La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 1

Diagramas de dispersión y mapas

Primera edición: julio 2008
ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana

www.fbbva.es

CAPÍTULO 1

Diagramas de dispersión y mapas

El análisis de correspondencias es un método de análisis de datos que representa gráficamente tablas de datos. El análisis de correspondencias es una generalización de una representación gráfica con la que todos estamos familiarizados, el *diagrama de dispersión*. Un diagrama de dispersión representa los datos en forma de puntos con relación a dos ejes de coordenadas perpendiculares: el eje horizontal, eje de las *x*, y el eje vertical, eje de las *y*. Para introducirnos poco a poco en el análisis de correspondencias, es conveniente que reflexionemos sobre lo que entendemos por diagrama de dispersión y sobre cómo interpretamos los datos que éste representa gráficamente. Haremos énfasis en cómo interpretar las distancias entre puntos y en averiguar cuándo podemos considerar que los diagramas de dispersión son *mapas de datos*.

Contenido

Conjunto de datos 1: mis viajes en 2005

A finales de 2005, cuando empecé a escribir este libro, reflexioné sobre los viajes que durante ese año había hecho a tres de mis países favoritos: Noruega, Canadá y Grecia. Según mi diario pasé 18 días en Noruega, 15 días en Canadá y 29 días en Grecia. Aparte de estas visitas, también hice algunos viajes cortos a Francia y a Alemania, en total 24 días. Podemos representar esta descripción numérica del tiempo que estuve de viaje en gráficos como los de la imagen 1.1. Este ejemplo, aparentemente trivial, esconde algunos conceptos importantes para la interpretación de gráficos en los que representamos los datos con relación a dos ejes de coordenadas, y que eventualmente nos pueden ayudar a comprender el análisis de correspondencias. Vamos a revisar estos conceptos uno a uno.

Variables continuas

El eje vertical situado a la izquierda, que hemos etiquetado como Días, es una escala con información numérica de una variable *continua*. La escala de este eje indica claramente el número de días que pasé en algunos países extranjeros. Hemos ordenado los valores numéricos desde 0 días, en la parte inferior de la escala, hasta 30 días en la parte superior de la misma. En el diagrama de barras situado a la derecha de la imagen 1.1, mostramos una representación gráfica muy habitual de datos, en la cual la longitud de las barras es proporcional a los valores de la variable. Hemos redondeado el tiempo que pasé en cada país a número de días, sin embargo, seguimos considerando esta variable como continua, ya que el tiempo es esencialmente una variable continua.

Expresión de los datos en valores relativos

El eje vertical situado a la derecha de los dos gráficos de la imagen 1.1 expresa el número de días de viaje en cada país, como porcentaje, con relación al total de mis 86 días de viaje. Por ejemplo, 18 días en Noruega corresponde al 21% del tiempo total. El total de 86 días es la *base* con relación a la cual expresamos los valores relativos de los datos. En este caso tenemos un solo conjunto de datos, y en consecuencia sólo una base. En estos dos gráficos podemos representar, en el mismo gráfico, la escala absoluta original de la izquierda y la escala de valores relativos de la derecha.

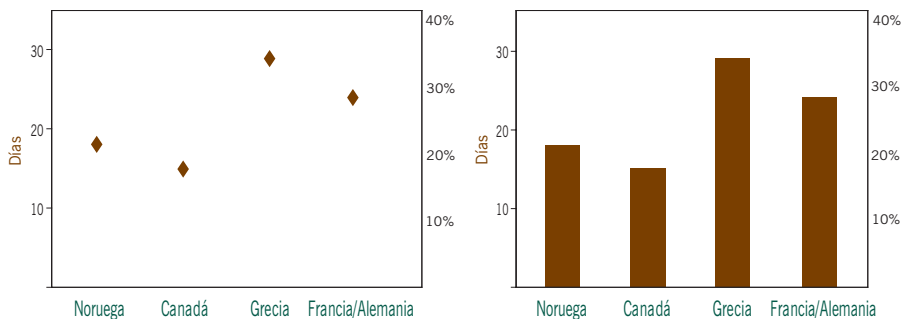


Imagen 1.1:
Gráficos sobre el número de días que pasé en países extranjeros en 2005, en forma de diagrama de dispersión y de diagrama de barras. A la derecha de cada gráfico, el eje vertical expresa el número de días en porcentaje con relación al total de 86 días de viaje

Variables categóricas

A diferencia del eje vertical, eje *y*, el eje horizontal, eje *x*, corresponde claramente a una variable no numérica. En este eje, los cuatro puntos son sólo posiciones en las que hemos situado las etiquetas que indican el país visitado. La escala horizontal representa una variable *categórica*. Hay dos características de este eje horizontal que no tienen significado sustantivo alguno en el gráfico: la ordenación de las categorías y la distancia entre ellas.

