminante de las otras dos dimensiones es similar.

Del análisis de los gráficos de la Figura 6 se aprecia que:

-Las empresas grandes tienden a utilizar procedimientos centralizados, compras al por mayor y de tipo modificada o nueva

-Las empresas pequeñas tiende a utilizar procedimientos no centralizados, compras al por menor y de tipo simple

-El tipo de Industria es independiente respecto al resto de las variables

Tabla 9
Matriz de Burt

	Pequeña	Grande	Por mayor	Por menor	No Centralizada	Centralizada	Otras Industrias	Tipo A	Nueva	Modificada	Simple
Pequeña	60	0	0	60	50	10	30	30	10	16	34
Grande	0	40	40	0	0	40	20	20	24	16	0
Por mayor	0	40	40	0	0	40	20	20	24	16	0
Por menor	60	0	0	60	50	10	30	30	10	16	34
No Centralizada	50	0	0	50	50	0	26	24	8	10	32
Centralizada	10	40	40	10	0	50	24	26	26	22	2
Otras Industrias	30	20	20	30	26	24	50	0	18	16	16
Tipo A	30	20	20	30	24	26	0	50	16	16	18
Nueva	10	24	24	10	8	26	18	16	34	0	0
Modificada	16	16	16	16	10	22	16	16	0	32	0
Simple	34	0	0	34	32	2	16	18	0	0	34

Tabla 10
Resultados del Análisis de Correspondencias Múltiples

Autovalores

Dimensión	Autovalores
1	.652
2	.205
3	.198

Medidas de discriminación

	Dimensión		
	1	2	3
TAMAÑO	.921	.001	.002
ESPCOM	.921	.001	.002
PROCOMP	.825	.012	.001
INDUSTR	.000	.427	.568
SITUACOM	.594	.585	.419

TAMAÑO

	Frecuencia marginal	Cuantificaciones de categorías		
		Dimensión		
		1	2	3
Pequeña	60	-.784	.026	-.034
Grande	40	1.176	-.039	.051
Perdidos	0			

ESPCOM

	Frecuencia marginal	Cuantificaciones de categorías		
		Dimensión		
		1	2	3
Mayor	40	1.176	-.039	.051
Menor	60	-.784	.026	-.034
Perdidos	0			

PROCOMP

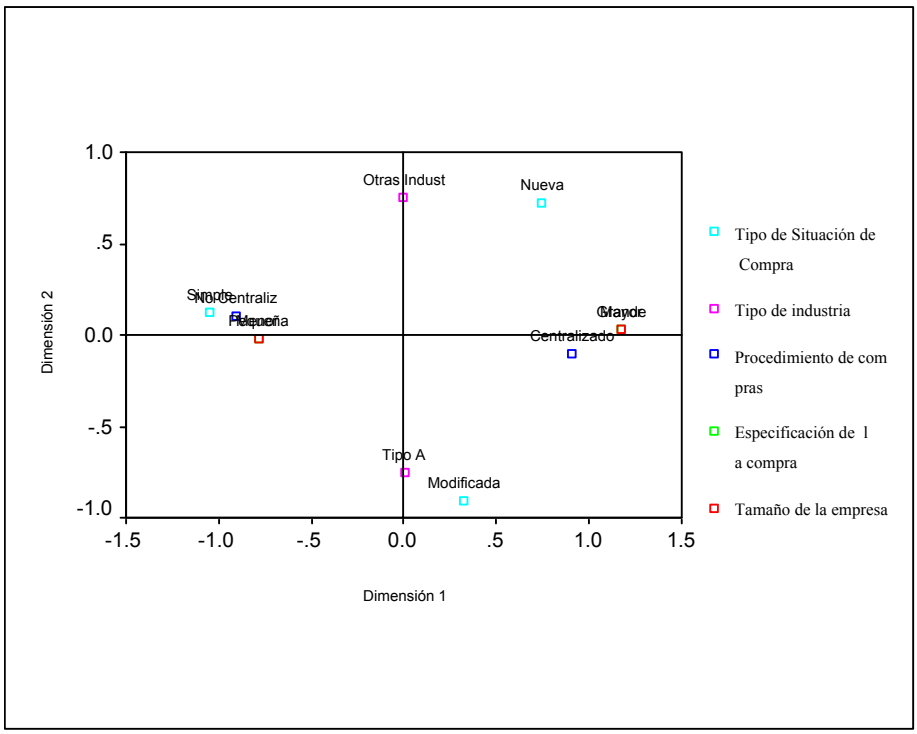
	Frecuencia marginal	Cuantificaciones de categorías		
		Dimensión		
		1	2	3
No Centralizado	50	-.908	-.109	.025
Centralizado	50	.908	.109	-.025
Perdidos	0			

INDUSTR

	Frecuencia marginal	Cuantificaciones de categorías Dimensión 1	2	3
Otras Industrias	50	.000	-.654	-.754
Tipo A	50	.000	.654	.754
Perdidos	0			

