El análisis de correspondencias

Ana María López Jiménez
Dept. Psicología Experimental (USE)

4. El análisis de correspondencias

- 4.1. Introducción
- 4.2. Tabla de correspondencias
- 4.3. Dependencia e independencia en tablas de contingencia: Análisis clásico
- 4.4. Perfiles marginales y perfiles condicionales
- 4.5. El análisis de correspondencias simple (ACS)
- 4.6. Análisis de correspondencias múltiple (ACM)
- 4.7. Reglas para interpretar los resultados de un AC.

4.1. Introducción

El análisis de correspondencias (AC), como el AFE, es una técnica descriptiva cuyo objetivo es la representación de tablas de contingencia en espacios de baja dimensión. En la clasificación inicial que realizamos de las técnicas multivariantes se enmarcaría dentro de las técnicas de interdependencia. Matemáticamente, es equivalente al análisis de componentes principales para variables cualitativas. A diferencia del AFE, identifica las dimensiones básicas mediante el análisis de tablas de contingencia o correspondencias obtenidas del cruce de las categorías de las variables cualitativas (escala nominal y ordinal) observadas en una muestra. Se distinguen, habitualmente, dos tipos de AC: Análisis de Correspondencias Simple (ACS) y Análisis de Correspondencias Múltiple (ACM). El ACS nos permite identificar las dimensiones básicas subyacentes a la combinación de modalidades o niveles de dos variables cualitativas. El número máximo de dimensiones que se pueden identificar en un ACS depende del número de categorías de cada variable. Concretamente, si una variable tiene I categorías y la otra tiene J categorías, el número de dimensiones (o factores) es

$$\min \{I-1, J-1\})$$

La generalización de está técnica a cualquier número de variables es lo que se conoce como Análisis de Correspondencias Múltiple. En ACM el número máximo de dimensiones es:

$$\text{Min } \{m, N-1\}$$

donde m es el número de categorías de las variables sin datos perdidos menos el número de dichas variables y N es el tamaño de la muestra.

4.1. Introducción

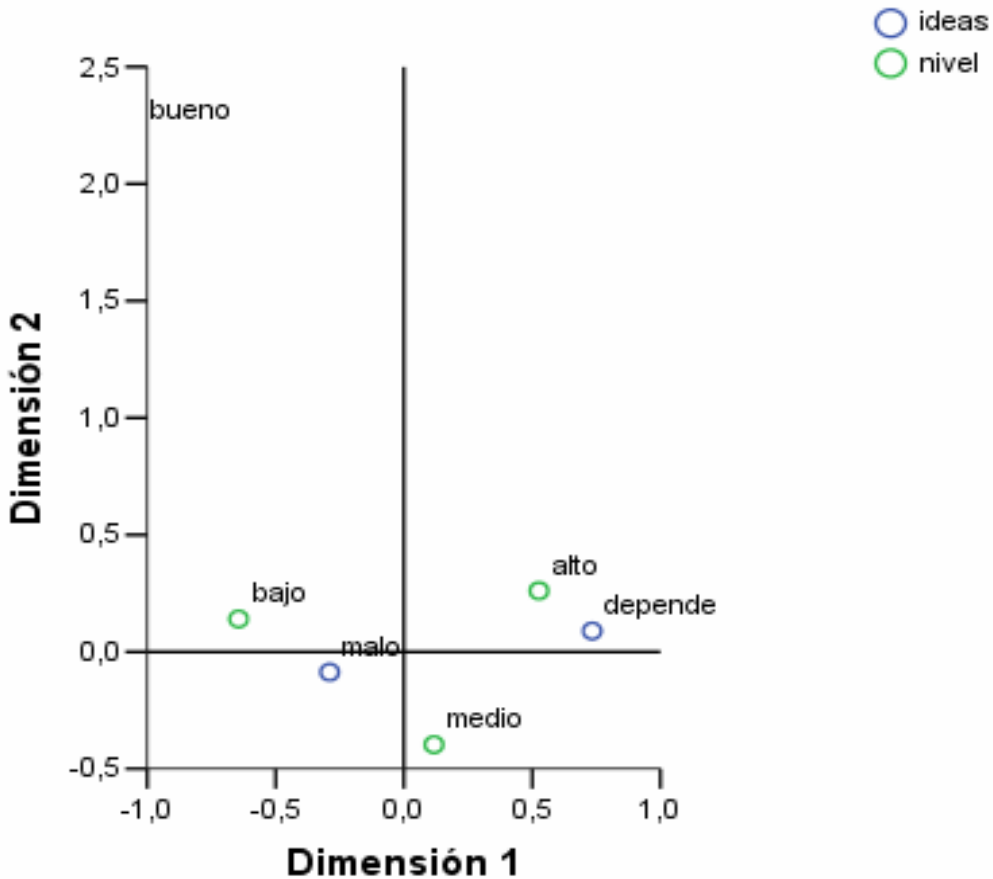
- Preguntas de investigación que se pueden resolver mediante AC
  - Existe alguna relación entre la opinión de los padres acerca de ser hijo único y el nivel cultural
  - Están determinados atributos de los coches relacionados con determinadas marcas.
  - Existe alguna relación entre tener o no estrés laboral y el sector al que se pertenece en la universidad.
  - Está relacionada la opinión de los padres acerca del consumo de drogas blandas con el hecho de tener hijos adolescentes y con el género.
  - Existe relación entre ser fumador con el género, con la hipertensión y con la presencia de enfermedades cardiovasculares.
  - Existe relación entre las diferentes estrategias de búsqueda de empleo, la provincia, el estrato de edad y el género