Ordenación de las categorías

En primer lugar, no hay ninguna razón de peso por la cual hayamos situado a Noruega en primer lugar, a Canadá en segundo y a Grecia en tercer lugar; quizás el hecho de que visité estos países por este orden. Como la etiqueta Francia/Alemania indica un conjunto de viajes cortos que realicé en distintos momentos del año, hemos situado esta etiqueta después de las otras. Sin embargo, en este tipo de representaciones gráficas en las que el orden es irrelevante, siempre es bueno reordenar las categorías de manera que tengan algún significado sustantivo, por ejemplo, los valores de la variable. Así, podríamos ordenar los países en orden descendiente de acuerdo con el tiempo que pasé en cada país. En tal caso habríamos situado los países en el siguiente orden: Grecia, Francia/Alemania, Noruega y Canadá. Esta sencilla reordenación facilita la interpretación de los datos, especialmente cuando tenemos muchos. Por ejemplo, si hubiera visitado 20 países distintos, la ordenación contendría información relevante que no obtendríamos de forma rápida a partir de la ordenación original.

Distancias entre las categorías

En segundo lugar, no existe razón alguna por la cual hayamos situado los cuatro puntos a intervalos iguales en el eje de las *y*. Asimismo, no existe tampoco razón por la cual hayamos de situarlos a intervalos distintos; en realidad los hemos situado a intervalos iguales por conveniencia y estética. Cuando utilicemos el análisis de correspondencias, veremos que existen distintas maneras de definir intervalos entre las categorías de las variables como la que acabamos de comentar. Es más, presentaremos el análisis de correspondencias como un procedimiento para la cuantificación de las categorías de una variable y, así, tanto las distancias entre categorías como su ordenación tendrán un significado importante.

Interpretación de las distancias en los diagramas de dispersión

En el eje horizontal del gráfico de la izquierda de la imagen 1.1, tanto la ordenación de los países como la separación entre éstos son arbitrarias, por tanto, no tiene ningún sentido que midamos e interpretemos las distancias entre los puntos mostrados en el gráfico de la izquierda. Dada la naturaleza numérica del eje vertical que indica frecuencia (o frecuencia relativa), las únicas medidas de distancia que tienen sentido son estrictamente las distancias en dirección vertical.

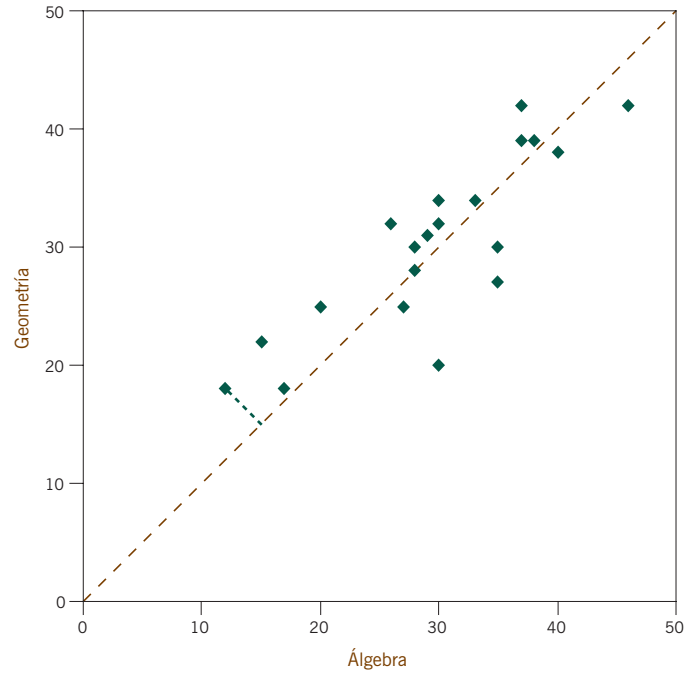


Imagen 1.2:
Diagrama de dispersión de las calificaciones de 20 estudiantes en dos materias (álgebra y geometría) en un examen de matemáticas. Los puntos tienen propiedades especiales. Así podemos obtener la calificación total de los estudiantes proyectando los puntos perpendicularmente sobre la bisectriz que hemos calibrado de 0 (abajo a la izquierda) a 100 (arriba a la derecha)

