Capítulo 2

Medidas de Asociación.

2.1. Introducción.

Una vez considerado que el objetivo del Análisis Cluster consiste en encontrar agrupaciones naturales del conjunto de individuos de la muestra, es necesario definir qué se entiende por agrupaciones naturales y, por lo tanto, con arreglo a qué criterio se puede decir que dos grupos son más o menos similares. Esta cuestión conlleva otras dos, a saber:

1. Cómo se puede medir la similitud entre dos individuos de la muestra.
2. Cómo se puede evaluar cuándo dos clusters pueden ser o no agrupados.

A continuación vamos a centrarnos en las posibles funciones que pueden elegirse para medir la similitud entre los grupos que sucesivamente se van formando, distinguiendo primeramente entre distancias métricas y similaridades.

2.2. Distancias y Similaridades. Definiciones preliminares.

2.2.1. Distancias.

Definición 2.1 *Sea U un conjunto finito o infinito de elementos. Una función $d : U \times U \longrightarrow \mathbb{R}$ se llama una distancia métrica si $\forall x, y \in U$ se tiene:*

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \Leftrightarrow \ x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)\ ,\ \forall z \in U$

Comentario 2.2.1

La definición anterior de distancia métrica puede exponerse sin necesidad de tantos axiomas. En efecto se puede comprobar que una distancia métrica es una función $d : U \times U \longrightarrow \mathbb{R}$ que verifica los siguientes axiomas

1. $d(x, y) = 0 \ \Leftrightarrow \ x = y$
2. $d(y, z) \leq d(x, y) + d(x, z)\ ,\ \forall x, y, z \in U$

Comentario 2.2.2

Ciertos autores realizan una cierta distinción entre lo que es una función distancia y lo que es una distancia métrica. Para ello definen una distancia como aquella función $d : U \times U \longrightarrow \mathbb{R}$ que verifica

1. $d(x, y) \geq 0$
2. $d(x, x) = 0$

3. $d(x,y) = d(y,x)$

y reservan el nombre de distancia métrica a aquellas distancias que además verifican

1. $d(x,y) = 0 \implies x = y$
2. $d(x,z) \leq d(x,y) + d(y,z)\ ,\ \forall z \in U$

Comentario 2.2.3

Extendiendo el concepto clásico de distancia plasmado anteriormente, algunos autores definen distancias métricas que pueden tomar valores negativos. De esta manera una función distancia métrica sería una función $d : U \times U \longrightarrow \mathbb{R}$ *tal que cumple los siguientes axiomas*

1. $d(x,y) \geq d_0$
2. $d(x,y) = d_0 \iff x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)\ ,\ \forall z \in U$

donde d_0 *puede ser menor que cero. Tal definición la realizan amparándose en el hecho de que, dada una tal función distancia métrica d, se puede definir otra* $d^{'}$ *a partir de ella, de la forma* $d^{'}(x,y) = d(x,y) - d_0$*, demostrándose fácilmente que* d' *es una distancia métrica en el sentido expuesto en la definición 2.1*

Comentario 2.2.4

1. *Una función que verifique los tres primeros apartados de la definición 2.1, pero no así la desigualdad triangular, es llamada semimétrica.*
2. *Se llama ultramétrica a toda métrica que verifique adicionalmente la propiedad*

$$d(x,z) \leq \text{Max}\,\{d(x,y), d(y,z)\}$$

2.2.2. Similaridades.

De forma similar a las distancias, tenemos la siguiente definición de similaridad

Definición 2.2 *Sea U un conjunto finito o infinito de elementos. Una función* $s : U \times U \longrightarrow R$ *se llama similaridad si cumple las siguientes propiedades:* $\forall x, y \in U$

1. $s(x,y) \leq s_0$
2. $s(x,x) = s_0$
3. $s(x,y) = s(y,x)$

donde s_0 *es un número real finito arbitrario.*

Definición 2.3 *Una función s, verificando las condiciones de la definición 2.2, se llama similaridad métrica si, además, verifica:*

1. $s(x,y) = s_0 \implies x = y$
2. $|s(x,y) + s(y,z)|s(x,z) \geq s(x,y)s(y,z)\ ,\ \forall z \in U$

Notemos que el segundo apartado de la definición anterior corresponde al hecho de que la máxima similaridad sólo la poseen dos elementos idénticos.

En las siguientes secciones expondremos algunas de las distancias y similaridades más usuales en la práctica.

Consideraremos, en general, m individuos sobre los cuales se han medido n variables $X_1, \ldots, X_n$. Con ello tenemos $m \times n$ datos que colocaremos en una matriz $m \times n$ dimensional

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

Distinguiremos entre medidas de asociación para individuos y para variables, aunque, técnicamente hablando, son válidas tanto para individuos como para variables (basta, para ello, considerar dichas medidas en un espacio n-dimensional o m-dimensional, esto es, trasponer la matriz).

2.3. Medidas de asociación entre variables.

Para poder *unir* variables es necesario tener algunas medidas numéricas que caractericen las relaciones entre variables. La base de trabajo de todas las técnicas cluster es que las medidas numéricas de asociación sean comparables, esto es, si la medida de asociación de una par de variables es 0,72 y el de otro par es 0,59, entonces el primer par está más fuertemente asociado que el segundo. Por supuesto, cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando.