Estos son algunos ejemplos en los que podemos utilizar AC.

4.1. Introducción

- En los tres primeros ejemplos se pregunta sobre la relación entre dos variables cualitativas y el AC se denomina Análisis de Correspondencias Simple o Binario.
- Los tres últimos casos relacionan más de dos variables cualitativas y el AC se denomina Análisis de Correspondencias Múltiple.
- Para los objetivos didácticos que pretendemos con esta presentación describiremos el Análisis de Correspondencias Simple (ACS) sabiendo que el ACM es una generalización del mismo.
- En resumen: el AC hay que entenderlo como una técnica descriptiva que nos va a permitir elaborar un mapa perceptual de las categorías de las variables analizadas en un espacio de pocas dimensiones (habitualmente 2). La mayor o menor distancia entre los puntos representados reflejan relaciones de dependencia y semejanza más o menos fuertes entre las categorías representadas (Peña, 2002, Figueras, 2003).

4.1. Introducción



4.2. Tabla de correspondencias

- Para los objetivos didácticos que se pretenden con esta presentación vamos a describir los elementos básicos de un AC partiendo de un ejemplo sencillo: la relación entre ideas de los padres acerca de si ser hijo único es bueno o malo y el nivel educativo.
- La variable nivel cultural es una variable ordinal con tres niveles bajo, medio y alto.
- La variable opinión se codificó como bueno, malo, depende y no sabe.

4.2. Tabla de correspondencias

El punto de partida del AC es una tabla de contingencia o de correspondencias. Supongamos que queremos estudiar las ideas de los padres acerca de si ser hijo único es bueno o malo y el nivel cultural (Palacios,1987). Para ello se seleccionó una muestra de 600 madres y padres de la Comunidad Autónoma Andaluza de la que se obtuvo la siguiente información:

	Bueno	Malo	Depende	No sabe	Marginal nivel
Bajo	6	158	31	4	199
Medio	0	136	61	3	200
alto	3	111	81	6	201
Marginal opinión	9	405	173	13	600

4.2. Tabla de correspondencias

- En la tabla anterior los valores que se encuentran en la intersección de cada fila y columna corresponden a las frecuencias absolutas (número de sujetos) de cada combinación de opinión por nivel cultural. A dichos valores se les denota genéricamente por n_{ij}. Donde *i* representa a las categorías de la variable representada en las filas y *j* a las categorías de la variable representada en las columnas.
- La tabla de frecuencias anterior se le denomina tabla de correspondencias o tabla de contingencia.
- Además de las frecuencias para cada combinación de las categorías de las variables en la tabla aparecen varios totales:
  - Marginal de fila: $n_{i.}$: son los totales de cada fila
  - Marginal de columna: $n_{.j}$: son los totales de cada columna
  - Total: N es la suma de las frecuencias absolutas de todas las casillas