Los diagramas de dispersión como mapas

En algunos casos especiales, las dos variables que definen a los ejes de los diagramas de dispersión tienen la misma naturaleza numérica y escalas similares. Por ejemplo, supongamos que 20 estudiantes han realizado un examen de matemáticas que consta de dos partes, álgebra y geometría. Supongamos que cada parte representa el 50% de la nota final. En la imagen 1.2, hemos representado gráficamente los pares de calificaciones de los estudiantes. Es importante que los dos ejes, que representan las respectivas calificaciones, tengan escalas con unidades de la misma longitud. Dada la naturaleza similar de las dos variables y de sus dos escalas, en esta representación gráfica podemos medir distancias en cualquier dirección; no solamente horizontal o verticalmente —igual que en un mapa en el que podemos medir distancias entre poblaciones—. Dos puntos que se hallen cerca tendrán calificaciones similares. Por tanto, tiene sentido que nos fijemos en la forma de la distribución de los puntos y, en particular, remarcar que hay un pequeño grupo de cuatro estudiantes con calificaciones elevadas y sólo un estudiante con calificaciones muy elevadas. Podemos considerar la imagen 1.2 un *mapa*, ya que las posiciones de los estudiantes vienen definidas por posiciones bidimensionales, de la misma manera que, en una región, las localizaciones geográficas vienen definidas por la longitud y la latitud.

Calibración de una dirección en un mapa

Los mapas tienen interesantes propiedades geométricas. Por ejemplo, en la imagen 1.2, la bisectriz, que hemos representado como una línea discontinua, define un eje que expresa las calificaciones finales de los estudiantes, combi-

nando las calificaciones de álgebra y de geometría. Si calibramos este eje de 0 (abajo izquierda) hasta 100 (arriba a la derecha), podemos leer las calificaciones finales de los estudiantes en el mapa, proyectando de forma perpendicular sobre el mencionado eje los puntos que representan sus calificaciones. En la representación gráfica podemos ver un ejemplo para un estudiante que obtuvo 12 puntos sobre 50 en álgebra y 18 sobre 50 en geometría. A la proyección de este punto sobre la bisectriz, de coordenadas 15 y 15, le corresponde una calificación final de 30.

Transformación de la información en la representación gráfica

Los diagramas de dispersión de las imágenes 1.1 y 1.2 son dos maneras distintas de expresar, de forma gráfica, la información numérica contenida en dos tablas que contienen datos sobre viajes y calificaciones, respectivamente. En ambos casos, no hay pérdida de información entre los datos y las representaciones gráficas. Dados los gráficos, es fácil recuperar exactamente la información numérica. Decimos que los diagramas de dispersión o los mapas son «instrumentos de transformación de la información» en los que, en absoluto, se produce un procesado de los datos; simplemente expresamos los datos de forma visual, es decir, se trata de una manifestación alternativa de la misma información.

Variables nominales y variables ordinales

En el ejemplo sobre mis viajes, la variable categórica «país» tiene cuatro categorías, y dado que no existe una ordenación intrínseca de las categorías, llamamos a esta variable *nominal*. En cambio, si podemos ordenar de forma natural las categorías de una variable categórica, llamamos a la variable *ordinal*. Por ejemplo, podemos clasificar los días en tres categorías de acuerdo con el tiempo que dediqué cada día a trabajar: a) menos de una hora («festivos»), b) más de una pero menos de seis horas («medias jornadas») y c) más de seis horas («jornadas completas»). Por tanto, hemos ordenado estas categorías de acuerdo con una variable continua «tiempo diario de trabajo» que hemos dividido en intervalos. Tendremos en cuenta esta ordenación en cualquier representación gráfica de las variables. En muchas encuestas sociales, se dan las respuestas en una escala ordinal. Por ejemplo, una escala ordinal sobre valoración de la importancia: nada importante/algo importante/muy importante. Otro ejemplo típico es la escala de acuerdo/desacuerdo: muy de acuerdo/algo de acuerdo/ni de acuerdo ni en desacuerdo/algo en desacuerdo/muy en desacuerdo. Aquí la posición ordinal de la categoría «ni de acuerdo ni en desacuerdo» puede no estar situada entre «algo de acuerdo» y «algo en desacuerdo», podría ser, por ejemplo, una categoría utilizada por algunos encuestados para expresar que «no sabe» cuando éstos o bien no comprenden la pregunta o bien no tienen una respuesta clara. Veremos este tema más adelante (cap. 21), una vez hayamos desarrollado las herramientas que nos permitan estudiar las asociaciones entre las respuestas en cuestionarios de datos multivariantes.

