Introducción:

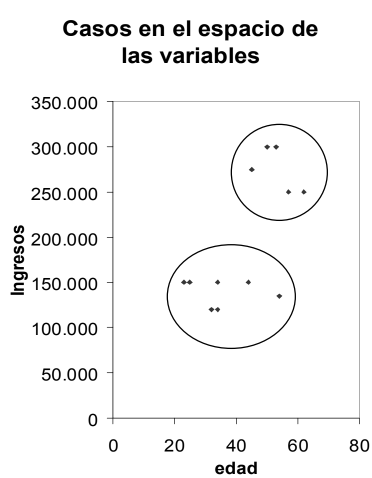
Ejemplo

Caso	Edad	Sexo	Ingresos	Part. Ingre.	Niv. Estud	Miem-bros	IRPF
1	34	1	120.000	100	1	3	22,1
2	45	1	275.000	85	2	3	24,5
3	34	2	150.000	50	1	4	18,0
4	25	1	150.000	35	3	2	23,1
5	62	2	250.000	99	1	2	32,3
6	53	1	300.000	75	1	3	34,1
7	32	2	120.000	100	2	3	22,1
8	54	2	135.000	85	2	3	24,5
9	23	2	150.000	50	3	4	18,0
10	44	1	150.000	35	1	2	23,1
11	57	1	250.000	100	2	2	32,3
12	50	2	300.000	75	1	3	34,1

Introducción:

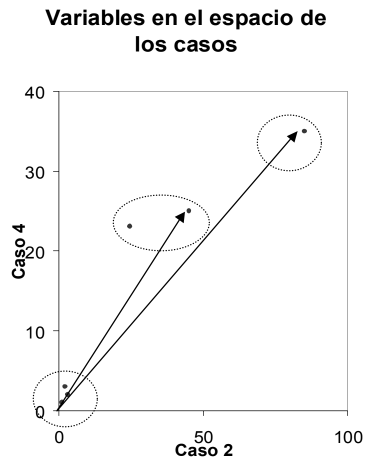
Ejemplo

- Casos representados en el espacio de las Variables
- **Ejemplo:**
  - **Casos en el espacio (Edad, Salarios)**
- **Proximidades:**
  - **Distancias,**
  - **Disimilaridades,...**



Introducción:
Ejemplo

- Variables representadas en el espacio de los Casos
- **Ejemplo:**
  - **Variables en el espacio (Caso 2, Caso 4)**
- **Proximidades:**
  - **Asociación,**
  - **Correlación, ...**



Dpto. de Estadística Económica, Estructura Económica y O.E.I.
© Fco. Javier Callealta Barroso

Proximidades:

- Miden el grado de semejanza que presentan dos elementos cualesquiera (i,j) obtenido a partir de un cierto número de sus características, y que notaremos genéricamente por d(i,j).
- Para medirlo, generalmente se utilizan medidas de distancia y de disimilaridad, que aumentan de valor al decrecer la semejanza.
- Alternativamente, se utilizan medidas de similaridad que aumentan de valor al crecer la semejanza.

Proximidades:
Definiciones

- Medida de Distancia o Métrica
  - Función d:ExE ———>R^+, tal que

    $d(i,j) \geq 0$, $\forall$ i,j

    $d(i,j) = 0 \Leftrightarrow i=j$, $\forall$ i,j

    $d(i,j) = d(j,i)$, $\forall$ i,j (Simetría)

    $d(i,j) \leq d(i,k) + d(k,j)$, $\forall$ i,j,k (Triangular)

  - Como consecuencia verifica:

    $d(i,i) = 0$, $\forall$ i

Proximidades:
Definiciones

- Medida de Disimilaridad (General):
  - Función d:ExE ———>R, tal que

    d(i,j) disminuye cuando aumenta la similitud entre i y j

- Medida de Disimilaridad (Jardine & Sibson):
  - Función d:ExE ———>R^+, tal que

    $d(i,j) \geq 0$, $\forall$ i,j

    $d(i,i) = 0$, $\forall$ i

    $d(i,j) = d(j,i)$, $\forall$ i,j (Simetría)

Proximidades:
Definiciones

- Medida de Similaridad (General)
  - Función d:ExE ———>R, tal que

    d(i,j) aumenta cuando aumenta la similitud entre i y j.