2.3.1. Coseno del ángulo de vectores.

Consideremos dos variables X_i y X_j, muestreadas sobre m individuos, y sean x_i y x_j los vectores cuyas $k-$ésimas componentes indiquen el valor de la variable correspondiente en el $k-$ésimo individuo:

$$x_i = (x_{1i}, \ldots, x_{mi})^{'} \quad ; \quad x_j = (x_{1j}, \ldots, x_{mj})^{'}$$

Como es conocido, el producto escalar de dos vectores es:

$$x_i^{'} x_j = \sum_{l=1}^{m} x_{li} x_{lj}$$

que en Estadística se conoce como la suma de los productos cruzados entre x_i y x_j, mientras que el producto escalar de un vector por sí mismo, norma al cuadrado del vector, se llama suma de cuadrados.

Así se tiene:

$$x_i^{'} x_j = ||x_i|| ||x_j|| \cos(\beta) \tag{2.1}$$

donde β es el ángulo entre los vectores x_i y x_j.

Observando la figura (2.1), la distancia desde el origen (O) a B vale $||x_i|| \cos(\beta)$, siendo esta cantidad la proyección ortogonal de x_i sobre x_j. Así el producto escalar puede interpretarse como el producto de la longitud del vector x_j por la longitud de la proyección de x_i sobre x_j.

A partir de (2.1) se tiene

$$\cos(\beta) = \frac{x_i^{'} x_j}{||x_i||\ ||x_j||} = \frac{\sum_{l=1}^{m} x_{li} x_{lj}}{\left(\sum_{l=1}^{m} x_{li}^2 \sum_{l=1}^{m} x_{lj}^2 \right)^{\frac{1}{2}}} \tag{2.2}$$

El coseno del ángulo es una medida de similaridad entre x_i y x_j, con valores entre -1 y 1 en virtud de la desigualdad de Schwarz. Además es la mejor
medida para establecer el paralelismo entre dos vectores, ya que dos vectores son paralelos cuando el coseno del ángulo que forman es uno en valor absoluto.

Figura 2.1: Ángulo entre vectores

Esta medida es independiente, salvo signo, de la longitud de los vectores considerados. Algebráicamente, sean b y c dos escalares cualesquiera y definamos

$$\widehat{x}_i = bx_i \quad ; \quad \widehat{x}_j = cx_j \quad ; \quad b, c \neq 0$$

Entonces:

$$\cos(\widehat{x}_i, \widehat{x}_j) = \frac{\widehat{x}_i' \widehat{x}_j}{||\widehat{x}_i|| ||\widehat{x}_j||} = \frac{\sum_{l=1}^{m} bx_{li} cx_{lj}}{\left(\sum_{l=1}^{m} b^2 x_{li}^2 \sum_{l=1}^{m} c^2 x_{lj}^2 \right)^{\frac{1}{2}}} =$$

$$= \frac{bc \sum_{l=1}^{m} x_{li} x_{lj}}{|bc| \left(\sum_{l=1}^{m} x_{li}^2 \sum_{l=1}^{m} x_{lj}^2 \right)^{\frac{1}{2}}} \operatorname{Sgn}(bc) \cos(x_i, x_j)$$

con lo cual el coseno entre x_i y x_j es invariante ante homotecias, excepto un eventual cambio de signo.

2.3.2. Coeficiente de correlación.

Consideremos ahora las variables X_i y X_j, anteriores y centrémoslas respecto de sus medias, obteniendo unas nuevas variables cuyos valores para la muestra de los m individuos serán

$$\widehat{x}_i = (x_{1i} - \overline{x}_i, \ldots, x_{mi} - \overline{x}_i)' \quad ; \quad \widehat{x}_j = (x_{1j} - \overline{x}_j, \ldots, x_{mj} - \overline{x}_j)'$$

El producto escalar de las dos variables $\widehat{x}_i$ y $\widehat{x}_j$ se llama dispersión (scatter en la literatura anglosajona) de x_i y x_j. El producto escalar de $\widehat{x}_i$ por sí mismo es llamado la dispersión de x_i o la suma de los cuadrados de las desviaciones respecto a $\overline{x}_i$. Dividiendo por m ambas expresiones obtenemos la covarianza y la varianza, respectivamente.

$$\operatorname{Cov}(x_i, x_j) = \frac{\widehat{x}_i' \widehat{x}_j}{m} = \frac{1}{m} \sum_{l=1}^{m} (x_{li} - \overline{x}_i)(x_{lj} - \overline{x}_j)$$

$$\operatorname{Var}(x_i) = \frac{\widehat{x}_i' \widehat{x}_i}{m} = \frac{1}{m} \sum_{l=1}^{m} (x_{li} - \overline{x}_i)^2$$

La correlación muestral entre x_i y x_j se define como

$$r = \frac{\operatorname{Cov}(x_i, x_j)}{(\operatorname{Var}(x_i) \operatorname{Var}(x_j))^{\frac{1}{2}}} = \frac{\sum_{l=1}^{m} (x_{li} - \overline{x}_i)(x_{lj} - \overline{x}_j)}{\left(\sum_{l=1}^{m} (x_{li} - \overline{x}_i)^2 \sum_{l=1}^{m} (x_{lj} - \overline{x}_j)^2 \right)^{\frac{1}{2}}} \tag{2.3}$$

lo cual muestra que r es el coseno del ángulo entre los vectores centrados $\widehat{x}_i$ y $\widehat{x}_j$.