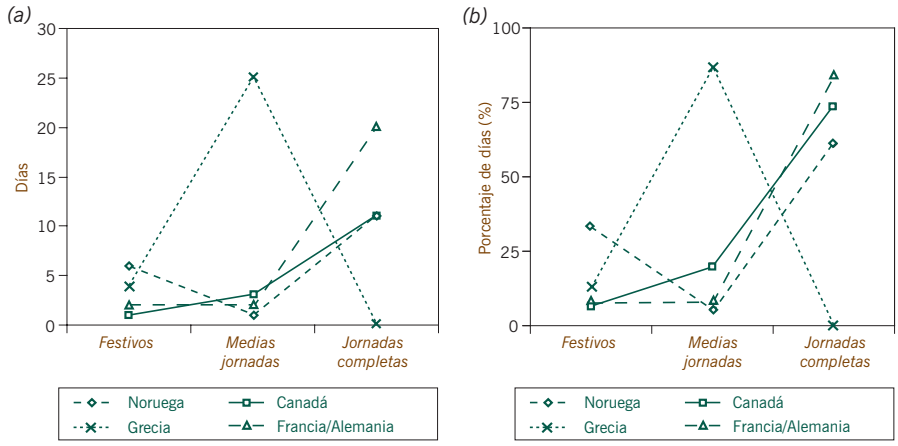
Imagen 1.3: *Frecuencias de los tipos de día en los cuatro viajes*

País	*Festivos*	*Medias jornadas*	*Jornadas completas*	*Total*
Noruega	6	1	11	*18*
Canadá	1	3	11	*15*
Grecia	4	25	0	*29*
Francia/Alemania	2	2	20	*24*
Total	*13*	*31*	*42*	*86*

Representación gráfica de más de un conjunto de datos

Supongamos que clasificamos mis 86 días de viaje en el extranjero de acuerdo con las categorías, *festivos*, *medias jornadas* y *jornadas completas*. En la imagen 1.3 se muestra una tabla que corresponde a la *clasificación cruzada* de país por tipo de día. Podemos ver esta tabla de dos formas distintas: como un conjunto de filas o como un conjunto de columnas. En este caso, las columnas son conjuntos de frecuencias que caracterizan a los respectivos tipos de día; mientras que las filas caracterizan a los respectivos países. En la figura (a) de la imagen 1.4, se muestra un diagrama de frecuencias de los distintos países (filas), en el que hemos situado el tipo de día (las columnas) en el eje horizontal. Dado que hemos ordenado las categorías de la variable «tipo de día», tiene sentido unir los valores de las categorías de esta variable mediante líneas. Sin embargo, si queremos comparar los países entre sí, hemos de tener en cuenta que el número de días que pasé en cada país no fue el mismo. El número total de días que pasé en cada país nos proporciona una base sobre la que podemos reexpresar los valores de las filas de la imagen 1.3, como porcentajes con relación a estos totales (imagen 1.5). En la representación gráfica de la imagen 1.4(b), hemos visualizado estos porcentajes, y ahora sí podemos comparar los tipos de día de los distintos viajes.

Imagen 1.4: *Diagramas de frecuencias absolutas (a) y de frecuencias relativas (b), expresadas como porcentajes de las filas de la imagen 1.3*



País	*Festivos*	*Medias jornadas*	*Jornadas completas*
Noruega	33%	6%	61%
Canadá	7%	20%	73%
Grecia	14%	86%	0%
Francia/Alemania	8%	8%	83%
Global	*15%*	*36%*	*49%*

Imagen 1.5: *Porcentajes correspondientes a los tipos de día en cada país, así como los porcentajes globales de los países, donde la suma de los valores de las filas es el 100%*

Interpretación de las frecuencias absolutas y de las frecuencias relativas

De estas representaciones gráficas tenemos que extraer una lección fundamental para el análisis de frecuencias de datos. Cada viaje ha implicado un diferente número de días y, por tanto, corresponde a una base distinta sobre la que expresar la frecuencia de los tipos de día. Sólo podemos comparar los 6 *festivos* en Noruega, con los 4 en Grecia, con relación al número total de días que pasé en cada uno de estos países. Como porcentajes, estos valores se transforman en valores muy distintos; 6 de 18 es el 33%, mientras que 4 de 29 es el 14%. La visualización de las frecuencias relativas de la imagen 1.4(b) nos permite una comparación más precisa de cómo pasé mi tiempo en los diferentes países. También podemos expresar las frecuencias «marginales» (18, 15, 29 y 24, de los países y 13, 31, 42 del tipo de día) con relación a sus respectivos totales (por ejemplo, en la última fila de la imagen 1.5 mostramos los porcentajes correspondientes al tipo de día para la combinación de todos los países). Estas frecuencias marginales relativas, también las podíamos haber representado en la imagen 1.4 (b).

