



PASOS PARA SU APLICACIÓN

ANÁLISIS DE CONGLOMERADOS

MÉTODOS JERÁRQUICOS



I. Selección de variables

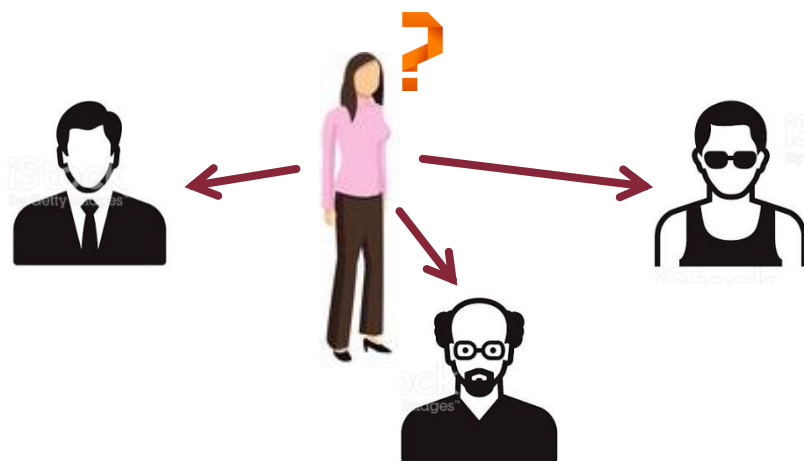
- Selección de un conjunto concreto de características usadas para describir a cada individuo que sirva de marco de referencia para establecer las agrupaciones. *Refleja la opinión del investigador acerca del propósito de la clasificación.*
- Una gran cantidad de variables puede ocasionar problemas y dificultar la identificación de la estructura de los grupos. Es posible utilizar previamente un Análisis de Componentes Principales o Análisis de Factores para reducir la dimensionalidad
- El *tipo de variable* y las *unidades* de medición influyen en la forma de tratar esos datos para generar las agrupaciones. Es recomendable trabajar con datos transformados, para eliminar el efecto de la escala, además facilitar la interpretación.



2. Escoger la medida de asociación

Medir la proximidad o similitud de los objetos en estudio puede expresarse en forma de **DISTANCIA**. Aquel par de observaciones que tengan más características en común, tendrán una distancia más corta. Una distancia más grande indica poca similitud entre el par de objetos considerando las variables seleccionadas previamente.

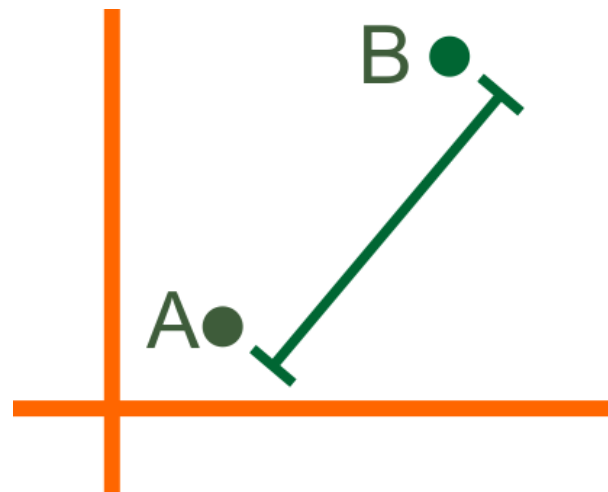
Antes de definir medidas de distancia, recuerde que muchas de las técnicas analíticas son particularmente sensitivas a los *outliers*. Por lo tanto, existen algunos chequeos preliminares para *outliers* y errores de dedo como el gráfico de dispersión, el diagrama de cajas, ...



2. Escoger la medida de asociación

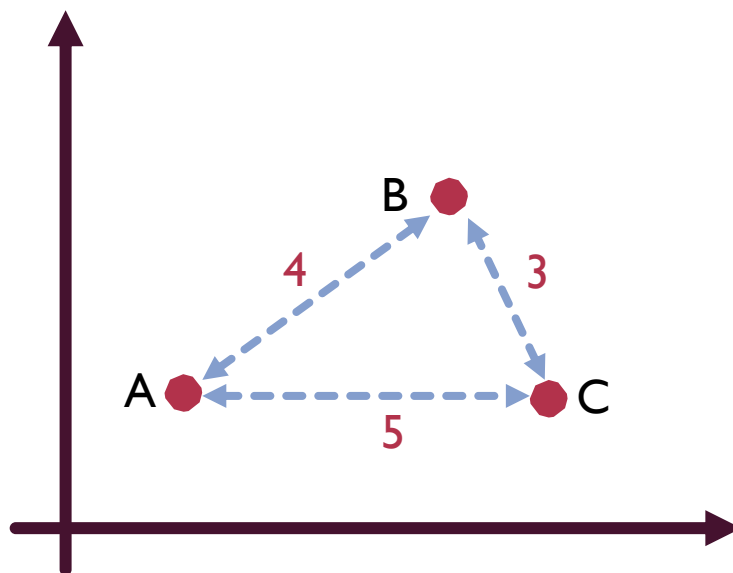
Matemáticamente se da el nombre de distancia entre dos puntos (A, B) , a toda medida que verifique los axiomas siguientes:

1. La distancia del punto A hacia B es positiva, $d(A, B) \geq 0$
2. La distancia hacia un mismo punto es cero, $d(A, A) = 0$
3. La distancia del punto A hacia B es la misma que si se parte del punto B hacia el punto A, $d(A, B) = d(B, A)$
4. La distancia del punto A hacia el punto B es menor o igual que ir del punto A al punto B pasando por un punto C,
$$d(A, B) \leq d(A, C) + d(C, B)$$



2. Escoger la medida de asociación

Ejemplo:



Distancia (A, B)

1. $d(A, B) = 4 \therefore d(A, B) \geq 0$ ✓
 2. $d(A, A) = 0$ y $d(B, B) = 0$ ✓
 3. $d(A, B) = 4$ y $d(B, A) = 4$ ✓
 4. $d(A, B) = 4, d(A, C) = 5, d(C, B) = 3$
- $$\underbrace{d(A, B)}_4 \leq \underbrace{d(A, C) + d(C, B)}_8 \quad \checkmark$$

Distancia (B, C)

- $$d(B, C) = 3 \therefore d(B, C) \geq 0 \quad \checkmark$$
- $$d(B, B) = 0 \text{ y } d(C, C) = 0 \quad \checkmark$$
- $$d(B, C) = 3 \text{ y } d(C, B) = 3 \quad \checkmark$$
- $$d(B, C) = 3, d(B, A) = 4, d(A, C) = 5$$
- $$\underbrace{d(B, C)}_3 \leq \underbrace{d(B, A) + d(A, C)}_9 \quad \checkmark$$

2. Escoger la medida de asociación

Existen diferentes medidas de asociación (formas de medir distancia) para variables cuantitativas:

- **Distancia Euclidiana.** Este tipo de distancia es probablemente el más usado. Se calcula así:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- **Distancia Euclidiana Cuadrada.** Uno puede desear elevar al cuadrado la Distancia Euclidiana Standard para ponderar progresivamente más objetos que están más lejos. Esta distancia se calcula así:

$$d(x, y) = \sum_i (x_i - y_i)^2$$

2. Escoger la medida de asociación

- ❑ **Distancia City-block (Manhattan).** Esta distancia es simplemente el promedio de las diferencias a lo largo de las dimensiones. En la mayoría de los casos, se obtienen resultados similares que con el método de la distancia Euclidiana. La distancia city-block se calcula así:

$$d(x, y) = \sum_i |x_i - y_i|$$

- ❑ **Distancia de Chebychev.** Esta distancia puede ser apropiada en casos cuando uno quiere definir si son diferentes en alguna dimensión. La distancia de Chebychev se calcula así:

$$d(x, y) = \max |x_i - y_i|$$

2. Escoger la medida de asociación

- ❑ **Distancia potencia.** Algunas veces uno puede desear incrementar o disminuir progresivamente el peso que se coloca en las dimensiones en los cuales los respectivos objetos son muy diferentes. Esto se puede lograr vía la *distancia potencia*. La distancia se calcula así:

$$d(x, y) = \sqrt[r]{\sum_i |x_i - y_i|^p}$$

- ❑ **Porcentaje de desacuerdo.** Esta distancia es particularmente útil si los datos incluidos en el análisis son de naturaleza categóricos. Esta distancia se calcula de la siguiente manera:

$$d(x, y) = \frac{(\text{número de } x_i \neq y_i)}{i}$$

El propósito es construir una *matriz de las distancias iniciales* a partir de cada par de individuos de la muestra considerando las variables seleccionadas para el estudio.