Alternativamente, si se tipifican las variables anteriores:

$$x^*_{li} = \frac{x_{li} - \overline{x}_i}{(\mathrm{Var}(x_i))^{\frac{1}{2}}} \quad ; \quad l = 1, \ldots, m$$

$$x^*_{lj} = \frac{x_{lj} - \overline{x}_j}{(\mathrm{Var}(x_j))^{\frac{1}{2}}} \quad ; \quad l = 1, \ldots, m$$

entonces la correlación entre x_i y x_j es la covarianza entre x^*_i y x^*_j.

Puesto que el coeficiente de correlación es el coseno del ángulo entre los vectores centrados, posee la propiedad vista con anterioridad, de invarianza, salvo signo, del coseno. Además, es invariante a las adiciones de una constante a cada elemento de x_i y x_j. En efecto, si llamamos $\widetilde{x}_i = x_i + b$, se tiene:

$$\widetilde{x}_i - \overline{\widetilde{x}}_i = (x_i + b) - (\overline{x}_i + b) = x_i - \overline{x}_i$$

Por ello, la correlación es invariante frente a transformaciones lineales, excepto posibles cambios de signo.

Así se observa que el coeficiente de correlación posee una invarianza más restrictiva que el coseno. Sin embargo, esta propiedad indica que el coeficiente de correlación discrimina menos que el coseno a la hora de establecer diferencias entre dos variables, ya que, dadas dos variables X e Y, hay muchos más elementos en la clase de equivalencia de todas las transformaciones lineales de X e Y que en la clase de equivalencia de las homotecias de X e Y.

La diferencia esencial entre las dos medidas (ángulo entre variables y el coeficiente de correlación) es que el coseno se basa en los datos originales y por ende emplea las desviaciones al origen mientras que el coeficiente de correlación usa los datos centrados y por tanto emplea las desviaciones respecto a la media. Si el origen está bien establecido y tiene sentido, entonces los datos originales tienen sentido de forma absoluta y el coseno es una medida apropiada de asociación; si, por el contrario, el origen es arbitrario o elegido a conveniencia, entonces los datos originales tienen sentido relativo respecto a su media, pero no respecto al origen. En tal caso es más apropiado el coeficiente de correlación.

2.3.3. Medidas para datos binarios o dicotómicos.

En ocasiones encontramos variables que pueden tomar dos valores (blanco-negro, si-no, hombre-mujer, verdadero-falso, etc). En tales casos se emplea, con frecuencia, el convenio de usar los valores dicotómicos 1 y 0 para ambos valores.

Al relacionar dos variables binarias, se forma una tabla de contingencia 2×2, que se puede esquematizar de la forma

$X_i \backslash X_j$	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$m = a+b+c+d$

(2.4)

En la anterior tabla se tiene:

1. a representa el número de individuos que toman el valor 1 en cada variable de forma simultánea.
2. b indica el número de individuos de la muestra que toman el valor 1 en la variable X_i y 0 en la X_j.
3. c es el número de individuos de la muestra que toman el valor 0 en la variable X_i y 1 en la X_j.
4. d representa el número de individuos que toman el valor 0 en cada variable, al mismo tiempo.
5. $a + c$ muestra el número de veces que la variable X_j toma el valor 1, independientemente del valor tomado por X_i.
6. $b + d$ es el número de veces que la variable X_j toma el valor 0, independientemente del valor tomado por X_i.
7. $a + b$ es el número de veces que la variable X_i toma el valor 1, independientemente del valor tomado por X_j.
8. $c + d$ es el número de veces que la variable X_i toma el valor 0, independientemente del valor tomado por X_j.

A continuación presentamos la versión binaria de las medidas introducidas anteriormente.

Medida de Ochiai

En el caso particular de variables dicotómicas, se tiene

$$x_i^{'}x_j = \sum_{l=1}^{m} x_{li}x_{lj} = a$$

$$x_i^{'}x_i = ||x_i||^2 = \sum_{l=1}^{m} x_{li}^2 = a+b$$

$$x_j^{'}x_j = ||x_j||^2 = \sum_{l=1}^{m} x_{lj}^2 = a+c$$

con lo cual el coseno del ángulo entre x_i y x_j queda en la forma:

$$\frac{a}{[(a+b)(a+c)]^{\frac{1}{2}}} = \left[\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)\right]^{\frac{1}{2}} \tag{2.5}$$

medida que es atribuida al zoólogo japonés Ochiai.

En el proceso seguido con las variables dicotómicas puede surgir una situación ambigua, como es el hecho de por qué y cómo asignar los valores 1 y 0 a los valores binarios. Puede ocurrir el caso de que intercambiando los papeles de dichos valores se llegue a resultados distintos, lo cual no es deseable. Por ello, en ocasiones, se toma la media geométrica de los cosenos obtenidos tomando ambos criterios y, más concretamente, se toma el cuadrado de dicha media geométrica, obteniéndose:

$$\left[\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)\left(\frac{d}{b+d}\right)\left(\frac{d}{c+d}\right)\right]^{\frac{1}{2}} \tag{2.6}$$

Hagamos notar que cada uno de los términos de la expresión anterior es una probabilidad condicionada. Así

1. $\frac{a}{a+b}$ es la probabilidad condicionada de que un individuo tome el valor 1 en la variable X_j dado que ha tomado el valor 1 en la variable X_i.
2. $\frac{a}{a+c}$ es la probabilidad condicionada de que un individuo tome el valor 1 en la variable X_i dado que ha tomado el valor 1 en la variable X_j.
3. $\frac{d}{b+d}$ es la probabilidad condicionada de que un individuo tome el valor 0 en la variable X_i dado que ha tomado el valor 0 en la variable X_j.
4. $\frac{d}{c+d}$ es la probabilidad condicionada de que un individuo tome el valor 0 en la variable X_j dado que ha tomado el valor 0 en la variable X_i.