SITUACOM

	Frecuencia marginal	Cuantificaciones de categorías Dimensión 1	2	3
Nueva	34	.743	-.816	.576
Modificada	32	.324	1.045	-.929
Simple	34	-1.047	-.168	.298
Perdidos	0			



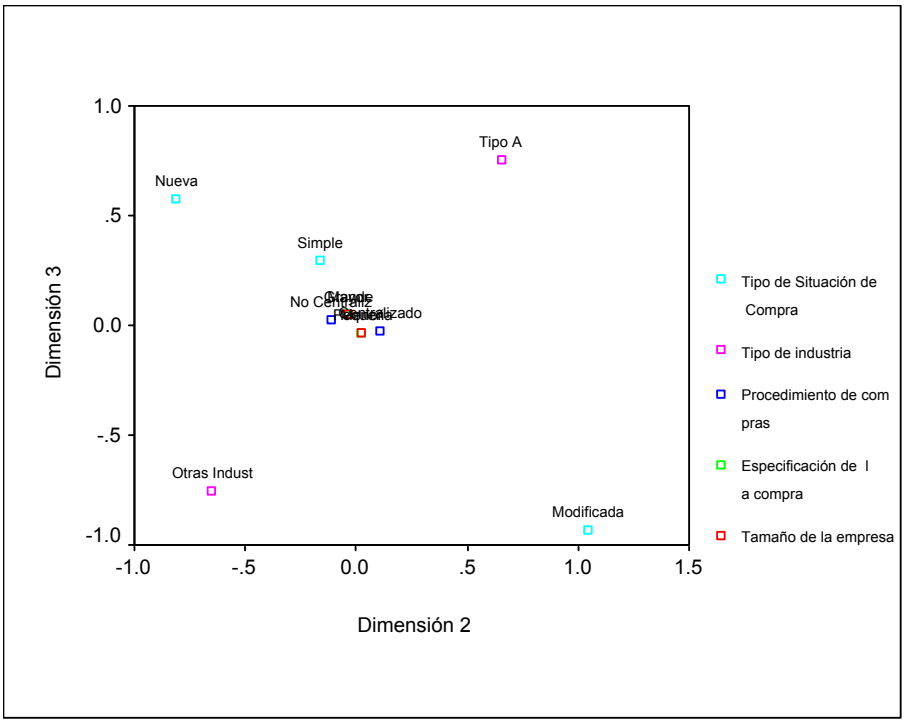


Figura 6: Gráfico de las puntuaciones obtenidas para cada modalidad de las variables de la Tabla 8 por el Análisis de Correspondencias Múltiples

Resumen

El Análisis de Correspondencias es una técnica estadística cuya finalidad es poner de manifiesto gráficamente las relaciones de dependencia existentes entre las diversas modalidades de dos o más variables categóricas a partir de la información proporcionada por sus tablas de frecuencias cruzadas.

Para ello asocia a cada modalidad un punto en el espacio $\mathbf{R}^k$ de forma que, cuanto más alejado del origen de coordenadas está el punto asociado a una modalidad de una variable, más diferente es su perfil condicional del perfil marginal correspondiente a las otras variables; además, los puntos correspondientes a dos modalidades diferentes de una misma variable estarán más cercanos cuanto más se parezcan sus perfiles condicionales y, finalmente, dichos puntos tenderán a estar más cerca (resp. más lejos) de aquéllas modalidades con las que tienen una mayor afinidad, es decir, aquéllas en los que las frecuencias observadas de la celda correspondiente tiende a ser mayor (resp. menor) que la esperada bajo la hipótesis de independencia de las variables correspondientes.

En mi opinión es una técnica complementaria al test de independencia de la χ^2 de Pearson y al estudio de los perfiles y residuos de dicho test y puede ser muy útil para interpretar los resultados obtenidos por dicho test.

Bibliografía

Desde un punto de vista práctico:

HAIR, J., ANDERSON, R., TATHAM, R. y BLACK, W. (1999). *Análisis Multivariante.* 5ª Edición. Prentice Hall.

Desde un punto de vista teórico-práctico:

GERI (1996) *L'Analyse des données évolutives: methods et applications.* Editions Technip. (Un buen libro sobre Análisis de Correspondencias Dinámico)
GIFI, A. (1990). *NonLinear Multivariate Analysis.* Wiley (Un buen libro para profundizar en el Análisis de Correspondencias Múltiples)
JOBSON, J.D. (1992) *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods.* Springer-Verlag.
LEBART, L; MORINEAU, A. and PIRON, M. (2000). *Statistique Exploratoire Multidimensionnelle.* 3ª Edition. DUNOD.

Enfocados hacia SPSS:

AGUILERA, A. (2000). *Tablas de Contingencia Bidimensionales.* Cuadernos de Estadística. Editorial La Muralla
JORAISTI, L. y LIZOSAIN, L.(2000) *Análisis de Correspondencias.* . Cuadernos de Estadística. Editorial La Muralla.
PEREZ, César (2001). *Técnicas Estadísticas con SPSS.* Prentice-Hall
VISAUTA, B. (1998) *Análisis Estadístico con SPSS para WINDOWS (Vol II. Análisis Multivariante).* Mc-Graw Hill.