2. Escoger la medida de asociación

- ❑ **EJEMPLO:** Obtener la matriz de distancia inicial usando primero la distancia euclidiana y en un segundo intento usando la distancia de Manhattan para la tabla de datos siguientes que representa 7 casos y 5 variables medidas:

Original Data					
Case	x_1	x_2	x_3	x_4	x_5
1	7	10	9	7	10
2	9	9	8	9	9
3	5	5	6	7	7
4	6	6	3	3	4
5	1	2	2	1	2
6	4	3	2	3	3
7	2	4	5	2	5

- ❑ Distancia Euclidiana

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + (x_5 - y_5)^2}$$

$$d(2, 1) = \sqrt{(9 - 7)^2 + (9 - 10)^2 + (8 - 9)^2 + (9 - 7)^2 + (9 - 10)^2} = \mathbf{3.32}$$

$$d(7, 2) = \sqrt{(2 - 9)^2 + (4 - 9)^2 + (5 - 8)^2 + (2 - 9)^2 + (5 - 9)^2} = \mathbf{12.17}$$

$$d(5, 4) = \sqrt{(1 - 6)^2 + (2 - 6)^2 + (2 - 3)^2 + (1 - 3)^2 + (2 - 4)^2} = \mathbf{7.07}$$

2. Escoger la medida de asociación

Caso	1	2	3	4	5	6	7
1	0.00	3.32	6.86	10.25	15.78	13.11	11.27
2	3.32	0.00	6.63	10.20	16.19	13.00	12.17
3	6.86	6.63	0.00	6.00	10.10	7.28	6.32
4	10.25	10.20	6.00	0.00	7.07	3.87	5.10
5	15.78	16.19	10.10	7.07	0.00	3.87	4.90
6	13.11	13.00	7.28	3.87	3.87	0.00	4.36
7	11.27	12.17	6.32	5.10	4.90	4.36	0.00

El resultado es una matriz simétrica que contiene todas la distancia de todos los posibles pares de observaciones. De la matriz anterior se observa que los caso 1 y 2 son el par con mayor similitud ya que tiene la distancia más corta. La pareja con mayor disimilitud se haya con los casos 2 y 5 ya que su distancia es la mayor.

Al observar el caso 7, se puede concluir que el orden de similitud con respecto a los demás casos es el siguiente: 6, 5, 4, 3, 1, 2.

2. Escoger la medida de asociación

□ Distancia de Manhattan

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| + |x_4 - y_4| + |x_5 - y_5|$$

$$d(2, 1) = |9 - 7| + |9 - 10| + |8 - 9| + |9 - 7| + |9 - 10| = \mathbf{7}$$

$$d(7, 2) = |2 - 9| + |4 - 9| + |5 - 8| + |2 - 9| + |5 - 10| = \mathbf{26}$$

$$d(5, 4) = |1 - 6| + |2 - 6| + |2 - 3| + |1 - 3| + |2 - 4| = \mathbf{14}$$

Matriz de distancias resultante

Caso	1	2	3	4	5	6	7
1	0.00	7.00	13.00	21.00	35.00	28.00	25.00
2	7.00	0.00	14.00	22.00	36.00	29.00	26.00
3	13.00	14.00	0.00	12.00	22.00	15.00	12.00
4	21.00	22.00	12.00	0.00	14.00	7.00	10.00
5	35.00	36.00	22.00	14.00	0.00	7.00	10.00
6	28.00	29.00	15.00	7.00	7.00	0.00	9.00
7	25.00	26.00	12.00	10.00	10.00	9.00	0.00

2. Escoger la medida de asociación

Para cualitativas dicotómicas (dos categorías):

A partir de la tabla de contingencia de 2×2

		Variable X_i		
		Presente (1)	Ausencia (0)	Total
Variable X_j	Presente (1)	a	b	a+b
	Ausencia (0)	c	d	c+d
	Total	a+c	b+d	N=a+b+c+d

a = # de individuos que toman el valor de 1 en cada variable de forma simultánea.

b = # de individuos que toman el valor de 0 en X_i y 1 en X_j .

c = # de individuos que toman el valor de 1 en X_i y 0 en X_j .

d = # de individuos que toman el valor de 0 en cada variable de forma simultánea.

2. Escoger la medida de asociación

Medidas de similitud

	Similitud o similaridad	Disimilaridad
Russel y Rao	$RR = \frac{a}{N}$	$\frac{N - a}{N}$
Parejas simples	$PS = \frac{a + d}{N}$	$\frac{N - (a + d)}{N}$
Jaccard	$J = \frac{a}{a + b + c}$	$\frac{b + c}{a + b + c}$
Dice y Sorensen	$D = \frac{2a}{2a + b + c}$	$\frac{b + c}{2a + b + c}$

Los valores se obtiene de la tabla de contingencia anterior.

2. Escoger la medida de asociación

Ejemplo: La siguiente tabla reúne la presencia/ausencia de 6 especímenes de bacterias en 7 lagos donde 1 indica presencia del espécimen y 0 indica ausencia.

Lago	Especie 1	Especie 2	Especie 3	Especie 4	Especie 5	Especie 6
La Irene	1	1	1	1	1	1
Pedro Luro	0	0	0	1	1	1
Loncoche	1	0	0	0	0	0
Lefipan	1	1	1	1	0	1
Salamanca	1	1	1	1	1	1
Dorotea	0	0	0	0	0	1
Paso del sapo	0	0	1	0	0	1

Paso 1 Generar la tabla de contingencia de entre dos pares de lagos. Se elige los dos primeros registros de la tabla

Lago	E 1	E 2	E 3	E 4	E 5	E 6
La Irene	1	1	1	1	1	1
Pedro Luro	0	0	0	1	1	1



		La Irene		Total
		Presente (1)	Ausente (0)	
Pedro Luro	Presente (1)	3	0	3
	Ausente (0)	3	0	3
Total		6	0	6

2. Escoger la medida de asociación

Paso 2 Seleccionar y calcular el indicador de medida de similitud. En este ejemplo se emplea el índice de Jaccard.

$$J = \frac{a}{a + b + c} = \frac{3}{3 + 0 + 3} = 0.500$$

Paso 3 Repetir paso 1 y 2 con todos los posibles emparejamientos.

Lago	E 1	E 2	E 3	E 4	E 5	E 6
Salamanca	1	1	1	1	1	1
Dorotea	0	0	0	0	0	1



		Salamanca		Total
		Presente (1)	Ausente (0)	
Dorotea	Presente (1)	1	0	1
	Ausente (0)	5	0	5
Total		6	0	6

$$J = \frac{a}{a + b + c} = \frac{1}{1 + 0 + 5} = 0.167$$

Paso 4 Construir la matriz de similitud con todos los índices calculados

2. Escoger la medida de asociación

Matriz de distancias resultante

	La Irene	Pedro Luro	Loncoche	Lefipan	Salamanca	Cerro Dorotea	Paso del sapo
La Irene	1.000	0.500	0.167	0.833	1.000	0.167	0.333
Pedro Luro	0.500	1.000	0.000	0.333	0.500	0.333	0.250
Loncoche	0.167	0.000	1.000	0.200	0.167	0.000	0.000
Lefipan	0.833	0.333	0.200	1.000	0.833	0.200	0.400
Salamanca	1.000	0.500	0.167	0.833	1.000	0.167	0.333
Cerro Dorotea	0.167	0.333	0.000	0.200	0.167	1.000	0.500
Paso del sapo	0.333	0.250	0.000	0.400	0.333	0.500	1.000

2. Escoger la medida de asociación

Transformación de valores:

Es importante recordar que debido a que el análisis de conglomerados emplea medidas de distancia, estas son muy sensibles a las diferencias de escala o magnitudes de las variables.

La forma mas común de transformación es la conversión de cada variable en su puntuación estándar (Z-score):

$$Z = \frac{x_i - \bar{x}}{S}$$

De esta forma se convierte los valores en puntuaciones estandarizadas con media = 0 y desviación estándar = 1, eliminado el sesgo originado por la diferencia de escalas.



2. Escoger la medida de asociación

Otras formas de transformación son:

- a) **rango -1 a 1**, los valores originales son divididos entre el rango de cada variable o caso,
- b) **rango 0 a 1**, se obtiene restando el mínimo y dividiendo por el rango de cada variable o caso,
- c) **magnitud máxima de 1**, se obtiene dividiendo los valores originales por el máximo de cada variable o caso, según corresponda al análisis,
- d) **media de 1**, se dividen los valores originales entre la media de cada variable o caso,
- e) **desviación típica 1**, se obtiene dividiendo los valores originales por la desviación típica de cada variable o caso.