4.2. Tabla de correspondencias

	1	2	3	4	Marginal fila
1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
Marginal columna	$n._1$	$n._2$	$n._3$	$n._4$	N

Matriz de frecuencias absolutas

4.2. Tabla de correspondencias

	1	2	3	4	Marginal fila
1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
Marginal columna	$f._1$	$f._2$	$f._3$	$f._4$	1

Matriz de frecuencias relativas

Dónde $f_{ij} = \frac{n_{ij}}{N}$

4.2. Tabla de correspondencias

La tabla de contingencia o correspondencias anterior es el resultado de multiplicar dos matrices de datos obtenidas a partir de la definición de I (i=1…I) variables binarias o dicotómicas correspondientes a las categorías de una de las variables incluidas en el análisis y J (j=1, …J) variables binarias correspondientes a las categorías de la segunda variable incluida en el análisis. Para los datos que estamos considerando las matrices serían:

$$X_f = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \ldots & \ldots & .. \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \qquad X_c = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ .. & .. & .. & .. \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

X_f es de orden 600x3. Las columnas corresponden a las tres categorías de la variable Nivel cultural (colocada en las filas de la tabla). La matriz X_c es de orden 600x4. Las columnas de corresponden a cuatro variables binarias (dicotómicas) definidas de las cuatro categorías de la variable opinión (colocada en las columnas). Multiplicando X'_aX_b (o bien X'_bX_a) sumamos a todos los padres y madres que tienen cada par de características y obtenemos la tabla de contingencia.

4.3. Dependencia e independencia en tablas de contingencia

El análisis clásico de la posible relación entre las variables cualitativas se realiza mediante una prueba de hipótesis nula. La H_0: establece que las variables son independientes,
La H_1: establece que las variables son dependientes. El estadístico de contraste es:

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ob} - n_{es}\right)_{ij}^2}{\left(n_{es}\right)_{ij}}$$

Donde n_{ob} son las frecuencias absolutas y n_{esp} las esperadas bajo la Ho.

4.3. Dependencia e independencia en tablas de contingencia

Las frecuencias esperadas se obtienen de la siguiente manera:

$$n_{esp} = \frac{n_{i.} \times n_{.j}}{N}$$

Para la tabla que estamos analizando las frecuencias observadas y esperadas Junto a los residuales tipificados vienen dadas en la siguiente tabla:

Tabla de contingencia nivel * ideas

			ideas				
			bueno	malo	depende	no sabe	Total
nivel	bajo	Recuento	6	158	31	4	199
		Frecuencia esperada	3,0	134,3	57,4	4,3	199,0
		Residuos corregidos	2,2	4,4	-5,0	-,2	
	medio	Recuento	0	136	61	3	200
		Frecuencia esperada	3,0	135,0	57,7	4,3	200,0
		Residuos corregidos	-2,1	,2	,6	-,8	
	alto	Recuento	3	111	81	6	201
		Frecuencia esperada	3,0	135,7	58,0	4,4	201,0
		Residuos corregidos	,0	-4,6	4,4	1,0	
Total		Recuento	9	405	173	13	600
		Frecuencia esperada	9,0	405,0	173,0	13,0	600,0

4.3. Dependencia e independencia en tablas de contingencia

El estadístico de contraste se distribuye con (I-1)x(J-1) grados de libertad y se rechaza la hipótesis nula si P<α. Los residuos tipificados corregidos se calculan con la expresión

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

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	37,250[a]	6	,000
Razón de verosimilitud	41,047	6	,000
Asociación lineal por lineal	27,371	1	,000
N de casos válidos	600		

a. 6 casillas (50,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 2,99.