De esta forma, la medida de Ochiai es la media geométrica de las probabilidades condicionadas asociadas con la celda con el valor a, mientras que la expresión (2.6) muestra el cuadrado de la media geométrica de las probabilidades condicionadas asociadas con la diagonal de la tabla (2.4).

Medida Φ

Esta medida se obtiene haciendo uso del coeficiente de correlación sobre dos variables dicotómicas.

$$r = \frac{\sum_{l=1}^{m} x_{li}x_{lj} - \frac{1}{m}\sum_{l=1}^{m} x_{li}\sum_{l=1}^{m} x_{lj}}{\left[\left\{\sum_{l=1}^{m} x_{li}^2 - \frac{1}{m}\left(\sum_{l=1}^{m} x_{li}\right)^2\right\}\left\{\sum_{l=1}^{m} x_{lj}^2 - \frac{1}{m}\left(\sum_{l=1}^{m} x_{lj}\right)^2\right\}\right]^{\frac{1}{2}}}$$

y teniendo en cuenta que

$$\sum_{l=1}^{m} x_{li}x_{lj} = a \qquad \sum_{l=1}^{m} x_{li} = a+b \qquad \sum_{l=1}^{m} x_{lj} = a+c$$

$$\sum_{l=1}^{m} x_{li}^2 = a+b \qquad \sum_{l=1}^{m} x_{lj}^2 = a+c$$

se tiene

$$r = \frac{a - \dfrac{(a+b)(a+c)}{m}}{\left[\left\{(a+b) - \dfrac{(a+b)^2}{m}\right\}\left\{(a+c) - \dfrac{(a+c)^2}{m}\right\}\right]^{\frac{1}{2}}} =$$

$$= \frac{am - (a+b)(a+c)}{\left[(a+b)\left\{m-(a+b)\right\}(a+c)\left\{m-(a+c)\right\}\right]^{\frac{1}{2}}} =$$

$$= \frac{ad-bc}{\left[(a+b)(c+d)(a+c)(b+d)\right]^{\frac{1}{2}}} \quad ; \quad (m = a+b+c+d) \tag{2.7}$$

Notemos, para finalizar, que, puesto que r es invariante bajo transformaciones lineales, los valores 0 y 1 son arbitrarios, ya que pueden ser transformados de forma lineal a otro par de valores.

Medidas basadas en coincidencias

Una forma intuitiva de medir la similaridad en variables dicotómicas es contar el número de veces que ambas variables toman el mismo valor de forma simultánea. Con ello dos variables serían más parecidas en tanto en cuanto mayor fuera el número de coincidencias a lo largo de los individuos.

No obstante, algunos factores influyen en las medidas que se pueden definir. Por ejemplo, una primera cuestión es qué hacer con las parejas del tipo $0-0$, ya que si las dicotomías son del tipo *presencia-ausencia*, los datos de la casilla d no poseen ningún atributo y no deberían tomar parte en la medida de asociación. Otra cuestión que surge es cómo ponderar las coincidencias y cómo las no coincidencias, o lo que es lo mismo, una diagonal u otra de la tabla (2.4).

A continuación exponemos algunas de las medidas que han ido surgiendo, atendiendo a varios criterios como los anteriormente expuestos.

1. **Medida de Russell y Rao**

$$\frac{a}{a+b+c+d} = \frac{a}{m} \tag{2.8}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar tenga el valor 1 en ambas variables. Notemos que este coeficiente excluye la pareja $0-0$, al contar el número de coincidencias pero no lo hace así al contar el número de posibles parejas. Asimismo, esta medida proporciona igual peso a las coincidencias y a las no coincidencias.

2. **Medida de parejas simples**

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{m} \tag{2.9}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar presente una coincidencia de cualquier tipo, pesando de igual forma las coincidencias y las no coincidencias.

3. **Medida de Jaccard**

$$\frac{a}{a+b+c} \tag{2.10}$$

Esta medida mide la probabilidad condicionada de que un individuo elegido al azar presente un 1 en ambas variables, dado que las coincidencias del tipo $0-0$ han sido descartadas primero y por lo tanto han sido tratadas de forma irrelevante.

4. **Medida de Dice**

$$\frac{2a}{2a+b+c} \tag{2.11}$$

Esta medida excluye el par $0-0$ de forma completa, pesando de forma doble las coincidencias del tipo $1-1$. Se puede ver este coeficiente como una extensión de la medida de Jaccard, aunque su sentido probabilístico se pierde.

5. **Medida de Rogers-Tanimoto**

$$\frac{a+d}{a+d+2(b+c)} \tag{2.12}$$

Este coeficiente puede interpretarse como una extensión de la medida de parejas simples, pesando con el doble valor las no coincidencias.