2. Escoger la medida de asociación

Encuesta en la ciudad de Aguascalientes entre la población económicamente inactiva

EDAD	ESTUDIANTES	HOGAR	JUBILADO	INCAP	OTRO	TOTAL
12 - 14	39,949	5,988	6	78	3,059	49,080
15 - 19	28,636	16,253	22	178	5,398	50,487
20 - 24	6,580	19,582	27	225	2,896	29,310
25 - 29	834	19,180	32	196	1,667	21,909
30 - 34	188	17,461	57	174	1,198	19,078
35 - 39	68	14,446	109	143	1,002	15,768
40 - 44	44	11,698	184	135	800	12,861
45 - 49	29	9,725	339	158	794	11,045
50 - 54	14	7,915	579	149	735	9,392
55 - 59	10	6,773	820	145	742	8,490
60 - 64	13	6,152	1,311	213	819	8,508
65 Y MAS	27	12,867	3,843	1,411	4,020	22,168

Como se aprecia en la tabla anterior, aquellas variables con una escala de medición mucho mayor tendrán una mayor influencia en el cálculo de las distancias entre pares de observaciones. Por lo tanto es recomendable hacer una transformación de los datos.

3. Elección de la técnica de agrupación

La elección del método de agrupación será relativamente natural dependiendo de la naturaleza de los datos usados y de los objetivos perseguidos.

Se recomienda probar varias técnicas y contrastar los resultados obtenidos con cada una de ellas. Si los resultados finales son parecidos, las conclusiones son mucho más válidas sobre la estructura natural de los datos. En caso contrario puede plantearse el hecho que tal vez los datos utilizados no obedezcan a una estructura bien definida.

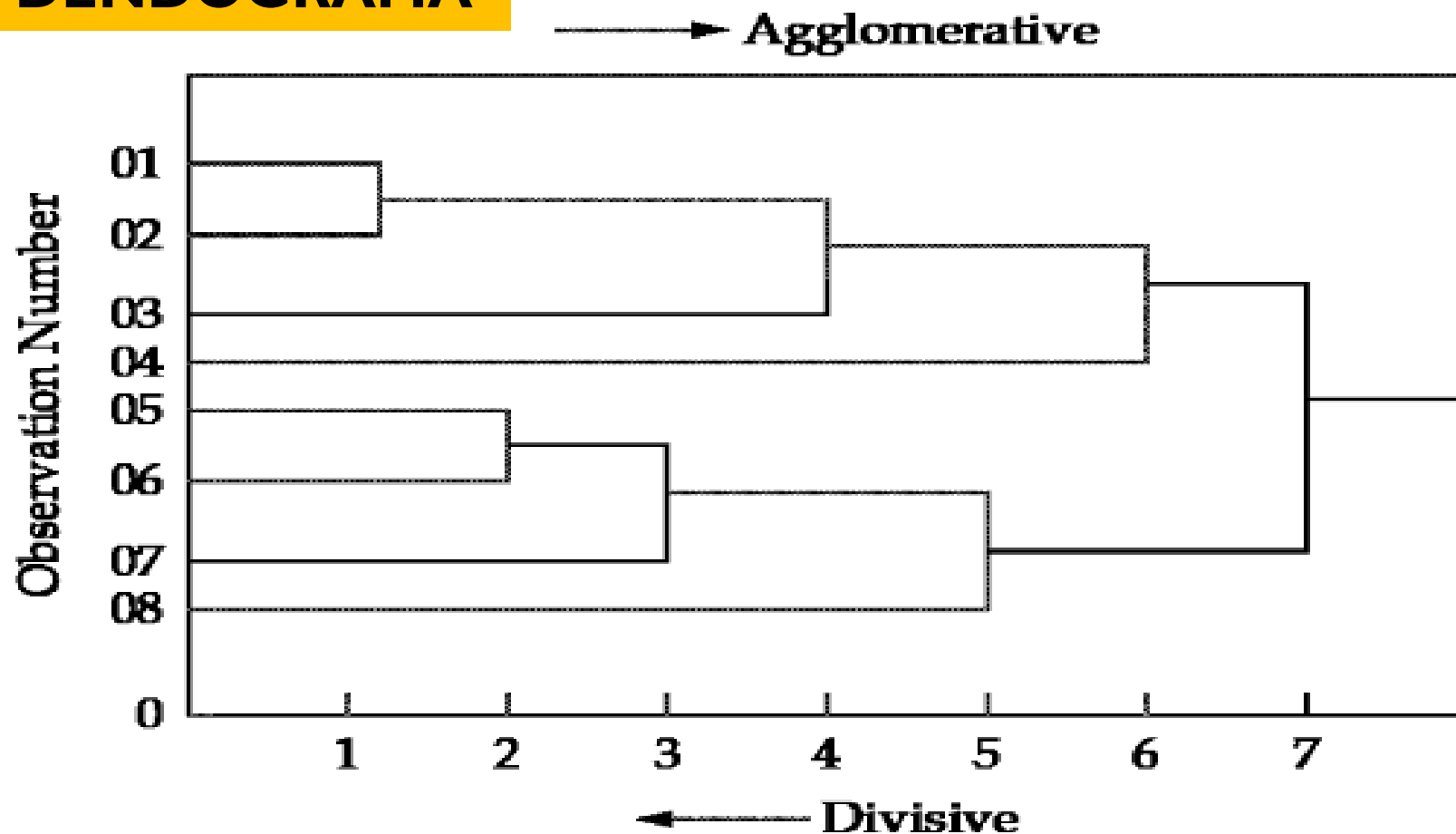
Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de DENDROGRAMA. El dendrograma:

- Permite *visualizar el proceso de agrupamiento de los clúster* en los distintos pasos.
- Ayuda a *decidir el número de grupos* que representan mejor la estructura de los datos considerando la forma en que se van anidando los clúster y la medida de similitud a la cual lo hacen.
- Permite al investigador “*seguir la pista*” de formación de los distintos clúster, que van englobándose o anidándose, hasta resumirse en sólo uno.

En biología a menudo es de mayor interés desvelar las distintas categorías en que van clasificándose los individuos estudiados, desde los grupos más particulares a los más generales.

Modelo de factor único

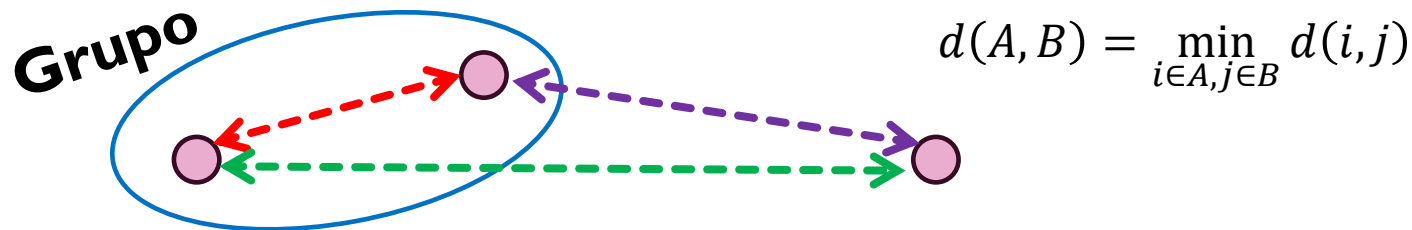
DENDOGRAMA



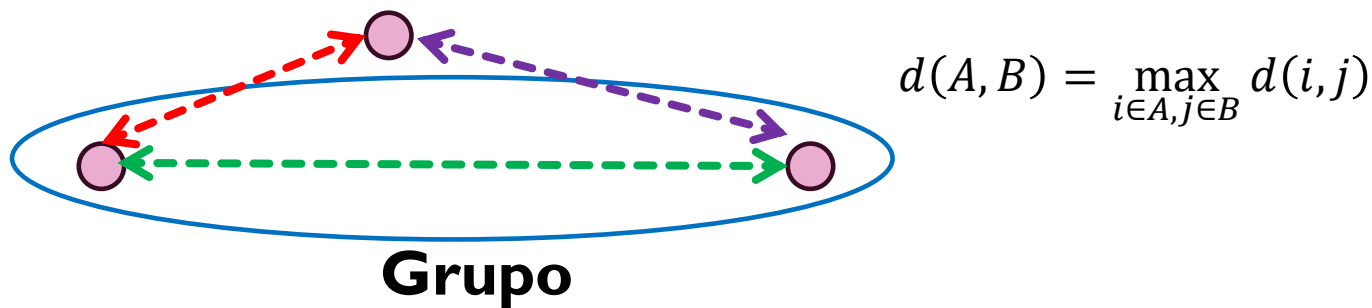
3. Elección de la técnica de agrupación

Algunas de las técnicas de agrupación más empleadas son:

- **Vecino más cercano** (*linkage simple*): Agrupa a los individuos con la *distancia* o *similaridad* más próxima.



