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

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

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:

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

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.