6. **Medida de Kulcynski**

$$\frac{a}{b+c} \tag{2.13}$$

Esta medida muestra el cociente entre coincidencias y no coincidencias, excluyendo los pares $0-0$.

No son éstas las únicas medidas de este tipo que existen. Podíamos seguir citando muchas más y, entre ellas, a modo de ejemplo:

$$\begin{array}{cccc} \dfrac{a+d}{b+c} & \dfrac{a+d}{a+b+c} & \dfrac{2a}{2(a+d)+b+c} & \dfrac{2(a+d)}{2(a+d)+b+c} \\ \dfrac{2(a+d)}{2a+b+c} & \dfrac{a}{a+d+2(b+c)} & \dfrac{a}{a+2(b+c)} & \dfrac{a+d}{a+2(b+c)} \end{array} \tag{2.14}$$

2.3.4. Medidas basadas en probabilidades condicionadas.

Notemos que, de entre las medidas citadas con anterioridad, (2.8), (2.10) y (2.11) poseen interpretaciones probabilísticas razonables. Hay otras medidas que también poseen fundamentos probabilísticos.

Así, como ya se ha comentado con anterioridad, $\dfrac{a}{a+b}$ es la probabilidad condicionada de que un individuo elegido al azar presente el valor 1 en la variable X_j dado que ha presentado un 1 en la variable X_i. Asimismo, $\dfrac{a}{a+c}$ es la probabilidad condicionada de que un individuo elegido al azar presente un 1 en la variable X_i dado que lo ha presentado en la variable X_j.

Así podríamos pensar en una medida que marcara la probabilidad de que un individuo presente un 1 en una variable, dado que ha presentado un 1 en la otra, surgiendo la medida

$$\frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right] \tag{2.15}$$

Como sabemos, no es claro que la codificación hecha sea la mejor. Por ello se puede optar por tener en cuenta también las otras coincidencias, dando lugar a la medida

$$\frac{1}{4}\left[\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right] \tag{2.16}$$

Estas expresiones son similares a las obtenidas a partir del coseno del ángulo entre variables en el caso de datos binarios, salvo que en lugar de tomar medias geométricas se toman medias aritméticas.

Por último se puede citar la medida de Hamann

$$\frac{2(a+d)}{a+b+c+d}-1=\frac{(a+d)-(b+c)}{a+b+c+d} \tag{2.17}$$

que indica la probabilidad de que un caso elegido al azar presente una coincidencia menos la probabilidad de que presente una diferencia en alguna de las variables.

2.4. Medidas de asociación entre individuos.

2.4.1. Distancias euclídea, de Minkowski y de Mahalanobis.

Consideremos ahora dos individuos tomados de la población, lo cual corresponde a tomar dos filas en la matriz de datos X:

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

La métrica más conocida, que corresponde a la generalización a más de dos dimensiones de la distancia entre dos puntos en el plano, es la derivada de la norma $\mathbf{L}_2$ de un vector: [1]

$$||x_i||_2 = \sqrt{x_i' x_i} = \sqrt{\sum_{l=1}^{n} x_{il}^2}$$

obteniéndose, a partir de ella, la distancia euclídea

$$d_2(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)' (x_i - x_j)} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \tag{2.18}$$

Esta métrica tiene la propiedad, al igual que la norma $\mathbf{L}_2$, de que todos sus valores son invariantes respecto de las transformaciones ortogonales $\widetilde{x}_i = \theta x_i$, donde θ es una matriz $n \times n$ que verifica $\theta'\theta = \theta\theta' = I$. En efecto:

$$||\theta x_i||_2 = \sqrt{x_i' \theta' \theta x_i} = \sqrt{x_i' x_i} = ||x_i||_2$$

y así se tiene

$$d_2(\theta x_i, \theta x_j) = d_2(x_i, x_j)$$

Además se verifica que estas transformaciones, además de las traslaciones, son las únicas para las cuales d_2 es invariante [2] .

En cuanto a las distancias de Minkowski, éstas proceden de las normas $\mathbf{L}_p$

$$||x_i||_p = \left(\sum_{l=1}^{n} |x_{il}|^p \right)^{\frac{1}{p}} \quad p \geq 1$$

dando origen a

$$d_p(x_i, x_j) = ||x_i - x_j||_p = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}} \tag{2.19}$$

Es fácil comprobar que esta distancia es invariante ante traslaciones, siendo éstas las únicas funciones para las cuales d_p posee esta propiedad.

Además se verifica la conocida relación

$$d_p(x_i, x_j) \leq d_q(x_i, x_j) \Leftrightarrow p \geq q$$

[1] Recordemos que dado un espacio vectorial X sobre un cuerpo K, una norma es una aplicación $||\cdot|| : X \longrightarrow K_0^+$ que verifica

1. $||x|| = 0 \Leftrightarrow x = 0$
2. $||\alpha x|| = |\alpha| ||x|| \quad \forall \alpha \in K \quad \forall x \in X$
3. $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in X$

[2] En efecto, si consideramos $\widehat{x}_i = a + x_i$ y $\widehat{x}_j = a + x_j$, entonces se tiene:

$$d_2(\widehat{x}_i, \widehat{x}_j) = ||\widehat{x}_i - \widehat{x}_j||_2 = ||(a + x_i) - (a + x_j)||_2 = ||x_i - x_j||_2 = d_2(x_i, x_j)$$

Algunos casos particulares para valores de p concretos son [3]