- **Vecino más lejano** (*linkage completo*): Agrupa a los individuos con la *distancia* o *similaridad* más lejana.



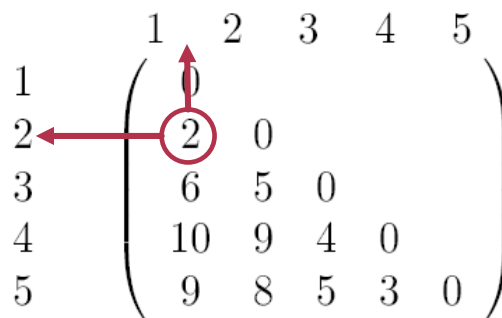
3. Elección de la técnica de agrupación

Ejemplo: A partir de la siguiente matriz de distancia emplear la técnicas de agrupación anteriores y obtener el dendograma respectivo.

$$\begin{array}{c} \begin{matrix} & 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

Vecino más cercano

Paso 1. Seleccionar la distancia más corta de la matriz para formar un grupo. De la matriz anterior se observa que es la $d(2, 1) = 2$, por lo tanto la observación 1 y 2 forman un clúster.


$$\begin{array}{c} \begin{matrix} & 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

3. Elección de la técnica de agrupación

Paso 2. Calcular la nueva matriz de distancias. Las distancia de los individuos que no fueron agrupados no varían, solamente aquellas que tienen relación con los individuos que son agrupados.

	(1-2)	3	4	5
(1-2)	0			
3		0		
4		4	0	
5		5	3	0

$$\begin{aligned}d([1,2], 3) &= \min(d(1,3), d(2,3)) = \min(6, 5) = 5 \\d([1,2], 4) &= \min(d(1,4), d(2,4)) = \min(10, 9) = 9 \\d([1,2], 5) &= \min(d(1,5), d(2,5)) = \min(9, 8) = 8\end{aligned}$$



	(1-2)	3	4	5
(1-2)	0			
3	5	0		
4	9	4	0	
5	8	5	3	0

Paso 3. Repetir los pasos 1 y 2 hasta agrupar a todos los individuos en un único grupo.

3. Elección de la técnica de agrupación

	(1-2)	3	4	5
(1-2)	0			
3	5	0		
4	9	4	0	
5	8	5	3	0



	(1-2)	3	(4-5)
(1-2)	0		
3	5	0	
(4-5)			0

$$d([1,2], [4,5]) = \min(d([1,2], 4), d([1,2], 5)) = \min(9, 8) = 8$$

$$d(3, [4,5]) = \min(d(3,4), d(3,5)) = \min(4, 5) = 4$$



	(1-2)	3	(4-5)
(1-2)	0		
3	5	0	
(4-5)	8	4	0



	(1-2)	3	(4-5)
(1-2)	0		
3	5	0	
(4-5)	8	4	0



	(1-2)	(3-4-5)
(1-2)	0	
(3-4-5)		0

$$d([1,2], [3,4,5])$$

$$= \min(d([1,2], 3), d([1,2], [4,5]))$$

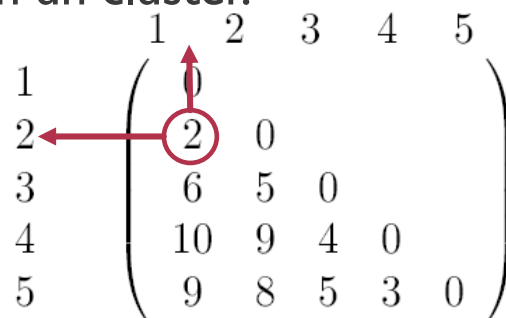
$$= \min(5, 8) = 5$$

	(1-2)	(3-4-5)
(1-2)	0	
(3-4-5)	5	0

3. Elección de la técnica de agrupación

Vecino más lejano

Paso 1. Seleccionar la distancia más corta de la matriz para formar un grupo. De la matriz anterior se observa que es la $d(2, 1) = 2$, por lo tanto la observación 1 y 2 forman un clúster.



	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Paso 2. Calcular la nueva matriz de distancias. Al igual que en el método anterior, las distancia de los individuos que no fueron agrupados no varían, solamente aquellas que tienen relación con los individuos que son agrupados.

	(1-2)	3	4	5
(1-2)	0			
3		0		
4		4	0	
5		5	3	0

$$\begin{aligned}d([1,2], 3) &= \max(d(1,3), d(2,3)) = \max(6, 5) = 6 \\d([1,2], 4) &= \max(d(1,4), d(2,4)) = \max(10, 9) = 10 \\d([1,2], 5) &= \max(d(1,5), d(2,5)) = \max(9, 8) = 9\end{aligned}$$

3. Elección de la técnica de agrupación

	(1-2)	3	4	5
(1-2)	0			
3	6	0		
4	10	4	0	
5	9	5	3	0

Paso 3. Repetir los pasos 1 y 2 hasta agrupar a todos los individuos en un único grupo.

	(1-2)	3	4	5
(1-2)	0			
3	6	0		
4	10	4	0	
5	9	5	0	0



	(1-2)	3	(4-5)
(1-2)	0		
3	6	0	
(4-5)			0

$$d([1,2], [4,5]) = \max(d([1,2], 4), d([1,2], 5)) = \max(10, 9) = 10$$

$$d(3, [4,5]) = \max(d(3,4), d(3,5)) = \max(4, 5) = 5$$

3. Elección de la técnica de agrupación

	(1-2)	3	(4-5)
(1-2)	0		
3	6	0	
(4-5)	10	5	0



	(1-2)	3	(4-5)
(1-2)	0		
3	6	0	
(4-5)	10	5	0



$$d([1,2], [3,4,5]) = \max(d([1,2], 3), d([1,2], [4,5])) \\ = \max(6, 10) = 10$$

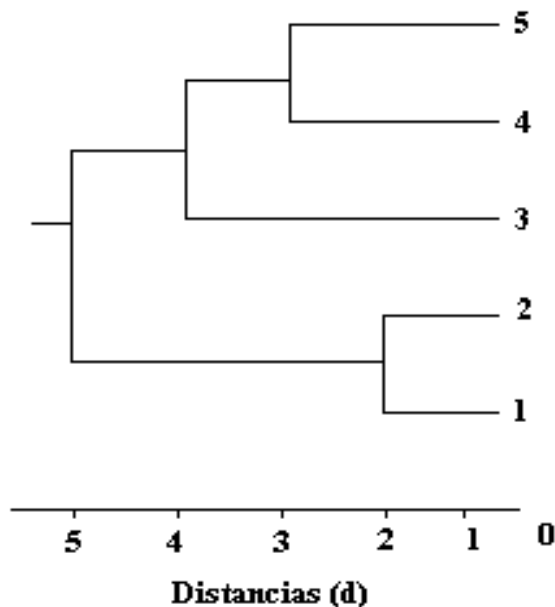
	(1-2)	(3-4-5)
(1-2)	0	
(3-4-5)		0



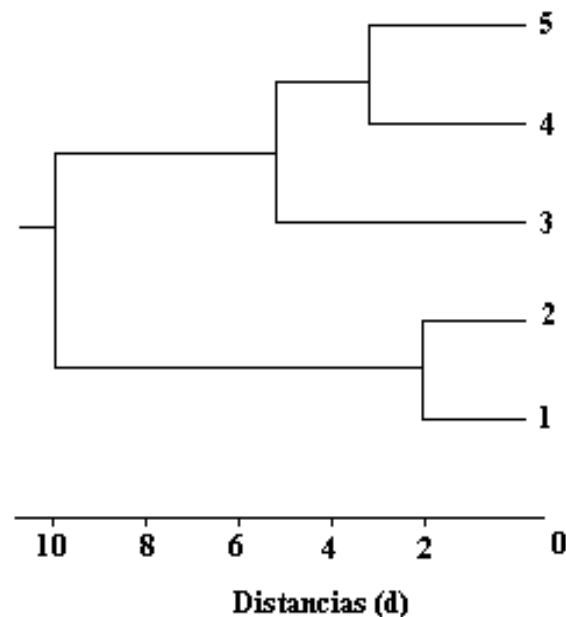
	(1-2)	(3-4-5)
(1-2)	0	
(3-4-5)	10	0

3. Elección de la técnica de agrupación

□ Ejemplo:



Vecino más cercano

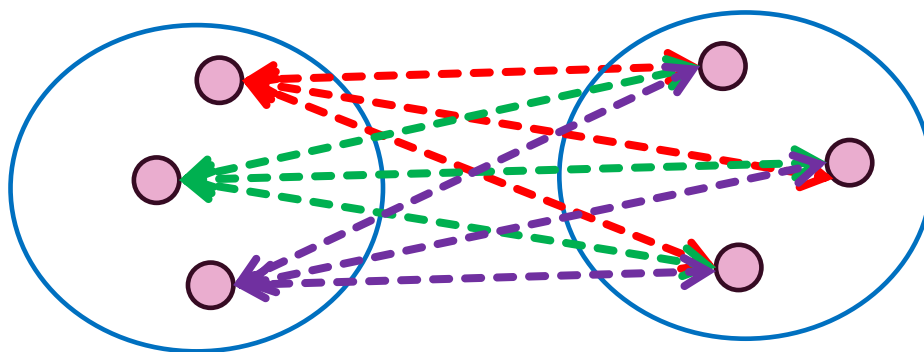


Vecino más lejano

La distancia (d) en ambos dendogramas indica el valor en que los individuos se fueron agrupando. En el caso del vecino más cercano la agrupación total ocurrió a una distancia de 5 unidades mientras que en el vecino más lejano se alcanzó hasta la distancia de 10 unidades.

3. Elección de la técnica de agrupación

- ❑ **Agrupamiento promedio:** Promedio de las distancias entre todos los pares de individuos.



$$d(A, B) = \frac{1}{n_A n_B} \sum_{i \in A, j \in B} d(i, j)$$

n_A = # individuos del grupo A

n_B = # individuos del grupo B

- ❑ **Centroide (centro de gravedad):** Con este método, una vez formados los grupos, son representados por su vector medio, y las distancias entre-grupos son ahora definidas en términos de distancias entre dos vectores medios.