- Medida de Similaridad (Jardine-Sibson)
  - Función d:ExE ———>R^+, tal que

    d(i,j) ≥ 0 , ∀ i,j

    d(i,j) = d(j,i) , ∀ i,j (Simetría)

    d(i,j) aumenta cuando aumenta la similitud entre i y j.
- Generalmente suelen construirse acotadas por 1.
- Si d(i,j) es una medida de disimilaridad acotada por M ⇒

  δ(i,j)=M-d(i,j) es una medida de similaridad.
- Si δ(i,j) es una medida de similaridad acotada por M ⇒

  d(i,j)=M- δ(i,j) es una medida de disimilaridad.

Proximidades:
Caso particular de interés teórico (1)

- Disimilaridad Euclidizable

  d(i,j) ≥ 0 , ∀ i,j

  d(i,j) = d(j,i) , ∀ i,j (Simetría)

  $$\forall i, \exists I = (x_{i1}, x_{i2}, ..., x_{im}) \in \Re^m \mid d(i,j) = d_2(I,J) = \sqrt{\sum_{h=1}^{m} (x_{ih} - x_{jh})^2}$$

  (Euclidizable)
  - Como consecuencia, verifica:

    d(i,j) = 0 ⇔ i=j , ∀ i,j (⇒ d(i,i) = 0 , ∀ i)

    d(i,j) ≤ d(i,k) + d(k,j) , ∀ i,j,k (Triangular)

Proximidades:
Caso particular de interés teórico (2)

- Disimilaridad Ultramétrica
  - Función d:ExE ———>R^+, tal que

    $d(i,j) \geq 0$, $\forall$ i,j
    $d(i,i) = 0$, $\forall$ i
    $d(i,j) = d(j,i)$, $\forall$ i,j (Simetría)
    $d(i,j) \leq \text{máx}_k\,(d(i,k) , d(k,j))$, $\forall$ i,j,k (Ultramétrica)

  - como consecuencia verifica,

    . $d(i,j) \leq d(i,k) + d(k,j)$, $\forall$ i,j,k (Triangular)
    . Cualesquiera tres puntos i,j,k forman un triangulo isósceles de base formada por los dos puntos que menos difieren

Medidas de Disimilaridad para casos:
Escalas de Intervalo

Distancia Euclídea:

$$d_2(i,j) = \sqrt{\sum_{h=1}^{p} (x_{ih} - x_{jh})^2}$$

Distancia de Minkowski:

$$d_q(i,j) = \left(\sum_{h=1}^{p} \left| x_{ih} - x_{jh} \right|^q \right)^{1/q}$$

- Casos Particulares de la Distancia de Minkowski:
  - Si q=1, se obtiene la Distancia de Manhatan o de “City-Block”
  - Si q=2, se obtienen la Distancia Euclídea
  - Si q $\to \infty$, se obtienen la Distancia de Chebychev

Medidas de Disimilaridad para casos:
Escalas de Intervalo

- **D^2 de Mahalanobis:**
- **entre 2 individuos**

$$d(i,j) = D^2(i,j) = (x_i - x_j)'\Sigma^{-1}(x_i - x_j)$$

- **entre un individuo y su grupo (centroide)**

$$d(i,\bar{x}) = D^2(i,\bar{x}) = (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x})$$

- **Entre 2 grupos**

$$d(\bar{x}_i,\bar{x}_j) = D^2(\bar{x}_i,\bar{x}_j) = (\bar{x}_i - \bar{x}_j)'\Sigma^{-1}(\bar{x}_i - \bar{x}_j)$$