1. Distancia d_1 o distancia ciudad (City Block) $(p = 1)$

$$d_1(x_i, x_j) = \sum_{l=1}^{n} |x_{il} - x_{jl}| \tag{2.20}$$

2. Distancia de Chebychev o distancia del máximo $(p = \infty)$

$$d_\infty(x_i, x_j) = \underset{l=1,\ldots,n}{\text{Max}} |x_{il} - x_{jl}| \tag{2.21}$$

Por otra parte, se puede generalizar la distancia euclídea, a partir de la norma

$$||x_i||_B = \sqrt{x_i' B x_i}$$

donde B es una matriz definida positiva. La métrica correspondiente a dicha norma es:

$$D_B(x_i, x_j) = \sqrt{(x_i - x_j)' B (x_i - x_j)} = \sqrt{\sum_{l=1}^{n}\sum_{h=1}^{n} b_{lh} x_{il} x_{jh}} \tag{2.22}$$

En el caso particular en que B sea una matriz diagonal, sus elementos son pesos positivos para las componentes del vector que corresponde a las variables en la matriz de datos.

Esta distancia se mantiene invariante frente a transformaciones (semejanzas) efectuadas por una matriz P que verifique $P'BP = B$. En efecto:

$$D_B(Px_i, Px_j) = \sqrt{(Px_i - Px_j)' B (Px_i - Px_j)} =$$

$$= \sqrt{(x_i - x_j)' P'BP (x_i - x_j)} = \sqrt{(x_i - x_j)' B (x_i - x_j)} = D_B(x_i, x_j)$$

La llamada métrica de Mahalanobis se obtiene tomando en 2.22 una matriz B determinada. Dicha matriz es la llamada matriz de varianzas-covarianzas de las variables (columnas de la matriz X de datos).

Los elementos de la matriz S, matriz de varianzas-covarianzas, se definen de la siguiente forma:

$$s_{uv} = \frac{1}{m}\sum_{l=1}^{m}(x_{lu} - \overline{x}_u)(x_{lv} - \overline{x}_v) \quad ; \quad u, v = 1, \ldots, n \tag{2.23}$$

Matricialmente tenemos dicha matriz expresada en la forma:

$$S = \frac{1}{m}\widetilde{X}'\widetilde{X} \quad \text{con} \quad \widetilde{X} = (\widetilde{x}_{ij}) \;\; ; \;\; \widetilde{x}_{ij} = x_{ij} - \overline{x}_j \;\; i = 1, \ldots, m \;\; ; \;\; j = 1, \ldots, n \tag{2.24}$$

A partir de la matriz S se puede definir la matriz de correlaciones, R, cuyos elementos son

$$\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \quad ; \quad i, j = 1, \ldots, n$$

Notemos que si $m \geq n$, entonces la matriz de varianzas-covarianzas S es definida positiva y tiene sentido definir la distancia de Mahalanobis, para individuos, como:

$$D_S(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} \tag{2.25}$$

Esta distancia es invariante frente a transformaciones regidas por una matriz $C_{n\times n}$ no singular. En efecto, [4]

[3] Notemos que esta distancia generaliza a la distancia euclídea, en tanto en cuanto, esta última es un caso particular para $p = 2$.

[4] Notemos que la matriz de varianzas-covarianzas de las variables transformadas queda de la forma:

$$S = \frac{1}{m} C\widetilde{X}'\widetilde{X}C'$$

$$D_S(Cx_i, Cx_j) = \sqrt{(Cx_i - Cx_j)' \left[\frac{1}{m} C\widetilde{X}'\widetilde{X}C'\right]^{-1} (Cx_i - Cx_j)} =$$

$$= \sqrt{(x_i - x_j)' C'(C')^{-1} \left[\frac{1}{m}\widetilde{X}'\widetilde{X}\right]^{-1} C^{-1}C (x_i - x_j)} =$$

$$= \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)} = D_S(x_i, x_j)$$

Si, en particular, C es una matriz diagonal con los elementos no nulos, la transformación de X por C significa que el valor de cada variable en X es multiplicado por una constante, o sea, se ha hecho un cambio de escala. Por ello la métrica de Mahalanobis es invariante frente a cambios de escala, propiedad que no posee, por ejemplo, la métrica euclídea.

En la aplicación de las técnicas cluster la métrica de Mahalanobis presenta la desventaja de que el cálculo de la matriz S está basado en todos los individuos de forma conjunta y no trata, como sería de desear, de manera separada los objetos de cada cluster; además, su cálculo es mucho más laborioso que el de otras métricas. Por estas razones no suele emplearse en las técnicas cluster, si bien puede utilizarse dentro de cada cluster formado en una etapa determinada.

2.4.2. Correlación entre individuos.

Formalmente hablando, el coeficiente de correlación entre vectores de individuos puede ser usado como una medida de asociación entre individuos.

$$\text{Individuo i} \qquad x_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$$

$$\text{Individuo j} \qquad x_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$$

$$r_{ij} = \frac{\sum_{l=1}^{n} (x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j} \tag{2.26}$$

donde se ha definido

$$\overline{x}_h = \frac{1}{n}\sum_{l=1}^{n} x_{hl} \qquad h = i, j \quad \text{Media de cada individuo}$$

$$s_h^2 = \sum_{l=1}^{n} (x_{hl} - \overline{x}_h)^2 \quad h = i, j \quad \text{Desviación cuadrática de cada individuo}$$

El principal problema de este coeficiente radica en el hecho de que en un vector de datos correspondiente a un individuo hay muchas unidades de medida diferentes, lo cual hace muy difícil comparar las *medias* y las *varianzas*.