$$d(A, B) = d(\bar{x}_A, \bar{x}_B)$$

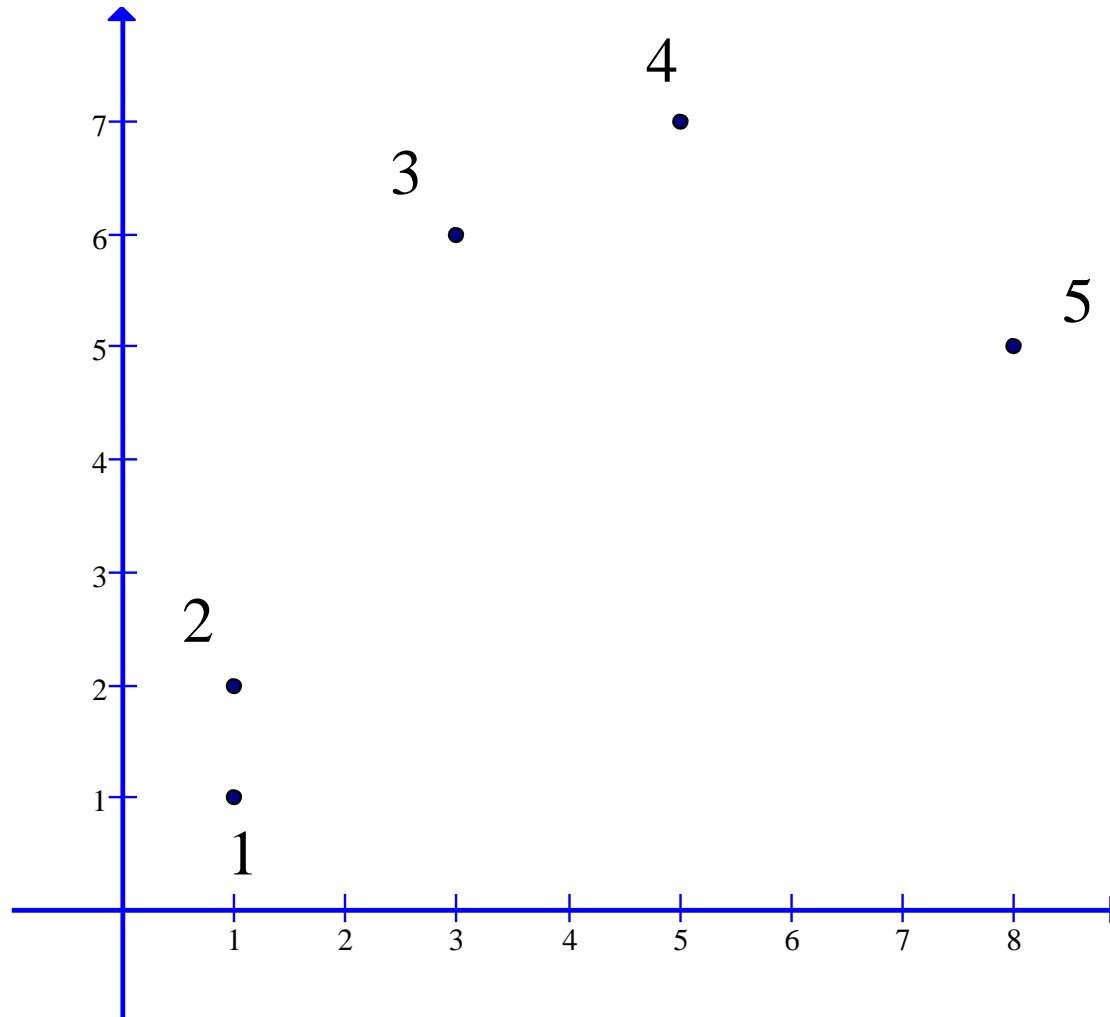
\bar{x}_A = centroide del grupo A

\bar{x}_B = centroide del grupo B

3. Elección de la técnica de agrupación

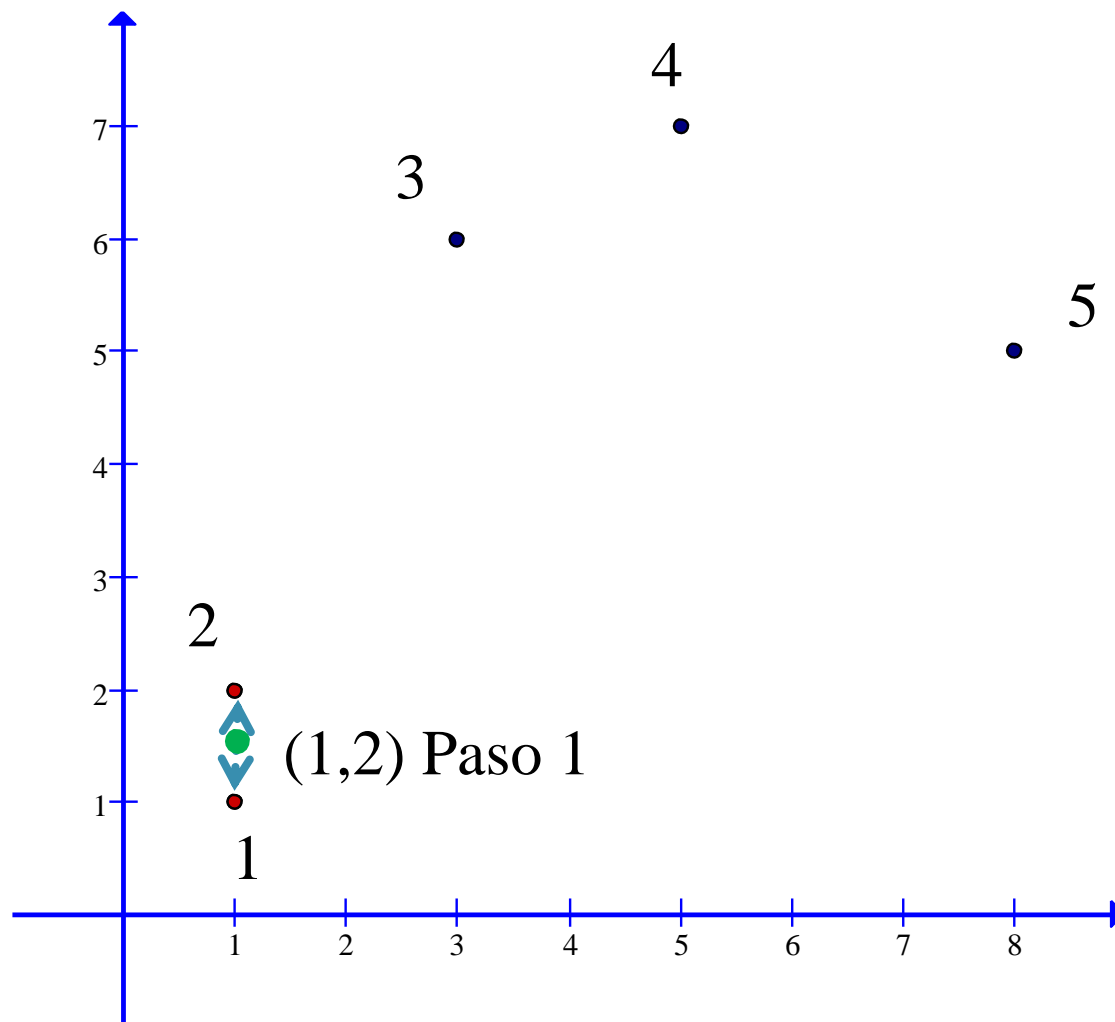
Centroide

Calcular centroides iniciales



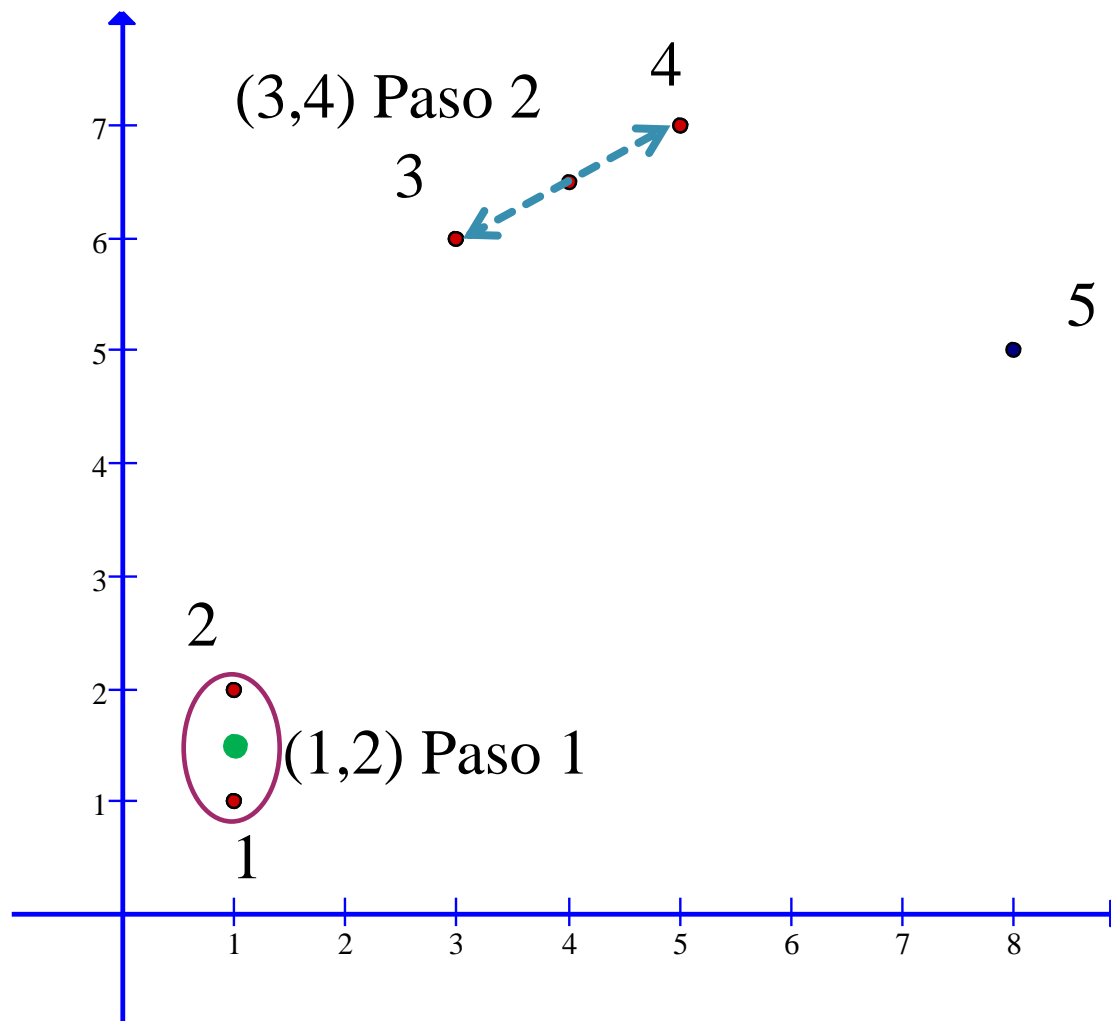
3. Elección de la técnica de agrupación

Seleccionar la pareja de menor distancia. En este ejemplo caso 1 y 2



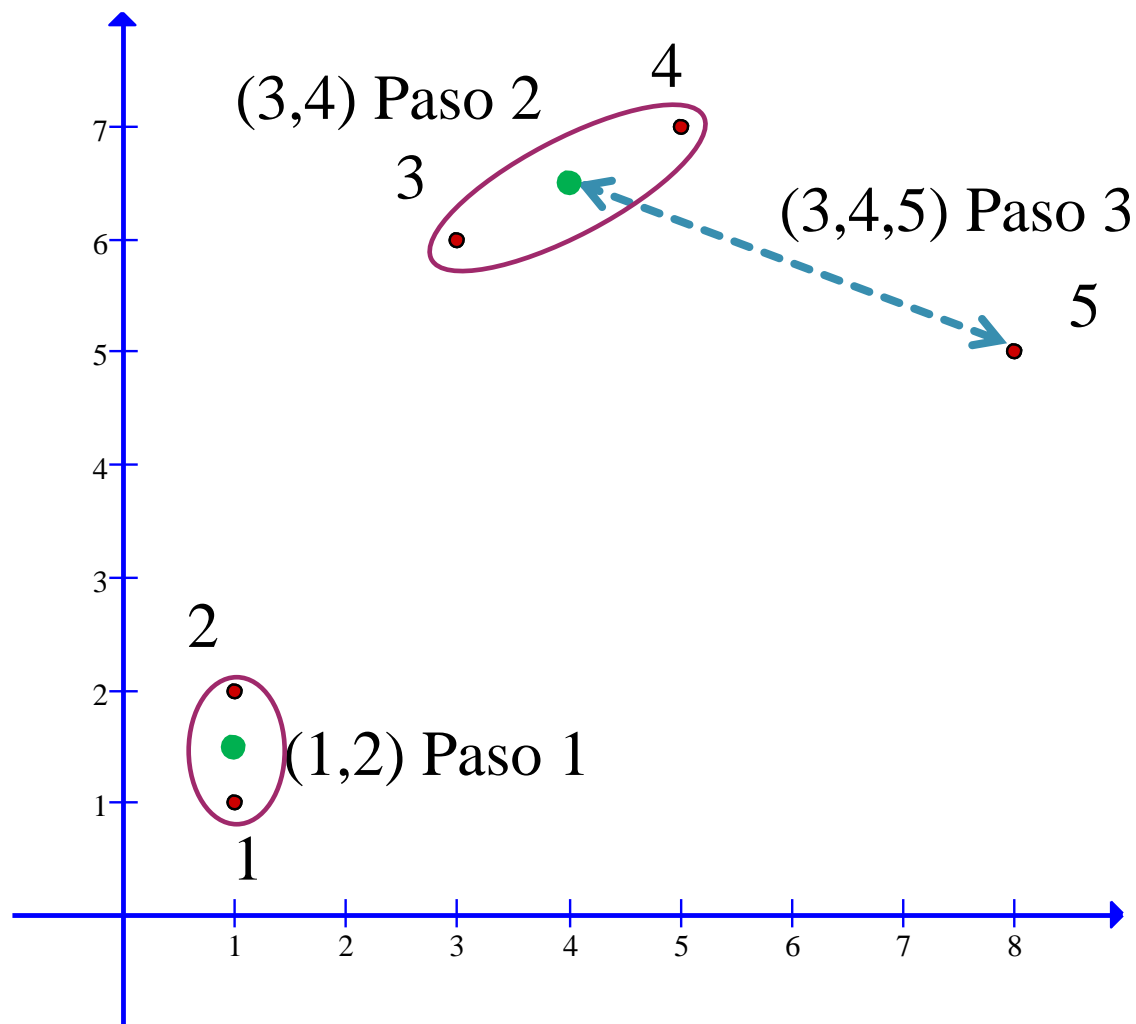
3. Elección de la técnica de agrupación

Calcular el centroide de la agrupación $\{1,2\}$, recalculer las distancias de los casos restantes hacia el nuevo centroide y seleccionar el de menor distancia. En el ejemplo $\{3,4\}$