Descripción e interpretación de los datos versus modelización e inferencia estadística

Cualquier conclusión que hayamos sacado sobre la posición de los puntos de la imagen 1.4(b) es sólo una interpretación de los datos, no es una afirmación sobre la significación estadística de lo que hemos observado. Estos aspectos estadísticos de las representaciones gráficas, los veremos solamente al final del libro (cap. 25). Por tanto, en la mayor parte del libro nos concentraremos en la descripción y en la interpretación de los datos. La deducción de que, en proporción, pasé más días festivos en Noruega que en ningún otro país es ciertamente verdadera, lo podemos ver en la imagen 1.4(b). Sin embargo, analizar si este fenómeno es estadísticamente comparable con un modelo o con una hipótesis sobre mi comportamiento que, por ejemplo, postule que la proporción de festivos fue la misma en todos mis viajes, es un tema completamente distinto. Gran parte de la metodología estadística existente se concentra en saber si los datos se ajustan, o se pueden comparar, con un determinado modelo teórico o con una hipótesis preconizada. Se dedica poca atención a desarrollar procedimientos para describir datos, para interpretarlos o para generalizar hipótesis. Un ejemplo típico, en ciencias sociales, es la utilización omnipresente del estadístico ji-cuadrado para contrastar asociaciones en tablas de contingencia. A menudo se hallan asociaciones estadísticamente significativas, pero en cambio no existen herramientas

sencillas para detectar qué partes de la tabla son las responsables de esta asociación. El análisis de correspondencias es una herramienta que puede contribuir a rellenar este vacío. Permite al analista visualizar las asociaciones existentes en los datos, y en consecuencia le permite formular hipótesis que éste puede contrastar en una etapa más avanzada de su investigación. En la mayor parte de las situaciones, podemos describir, interpretar y modelizar los datos. De todas formas, existen situaciones en las que la descripción y la interpretación de los datos tiene, por sí misma, una importancia capital, por ejemplo, cuando los datos representan a la totalidad de la población de interés.

Conjuntos de datos grandes

A medida que las tablas de datos aumentan de tamaño, debido al excesivo número de puntos, se hace difícil representar éstos de forma simple, como hemos hecho, por ejemplo, en la imagen 1.4. Supongamos que durante un año hubiera visitado 20 países, al clasificar el tiempo pasado en cada uno de ellos, hubiese obtenido una tabla de contingencia con muchas más filas. También podría haber registrado otros datos, como por ejemplo la meteorología de cada día («buen tiempo», «parcialmente nublado» o «lluvioso»), con el objetivo de estudiar posibles relaciones con el tipo de día. Tendría, pues, una tabla de datos con muchas más columnas y muchas más filas. Representar, de la misma manera como hemos hecho en la imagen 1.4, a los 20 conjuntos de puntos clasificados en muchas más categorías podría llevarnos a una gran confusión entre puntos y etiquetas. Resultaría absolutamente imposible identificar pauta alguna. Por tanto, en estas situaciones para resaltar las características esenciales de esos datos, tendríamos que buscar una alternativa a los diagramas de dispersión, el instrumento para la descripción de datos que hemos utilizado hasta ahora. Tal como veremos en el libro, el análisis de correspondencias, un método de representación gráfica de datos igual que los diagramas de dispersión, nos permitirá trabajar fácilmente con conjuntos de datos grandes.

RESUMEN: Diagramas de dispersión y mapas

1. Los diagramas de dispersión representan gráficamente dos variables con relación a un eje horizontal y un eje vertical, el eje *x* y el eje *y*, respectivamente.
2. A menudo, la naturaleza de la variable *x* es completamente distinta a la de la variable *y*, de manera que solamente podemos interpretar distancias en la dirección de unos de los dos ejes, de acuerdo con una determinada escala de medida con la que hayamos calibrado el eje. En estas situaciones, no tiene sentido medir o interpretar distancias en cualquier otra dirección del gráfico.
3. El algunos casos, las variables *x* e *y* son de naturaleza similar con escalas de medida comparables. En estas situaciones, podemos interpretar las distancias entre los puntos como una medida de la diferencia, o de la disimilitud, entre los puntos representados. En estos casos especiales consideramos que los diagramas de dispersión son *mapas*.

4. Cuando representamos valores positivos (en general, en nuestro contexto, frecuencias), estamos interesados tanto en los valores relativos como en los absolutos.
5. Cuanto más complejos sean los datos, menos conveniente será representarlos en forma de diagramas de dispersión.
6. Este libro, más que sobre la modelización de información compleja, trata sobre la descripción y la interpretación de la información.