No obstante, Cronbach y Gleser, en 1953, demostraron que este coeficiente posee un carácter métrico.

En efecto, sea x_{ik} el valor de la k-ésima variable sobre el i-ésimo individuo y transformemos ese dato en

$$\widehat{x}_{ik} = \frac{x_{ik} - \overline{x}_i}{s_i}$$

Entonces, la distancia euclídea al cuadrado entre dos individuos sobre los que se ha efectuado ese tipo de transformación será:

$$d_2^2(\widehat{x}_i, \widehat{x}_j) = \sum_{l=1}^{n} \left[\frac{x_{il} - \overline{x}_i}{s_i} - \frac{x_{jl} - \overline{x}_j}{s_j}\right]^2 =$$

$$= \sum_{l=1}^{n} \left[\frac{(x_{il} - \overline{x}_i)^2}{s_i^2} + \frac{(x_{jl} - \overline{x}_j)^2}{s_j^2} - 2\frac{(x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{s_i s_j}\right] = 2(1 - r_{ij})$$

Observemos que las dos medidas de la variable k-ésima, x_{ik} y x_{jk} son sometidas a transformaciones distintas

$$\widehat{x}_{ik} = \frac{x_{ik} - \overline{x}_i}{s_i}$$

$$\widehat{x}_{jk} = \frac{x_{jk} - \overline{x}_j}{s_j}$$

por lo que los nuevos valores no son comparables. Además, se observa que $1 - r$, complemento a uno del coeficiente de correlación, es una métrica (si $r_{ij} \longrightarrow 1 \Longrightarrow d(\widehat{x}_i, \widehat{x}_j) \longrightarrow 0$), pero lo es en el espacio en el que los datos se han transformado al tipificarlos.

Otra observación a hacer es que si se cambia la unidad de medida de una variable, cambia una componente en cada uno de los vectores de individuos: así si cambiamos la unidad de medida en la variable k-ésima, cambian los datos x_{ik} y x_{jk}; en consecuencia, cambian $\overline{x}_i$, $\overline{x}_j$, s_i y s_j y así cambia el coeficiente de correlación. Así pues, r_{ij}, es dependiente de cambios en unidades de medida. Es decir, estos cambios sopesan de manera distinta a las variables.

Por último, los valores de cada individuo pueden ser transformados de la siguiente manera

$$\widehat{x}_{ik} = \frac{x_{ik}}{\left(\sum_{l=1}^{n} x_{il}^2\right)^{\frac{1}{2}}}$$

Al igual que antes se puede demostrar, lo cual se deja como ejercicio al lector, que

$$d_2^2\left(\widehat{x}_i, \widehat{x}_j\right) = 2(1 - \cos(\alpha_{ij}))$$

donde

$$\cos(\alpha_{ij}) = \frac{\sum_{l=1}^{n} x_{il} x_{jl}}{\left(\sum_{l=1}^{n} x_{il}^2\right)^{\frac{1}{2}} \left(\sum_{l=1}^{n} x_{jl}^2\right)^{\frac{1}{2}}}$$

y, por lo tanto, $1 - \cos(\alpha_{ij})$ es una métrica.

2.4.3. Distancias derivadas de la distancia χ^2.

Hay muchas medidas de asociación que se basan en el estadístico χ^2, de uso familiar en el análisis de tablas de contingencia. Notemos

o_{ij} = valor observado en la celda i, j

e_{ij} = valor esperado bajo la hipótesis de independencia

Con dicha notación se define el estadístico χ^2 como

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{2.27}$$

donde p y q son el número de modalidades de las variables estudiadas.

Var **A**\Var **B**	1	$\cdots$	j	$\cdots$	q	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\cdot}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\cdot}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot q}$	$n_{\cdot\cdot}$

(2.28)

Bajo la hipótesis de independencia de ambas variables, el valor esperado en la celda i, j es

$$e_{ij} = f_{i\cdot} f_{\cdot j} n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$$

pero, por otra parte:

$$o_{ij} = n_{ij} = f_{ij} n_{..}$$

con lo cual

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \dfrac{n_{i\cdot} n_{\cdot j}}{n_{..}}\right)^2}{\dfrac{n_{i\cdot} n_{\cdot j}}{n_{..}}} =$$

$$= n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} n_{..} - f_{i\cdot} f_{\cdot j} n_{..})^2}{f_{i\cdot} f_{\cdot j}} = n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2 n_{..}^2}{f_{i\cdot} f_{\cdot j} n_{..}^2} =$$

$$= n_{..} \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = n_{..} \left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{f_{ij}^2}{f_{i\cdot} f_{\cdot j}} - 1\right] =$$

$$= n_{..} \left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right]$$

Ahora bien, esta cantidad, que es muy útil para contrastes en tablas de contingencia, no lo es tanto como medida de asociación, puesto que aumenta cuando $n_{..}$ crece. Por ello se considera la medida Φ^2, llamada contingencia cuadrática media, definida como