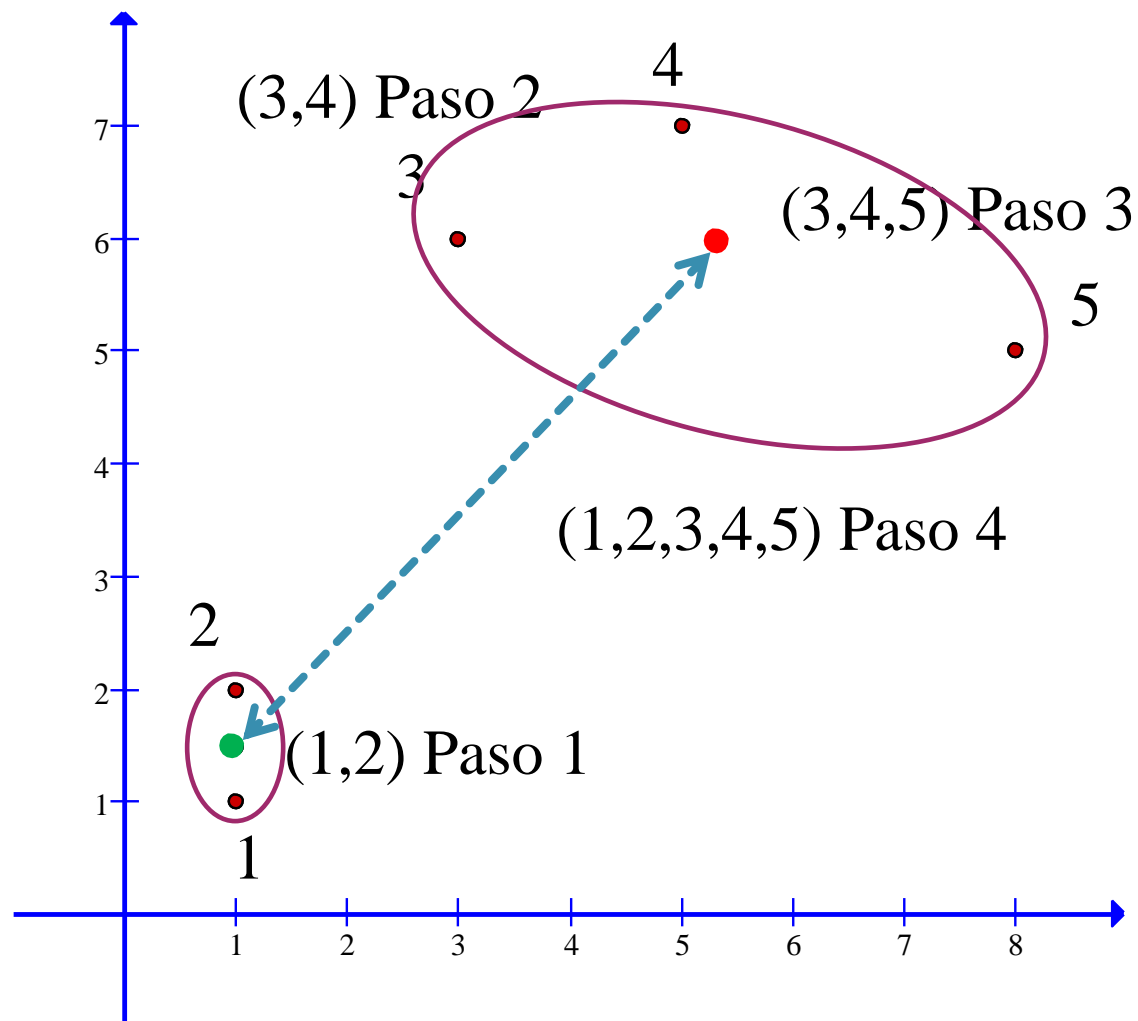


3. Elección de la técnica de agrupación

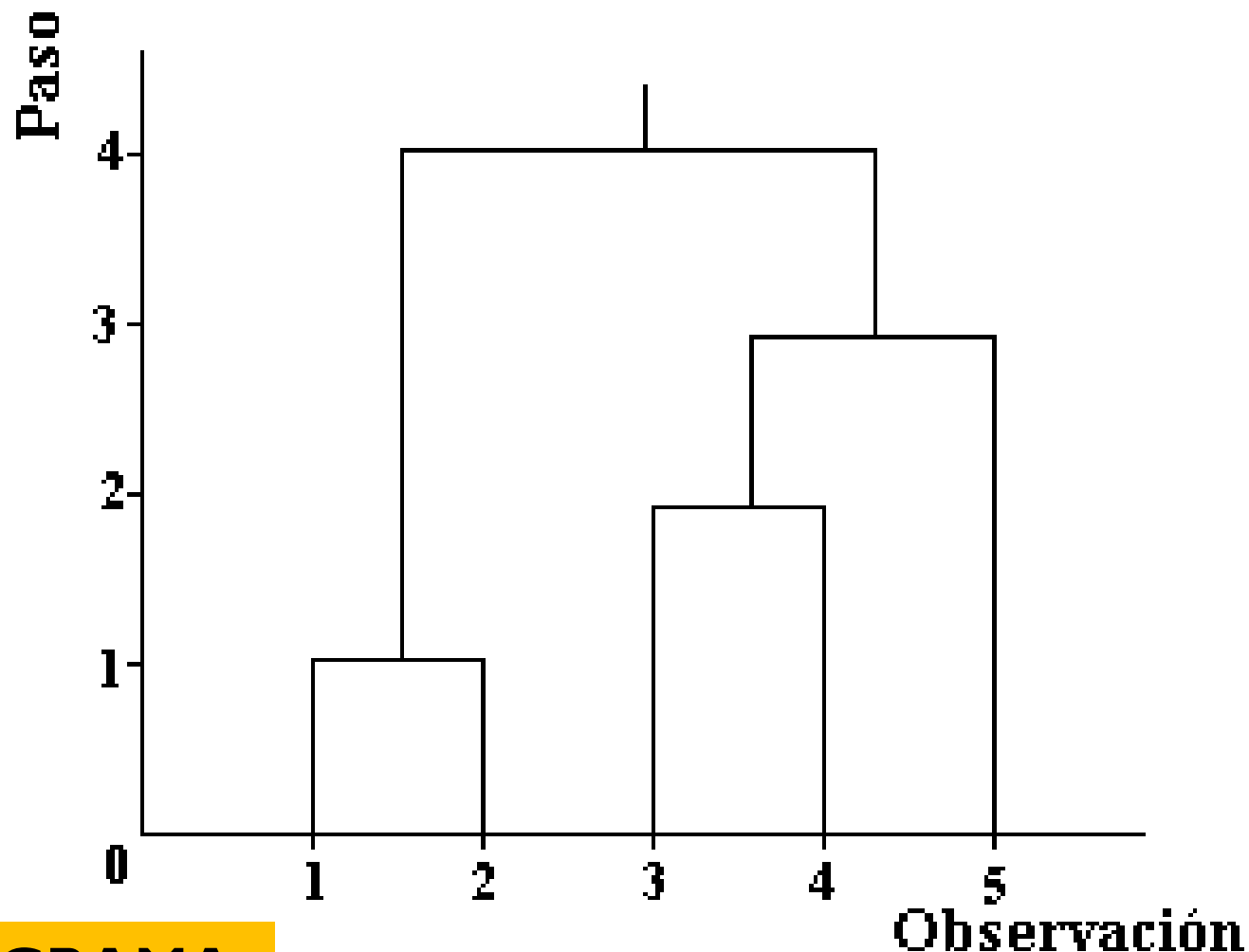
Repetir la mecánica anterior hasta agrupar todos los casos en un solo grupo.