P<α en consecuencia se rechaza la H_0

4.4. Perfiles marginales y condicionales

Los perfiles marginales describen la distribución marginal de las variables y vienen dados por:

Perfil Marginal de Fila
$n_{1.}/N = f_{1.}$ (199/600=0,3317)
$n_{2.}/N = f_{2.}$ (200/600=0,3333)
$n_{3.}/N = f3.$ (201/600=0,3350)
1

El conjunto de marginales fila ($f_{1.}$, $f_{2.}$, $f_{3.}$) corresponde a la columna promedio se le denomina centro de gravedad o centroide de las columnas.

4.4. Perfiles marginales y condicionales

Perfil Marginal Columna	$n._1/N=f._1$	$n._2/N=f._2$	$n._3/N=f._3$	$n._4/N=f._4$	1
	(9/600=0,02)	(405/600=0,67)	(173/600=0,29)	(13/600=0,02)	

El conjunto de marginales columna ($f._1$, $f._2$, $f._3$, $f._4$) corresponde a la fila promedio se le denomina centro de gravedad o centroide de las filas.

4.4. Perfiles marginales y condicionales

Los perfiles condicionales describen la distribución conjunta asociada a la tabla de correspondencias. Se pueden construir dos tablas de perfiles condicionales: la tabla de perfiles-fila que describe la distribución condicionada de las columnas para cada fila y se obtiene dividiendo las frecuencias absolutas de la tabla de contingencia entre los marginales de fila.

Así la tabla de perfiles-fila de opinión por nivel sería:

F(J/I) $\frac{f_{ij}}{f_{i.}} = \frac{n_{ij}}{n_{i.}}$	Bueno	Malo	Depende	No sabe	Marginal nivel
Bajo	6/199=0,03	158/199=0,79	31/199=0,16	4/199=0,02	1
Medio	0	136 /200=0,68	61/200=0,30	3/200=0,01	1
alto	3/201=0,01	111/201=0,55	81/201=0,40	6/201=0,30	1

Perfil Marginal Columna	0,01	0,67	0,29	0,02	

4.4. Perfiles marginales y condicionales

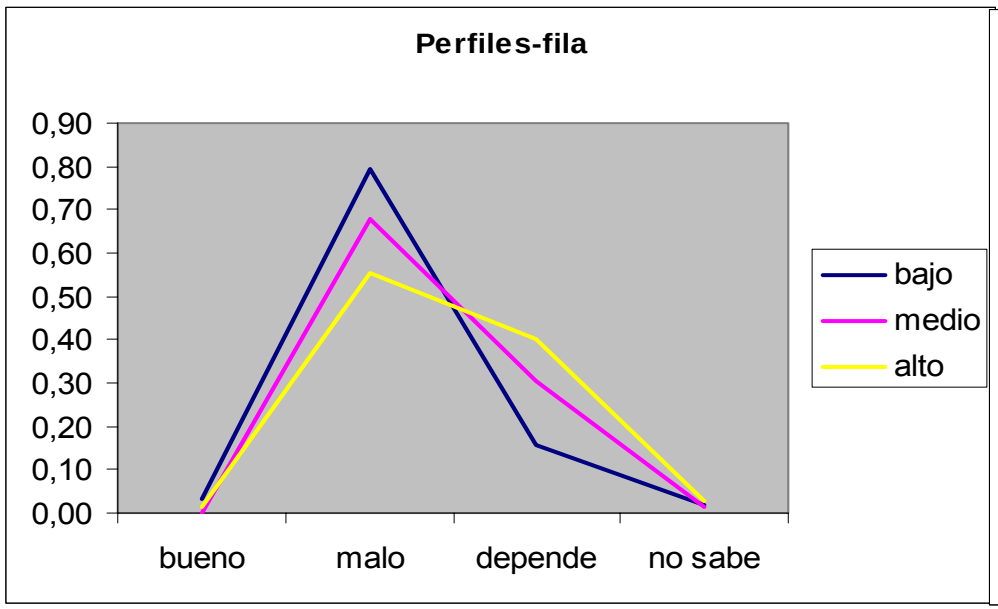
De la misma manera una tabla de perfiles-columna describe la distribución condicionada de la variable nivel para cada categoría de la variable opinión y se obtiene dividiendo las frecuencias absolutas entre los marginales de columna y
multiplicando por 100:

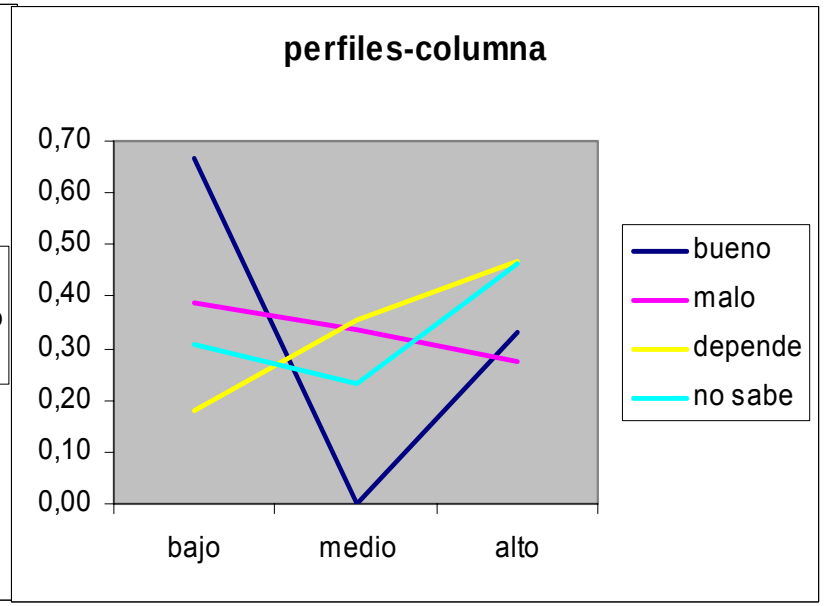
F(I/J) $\frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n_{.j}}$	Bueno	Malo	Depende	No sabe
Bajo	6/9=0,67	158/405=0,39	31/173=0,18	4/13=0,31
Medio	0	136/405=0,34	61/173=0,35	3/13=0,23
alto	3/9=0,33	111/405=0,27	81/173=0,47	6/13=0,46
	1	1	1	1

Perfil Marginal deFila
0,33
0,33
0,33

4.4. Perfiles marginales y condicionales

Para analizar el parecido de los perfiles-fila podemos construir un gráfico de líneas colocando en el eje de abscisas la variable opinión (J). De la misma manera Podemos representar los perfiles-columna colocando en el eje de abscisas el nivel cultural (I).





4.4. Perfiles marginales y condicionales

A la tabla (matriz) de perfiles-fila la denotamos como F. A la tabla (matriz) de perfiles-columna la denotamos como C. Las filas de F pueden considerarse como puntos en el espacio de las columnas (R^J). De la misma manera las columnas de C pueden considerarse como puntos en el espacio de las filas (R^I). Para medir la distancia entre los puntos representados indistintamente en el espacio R^J o R^I se utiliza la distancia la distancia P2.

La distancia P2 entre la fila 1 y la fila 2 de F (correspondientes a los perfiles de nivel bajo y medio) vendría dada por.

$$\chi^2_{(f_1, f_2)} = \sum_{j=1}^{J} \frac{\left(f_{1j} - f_{2j}\right)^2}{f_{.j}}$$

La matriz de distancias P2 obtenida a partir de F viene dada por:

$$D_f = \begin{pmatrix} 0 & 0{,}16 & 0{,}32 \\ 0{,}16 & 0 & 0{,}08 \\ 0{,}32 & 0{,}08 & 0 \end{pmatrix}$$

4.4. Perfiles marginales y condicionales

- De la misma manera podemos calcular la matriz de distancias P2 entre los perfiles columna. La distancia entre las columna 1 y 2 de C viene dada por:

$$\chi^2_{(c_1,c_2)} = \sum_{i=1}^{I} \frac{\left(c_{i1} - c_{i2}\right)^2}{f_{i.}}$$

y la matriz de distancia es:

$$D_c = \begin{pmatrix} 0 & 0{,}58 & 1{,}14 & 0{,}60 \\ 0{,}58 & 0 & 0{,}25 & 0{,}16 \\ 1{,}14 & 0{,}25 & 0 & 0{,}09 \\ 0{,}60 & 0{,}16 & 0{,}09 & 0 \end{pmatrix}$$