$$\Phi^2 = \frac{\chi^2}{n_{..}} \tag{2.29}$$

Sin embargo, este coeficiente depende del tamaño de la tabla. Por ejemplo, supongamos que $p = q$ y que las variables están asociadas de forma perfecta, o sea, $n_{i\cdot} = n_{\cdot i} = n_{ii}\ \forall i$ (notemos que en tal caso sólo hay p casillas con valores distintos de cero). En este caso

$$\chi^2 = n_{..}(p-1)$$

$$\Phi^2 = p - 1$$

En el caso de una tabla rectangular con las variables perfectamente relacionadas, el número de casillas no nulas es $\text{Min}(p, q)$, por lo que

$$\chi^2 = n_{..} \,\text{Min}(p-1, q-1)$$

$$\Phi^2 = \text{Min}(p-1, q-1)$$

Con estas ideas en mente, se han hecho algunos intentos para normalizar la medida Φ^2 al rango $[0, 1]$. Por ejemplo:

$$\begin{aligned}
&\text{Medida de Tschuprow:} && T = \left(\frac{\Phi^2}{[(p-1)(q-1)]^{\frac{1}{2}}}\right)^{\frac{1}{2}} \\
&\text{Medida de Cramer:} && C = \left(\frac{\Phi^2}{\text{Min}(p-1, q-1)}\right)^{\frac{1}{2}} \\
&\text{Coeficiente de contingencia de Pearson:} && P = \left(\frac{\Phi^2}{1+\Phi^2}\right)^{\frac{1}{2}} = \left(\frac{\chi^2}{n_{..} + \chi^2}\right)^{\frac{1}{2}}
\end{aligned} \tag{2.30}$$

Obviamente, este tipo de medidas son empleadas en los casos en los que los datos que se poseen son conteos de frecuencias. Así, supongamos que tenemos m individuos sobre los que se han observado n variables. Sea x_{ij} la frecuencia observada de la j-ésima variable sobre el i-ésimo individuo.

	Var 1	$\cdots$	Var j	$\cdots$	Var n	
Ind. 1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	$x_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ind. i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{in}	$x_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ind. m	x_{m1}	$\cdots$	x_{mj}	$\cdots$	x_{mn}	$x_{m\cdot}$
	$x_{\cdot 1}$	$\cdots$	$x_{\cdot j}$	$\cdots$	$x_{\cdot n}$	$x_{\cdot\cdot}$

Consideremos dos individuos x_i y x_j y sea la tabla $2 \times n$ formada a partir de ellos

	Var 1	$\cdots$	Var n	
Ind. i	x_{i1}	$\cdots$	x_{in}	$\sum_{l=1}^{n} x_{il}$
Ind. j	x_{j1}	$\cdots$	x_{jn}	$\sum_{l=1}^{n} x_{jl}$
	$x_{i1} + x_{j1}$	$\cdots$	$x_{in} + x_{jn}$	$\sum_{l=1}^{n} (x_{il} + x_{jl})$

Obviamente, cada individuo presenta un total de frecuencia marginal distinto ($x_{\cdot i}$ y $x_{\cdot j}$), por lo que no son comparables uno a uno. En este caso hay que buscar la semejanza teniendo en cuenta la proporcionalidad entre ambos. Por ello el empleo de distancias basadas en la distancia χ^2 es útil.

En nuestro caso, la forma que adopta el estadístico es:

$$\chi^2 = \sum_{l=1}^{n} \left[\frac{(x_{il} - e_{il})^2}{e_{il}} + \frac{(x_{jl} - e_{jl})^2}{e_{jl}} \right] \tag{2.31}$$

donde

$$e_{kh} = \frac{\sum_{l=1}^{n} x_{kl} (x_{ih} + x_{jh})}{\sum_{l=1}^{n} (x_{il} + x_{jl})} \quad ; \quad k = i, j \quad ; \quad h = 1, \cdots, n$$

y así, si $\chi^2 \longrightarrow 0$ se tiene la proporcionalidad buscada entre las dos filas y, por lo tanto, los dos individuos presentan el mismo perfil a lo largo de las variables, con lo cual dichos individuos serán parecidos.

2.4.4. Medidas no métricas: Coeficiente de Bray-Curtis.

Dados dos individuos

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

el coeficiente de Bray-Curtis viene definido por la expresión

$$D_{i,j} = \frac{\sum_{l=1}^{n} |x_{il} - x_{jl}|}{\sum_{l=1}^{n} (x_{il} + x_{jl})} \tag{2.32}$$

El numerador no es otra cosa que la métrica $\mathbf{L}_1$, mientras que el denominador puede ser interpretado como una medida de la magnitud total de los dos individuos.

Hay que hacer notar que es aconsejable usar esta medida con datos no negativos, ya que pudiera haber cancelaciones en el denominador, pudiéndose obtener resultados poco aconsejables; por ejemplo, usando esta medida, no es aconsejable centrar los datos previamente. Además, puesto que para cada par de individuos se emplea un denominador distinto, esta medida no satisface siempre la desigualdad triangular.

2.4.5. Medidas para datos binarios.

Con alguna excepción, las medidas de asociación que se mencionaron para variables de tipo binario pueden ser aplicadas para medir la asociación entre individuos. En este caso la tabla de contingencia que se tiene es

Ind. I\Ind. J	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$n=a+b+c+d$

(2.33)

Evidentemente, ahora a representa el número de veces que los individuos i y j presentan, de forma simultánea, un 1 sobre una misma variable.