3. Elección de la técnica de agrupación



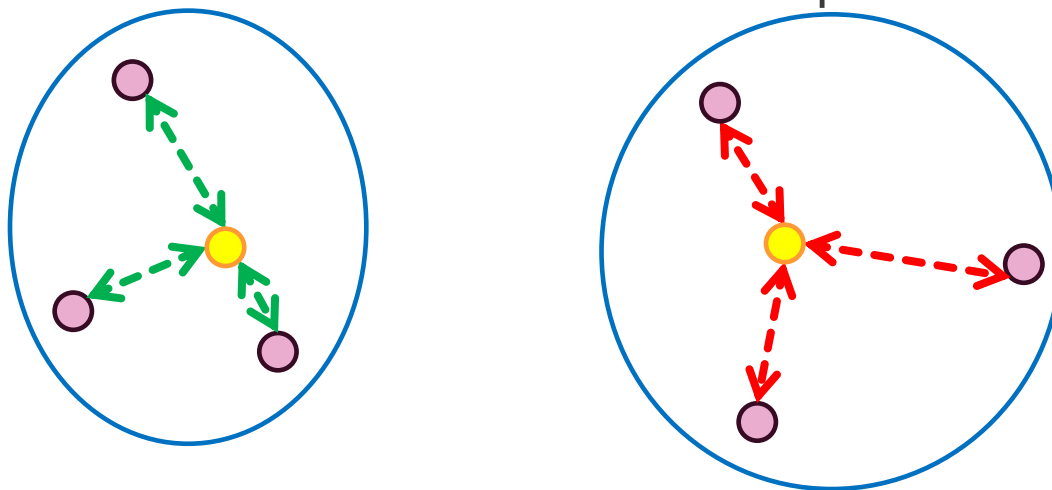
3. Elección de la técnica de agrupación



DENDOGRAMA

3. Elección de la técnica de agrupación

- **Ward:** procedimiento en el cual, en cada etapa, se unen los dos clústers para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias (E), dentro de cada clúster, de cada individuo al centroide del clúster. Es de los más empleados



$$E = \sum_{k=1}^h E_k$$
$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^n (x_{ij}^k - m_j^k)^2$$

x_{ij}^k = valor de la j -ésima variable sobre el i -ésimo clúster, suponiendo que posee n_k individuos.

m^k = centroide del clúster k con componentes m_j^k

E_k = suma del cuadrado de los errores del clúster (distancia euclidiana al cuadrado entre el individuo del clúster k a su centroide)

E = suma de los cuadrados de los errores para todas los h clúster

4. Validación e interpretación de resultados

Al plantear métodos jerárquicos se plantean los siguientes problemas:

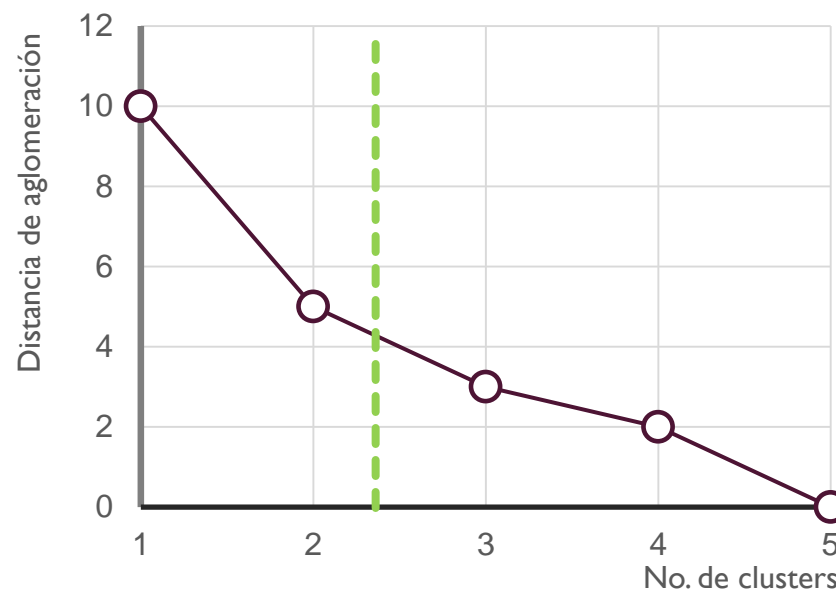
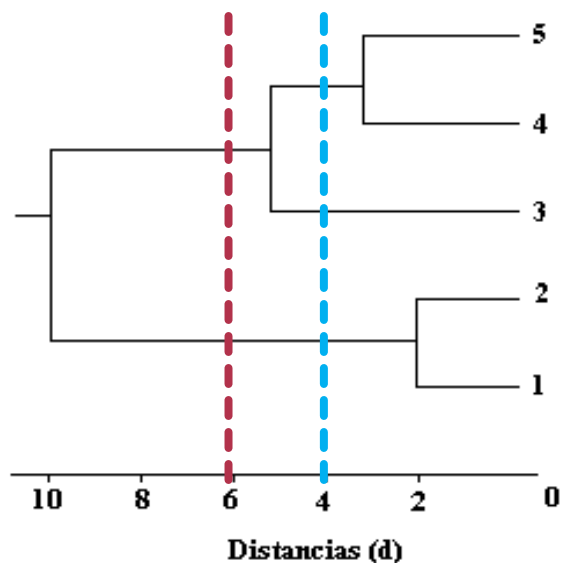
- a) ¿Cuál es el número idóneo de clústers que mejor representa la estructura natural de los datos?
- b) ¿En qué medida representa la estructura final las similitudes o diferencias entre los objetos?

Para contestar la primera pregunta, no existe una norma fija para establecer cuántos grupos pueden considerarse. Algunas estrategias empleados son:

- 1) Cortar el dendrograma. El dendrograma puede servir de ayuda visual para determinar dicho número dependiendo del coeficiente de proximidad usado. Este procedimiento generalmente presenta sesgos por la opinión y conocimiento que tiene el investigador sobre los datos.
- 2) Emplear un gráfico donde se represente la distancia de aglomeración y el número de grupos a esa distancia. En los primeros pasos la distancia es generalmente grande, mientras que los últimos es pequeño. El punto de corte será aquel en el que dejen de producirse saltos bruscos.

4. Validación e interpretación de resultados

Ejemplo:



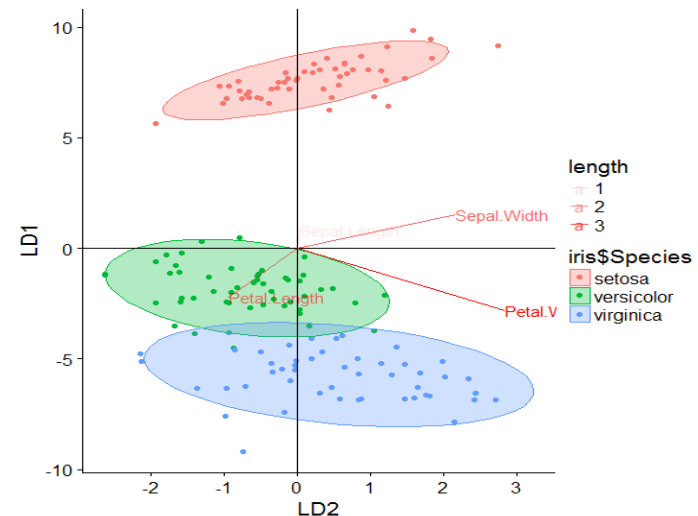
Si se utiliza el método del dendrograma, bastara con trazar una línea que corte el árbol, si se desea cortar a una distancia de 6 (línea roja) se tienen dos clústers: $\{1, 2\}$ y $\{3, 4, 5\}$. Pero si se decide cortar a una distancia de 4 (línea azul) se tienen 3 agrupaciones: $\{1, 2\}$, $\{3\}$ y $\{4, 5\}$. En el caso del gráfico, se aprecia que al pasar de 1 a 2 se tiene un salto brusco pero ya no al pasar de 2 a 3 por lo que se considerarían dos grupos.

4. Validación e interpretación de resultados

Para validar la pertinencia de la estructura seleccionada (número de agrupaciones elegidas) es necesario caracterizar cada grupo a partir de los individuos que fueron integrados. Para ello se realiza un análisis descriptivo de cada uno de los grupos en las variables que fueron consideradas para realizar el análisis.

Otra opción es utilizar alguna técnica grafica, como el grafico de dispersión o el grafico biplot, representando a los individuos de cada agrupación

En caso de no poder diferenciar de forma adecuada a cada uno de los grupos, se debe replantear el numero de grupos seleccionados e incluso si las variables medidas para realizar el análisis son las adecuadas.



Ventajas y desventajas

Ventajas:

- a) No requiere hacer inferencias previas sobre el número de clústers.
- b) Permite representar la secuencia de agrupaciones en forma de árbol (dendograma).

Desventajas:

- a) En ocasiones es alto el costo computacional.
- b) Sensible respecto a las primeras agrupaciones a realizar.
- c) Complicado de interpretar cuando el número de elementos a clasificar es grande.