4.5. El análisis de correspondencias simple

En AC existe una matriz similar a la matriz de correlaciones o de varianzas covarianzas en AFE denominada matriz de dispersión o matriz de inercia. La matriz de inercia se obtiene multiplicando la matriz X cuyo término general es;

$$x_{ij} = \frac{f_{ij} - \left(f_{i.} \times f_{j.}\right)}{\sqrt{\left(f_{i.} \times f_{j.}\right)}}$$

por la transpuesta de X. La matriz de inercia a partir de las filas viene dada por

$$S_F = X' X$$

4.5. El análisis de correspondencias simple

La suma de los elementos de la diagonal de S equivale a la varianza inicial a factorizar y viene dada por

$$I = \sum_{i,j} \frac{\left(f_{ij} - f_{i.} \times f_{.j} \right)^2}{f_{i.} \times f_{.j}}$$

El resto de los elementos de S equivalen a las covarianzas.

4.5. El análisis de correspondencias simple

Análogamente la matriz de dispersión para las columnas se puede obtener mediante el producto

$$S_c = X\ X'$$

La suma de la inercia de las filas (traza de la matriz X'X) es igual a la suma de la inercia de las columnas (traza de la matriz XX') y como se puede comprobar la inercia es igual al estadístico P^2 dividido por el número de sujetos N.

Una vez obtenida la matriz de inercia, el AC es equivalente al ACP. Si recordamos, se obtiene la primera componente de manera que explique la máxima varianza, la segunda componente de manera que explique la máxima varianza de la restante y así hasta obtener tantas componentes como (J-1) (I-1).

4.5. El análisis de correspondencias simple

La suma de los elementos de la diagonal de S equivale a la varianza inicial a factorizar y viene dada por

$$I = \sum_{i,j} \frac{\left(f_{ij} - f_{i.} \times f_{.j} \right)^2}{f_{i.} \times f_{.j}}$$

El resto de los elementos de S equivalen a las covarianzas. Una vez obtenida la matriz de inercia, el AC es equivalente al ACP. Si recordamos, se obtiene la primera componente de manera que explique la máxima varianza, la segunda componente de manera que explique la máxima varianza de la restante y así hasta obtener tantas componentes como (J-1) (I-1).

4.6. El análisis de correspondencias múltiple

El ACM es una generalización del ACS al caso en el que se cruzan más de dos variables cualitativas. Se parte de una matriz de datos X con N filas y tantas columnas como la suma de las categorías del conjunto de variables a analizar (p). La columnas son variables binarias codificadas con 1 ó 0. El ACM analiza la llamada tabla de Burt cuya expresión es

$$B=X'X$$

El número de dimensiones máximo que se pueden extraer en ACM es Min {m, N-1} donde m es el número de categorías de las variables sin datos perdidos menos el número de dichas variables y N es el tamaño de la muestra.

4.7.Reglas de interpretación del AC

- Existe asociación entre variables si se rechaza la hipótesis nula de independencia. Aún sin rechazarse la existencia de grandes diferencias en los porcentajes de varianza explicada de los distintos factores se interpretaría en términos de asociación de variables.
- Buscamos los puntos (categorías) que más contribuyan (contribuciones absolutas) a la inercia de la dimensión.
- Buscamos los puntos (categorías) mejor explicados por un factor (contribuciones relativas). Cuanto mayor sea la contribución relativa mejor representada está la categoría en el factor.
- Los cosenos al cuadrado permiten saber si un punto está bien representado sobre el eje factorial. La calidad de la representación de un punto sobre el eje será tanto mayor cuando más próximo a 1 sea el coseno al cuadrado.
- La proximidad entre categorías de variables se interpreta en términos de asociación o dependencia. Una regla que se suele utilizar es que se pueden considerar categorías próximas aquellas que forman ángulos menores de 60 grados.
- Para interpretar los factores se buscan categorías contrapuestas.