Medidas de Proximidad para casos:
Escalas Nominales - Casos en escalas binarias

Si comparamos casos en base a p variables binarias observadas:

$$\begin{array}{lccccccc} caso\ i: & 0 & 1 & 1 & 0 & 0 & ... & 1 \\ caso\ j: & 0 & 1 & 0 & 1 & 0 & ... & 0 \end{array}$$

Podemos resumir la información de su comparación en una tabla de contingencia 2x2, donde:

$$\begin{array}{cc} i/j & \begin{array}{cc} 1 & 0 \end{array} \\ \begin{array}{c} 1 \\ 0 \end{array} & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array}$$

a = número de variables donde caso i es 1 y caso j es 1
b = número de variables donde caso i es 1 y caso j es 0
c = número de variables donde caso i es 0 y caso j es 1
d = número de variables donde caso i es 0 y caso j es 0

- la asociación positiva representa similaridad entre los casos
- la asociación negativa representa disimilaridad entre los casos

Medidas de Similaridad para casos:
Escalas Nominales - Casos en escalas binarias

- Basadas en medidas de Asociación para Tablas 2x2.
  - **χ^2** $$\chi^2_{\exp} = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$
  - **χ^2 corregida de Yates:** $$\chi^2_{\exp} = \frac{N(|ad-bc|-0.5N)^2}{(a+b)(a+c)(c+d)(b+d)}$$
  - **Razón de Proporciones:** $$\psi = \frac{a \cdot d}{b \cdot c}$$
  - **Q de Yule** $$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$
  - **Demás medidas de asociación para Tablas 2x2.**

Dpto. de Estadística Económica, Estructura Económica y O.E.I.
© Fco. Javier Callealta Barroso

Medidas de Similaridad para casos:
Escalas Nominales - Casos en escalas binarias

- **Distancia Euclídea (disimilaridad)** $$d_2 = \sqrt{b+c}$$
- **Coef. de correlación** $$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a+d)(c+d)(a+c)(b+d)}}$$
- **Coef. de Similitud de parejas simples:** $$C_1 = \frac{a+d}{a+b+c+d}$$
- **Coef. Rassel y Rao** $$RR = \frac{a}{a+b+c+d}$$
- **Otras Muchas ...... (Bizquerra, pp. 49-54)**

Medidas de Similaridad para casos:

Escalas nominales y de intervalo simultáneamente

- **Coeficiente de Similaridad de Gower[1]:**

$$d_{ij}=\frac{\sum_{k=1}^{p} w_k \delta_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_k \delta_{ikj}}, \quad \text{siendo}: S_{ijk}=\begin{cases} \text{cuando } X_k \text{ es variable}: 1-\left|\dfrac{x_{ik}-x_{jk}}{\max_l\{x_{lk}\}-\min_l\{x_{lk}\}}\right| \\ \text{cuando } X_k \text{ es atributo}: \begin{cases} 1 \text{ si } x_{ik}=x_{jk} \\ 0 \text{ si } x_{ik}\neq x_{jk} \end{cases} \end{cases}$$

w_k = factor de ponderación de cada variable k-ésima

$$\delta_{ikj}=\begin{cases} 1, \text{ si la característica k puede compararse para los casos i y j} \\ 0, \text{ si la característica k no puede compararse para los casos i y j} \end{cases}$$

[1]: Gower, J.C. (1971) "A General Coefficient of Similarity and some of its Properties" *Biometrics, 27 pp.857-874.*

Medidas de Similaridad para variables:

Escalas de Intervalo

Medidas basadas en la relación: Y=a+bX

- **Coeficiente de Correlación lineal**

$$d(X_i,X_j)=\frac{S_{xy}}{S_x\cdot S_y}=\frac{\frac{1}{n}\sum_{h=1}^{n}(x_{hi}-\bar{x}_i)\cdot(x_{hj}-\bar{x}_j)}{\sqrt{\frac{1}{n}\sum_{h=1}^{n}(x_{hi}-\bar{x}_i)^2}\cdot\sqrt{\frac{1}{n}\sum_{h=1}^{n}(x_{hj}-\bar{x}_j)^2}}=\frac{1}{n}\sum_{h=1}^{n} z_{hi}\cdot z_{hj}$$

- **Coeficiente de Determinación**

$$d(X_i,X_j)=\frac{S_{xy}^2}{S_x^2\cdot S_y^2}=R_{xy}^2$$

Medidas de Similaridad para variables:

Escalas de Intervalo

Medidas basadas en la relación: Y=bX

- **Coseno entre variables**

$$d(X_i, X_j) = \frac{\sum_{h=1}^{n} x_{hi} \cdot x_{hj}}{\sqrt{\sum_{h=1}^{n} x_{hi}^2} \cdot \sqrt{\sum_{h=1}^{n} x_{hj}^2}} = \cos(\vec{X}_i, \vec{X}_j)$$

- **Cuadrado del Coseno entre variables**

$$d(X_i, X_j) = \frac{\left(\sum_{h=1}^{n} x_{hi} \cdot x_{hj}\right)^2}{\sum_{h=1}^{n} x_{hi}^2 \cdot \sum_{h=1}^{n} x_{hj}^2} = \cos^2(\vec{X}_i, \vec{X}_j)$$

Medidas de Similaridad para variables:

Escalas de Intervalo

Medida basadas en la relación: Y=X

- **Distancia Euclídea:**

$$d_2(X_i, X_j) = \sqrt{\sum_{h=1}^{n} (x_{hi} - x_{hj})^2}$$

- **Distancia de Minkowski:**

$$d_q(X_i, X_j) = \left(\sum_{h=1}^{n} \left| x_{hi} - x_{hj} \right|^q\right)^{1/q}$$

Medidas de Similaridad para variables:
Escalas Nominales-Tablas hxk

χ^2	$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{\left(e_{ij}-n_{ij}\right)^2}{e_{ij}}$
Coeficiente de Contingencia Cuadrático Medio	$\phi^2 = \frac{\chi^2}{N}$
Coeficiente de Contiigencia de Pearson	$P = \sqrt{\frac{\frac{\chi^2}{N}}{1+\frac{\chi^2}{N}}} = \sqrt{\frac{\phi^2}{1+\phi^2}}$
Coeficiente T de Tschuprov	$T = \left\{\frac{\frac{\chi^2}{N}}{\sqrt{(h-1)(k-1)}}\right\}^{1/2}$
Coeficiente V de Cramer	$V = \left\{\frac{\frac{\chi^2}{N}}{min(\mathrm{h-1,k-1})}\right\}^{1/2}$

Medidas de Similaridad para variables:
Escalas Nominales – Variables dicotómicas

Basadas en medidas de Asociación para Tablas 2x2.

- **χ^2** $\quad \chi^2_{\exp} = \frac{N\left(ad-bc\right)^2}{(a+b)(a+c)(c+d)(b+d)}$
- **χ^2 corregida de Yates:** $\quad \chi^2_{\exp} = \frac{N\left(\left|ad-bc\right|-0.5N\right)^2}{(a+b)(a+c)(c+d)(b+d)}$
- **Razón de Proporciones:** $\quad \psi = \frac{a \cdot d}{b \cdot c}$
- **Q de Yule** $\quad Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$
- **Demás medidas de asociación para Tablas 2x2.**

Preparación de los datos

- Homogeneización de las escalas
  - Al tipo de escala más informativo, ganando información subjetiva
  - Al tipo de escala menos informativa, perdiendo información
- Estandarización de las variables
  - Tipificación
  - Transformación condicionando media, varianza, máximo o rango
- Transformación de las Proximidades
  - Transformación condicionando media, varianza, máximo o rango