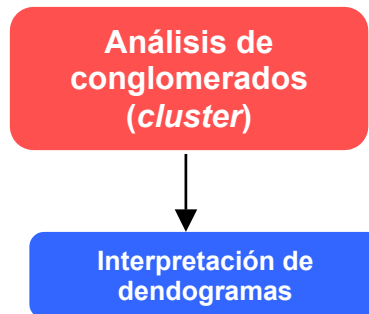


ANÁLISIS DE CONGLOMERADOS

Autor: Manuel Terrádez Gurrea (mterradez@uoc.edu).

ESQUEMA DE CONTENIDOS



INTRODUCCIÓN

El **análisis de conglomerados (cluster)** es una técnica multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencias entre los grupos.

Nos basaremos en los **algoritmos jerárquicos acumulativos** (forman grupos haciendo conglomerados cada vez más grandes), aunque no son los únicos posibles.

El **dendograma** es la representación gráfica que mejor ayuda a interpretar el resultado de un análisis *cluster*.

El análisis de conglomerados se puede combinar con el Análisis de Componentes Principales, ya que mediante ACP se puede homogeneizar los datos, lo cual permite realizar posteriormente un análisis *cluster* sobre los componentes obtenidos.

OBJETIVOS

- Entender por qué es importante agrupar elementos parecidos en bloques diferentes.
- Saber aplicar el análisis de conglomerados, con ayuda de Minitab.
- Interpretar el dendograma resultante del análisis.

CONOCIMIENTOS PREVIOS

Aparte de estar iniciado en el uso del paquete estadístico Minitab, resulta muy conveniente haber leído con profundidad los siguientes *math-blocks*:

- Estadística descriptiva.
- Correlación y regresión lineal múltiple.

CONCEPTOS FUNDAMENTALES

Medidas de disimilitud

Partimos de una matriz de información que contiene las observaciones de todas las variables sobre los diferentes elementos considerados (ver Tabla 1), y calculamos las diferencias entre dichos elementos mediante alguna de las medidas de disimilitud habituales: la **distancia**

euclidiana ($\sqrt{\sum_{j=1}^J (X_{rj} - X_{sj})^2}$), su cuadrado, la **distancia de City-Block** ($\sum_{j=1}^J |X_{rj} - X_{sj}|$),

la de Mahalanobis, la de Minkowski, la de Tchebychef, etc. Todas ellas proporcionan ordenaciones muy similares de las distancias en casi todos los casos.

Tabla 1

Elementos	X_1	X_2	...	X_J
1	X_{11}	X_{12}	...	X_{1J}
2	X_{21}	X_{22}	...	X_{2J}
...
K	X_{K1}	X_{K2}	...	X_{KJ}

Algoritmos de clasificación

Para clasificar los elementos en clusters utilizaremos **algoritmos jerárquicos**, que pueden ser **acumulativos** (se forman grupos haciendo *clusters* cada vez más grandes) o **disminutivos** (partiendo de un solo grupo se separan los elementos en *clusters* cada vez más pequeños).

Entre los algoritmos jerárquicos acumulativos destacan los siguientes métodos:

- Método de las distancias mínimas: se busca la mayor semejanza entre los elementos o grupos más cercanos.
- Método de las distancias máximas: se calcula la mínima distancia entre los elementos más alejados.
- Método de las distancias medias: se calcula la media de las distancias entre elementos.

□ **Presentación de los resultados**

Para representar la estructura jerárquica de la formación de los conglomerados se utiliza el **dendograma**, un gráfico que tiene forma de árbol invertido.

Así, a partir de los K elementos observados podemos identificar desde 1 hasta K *clusters*, según el número de grupo que queramos obtener, sin más que realizar la segmentación horizontal adecuada.

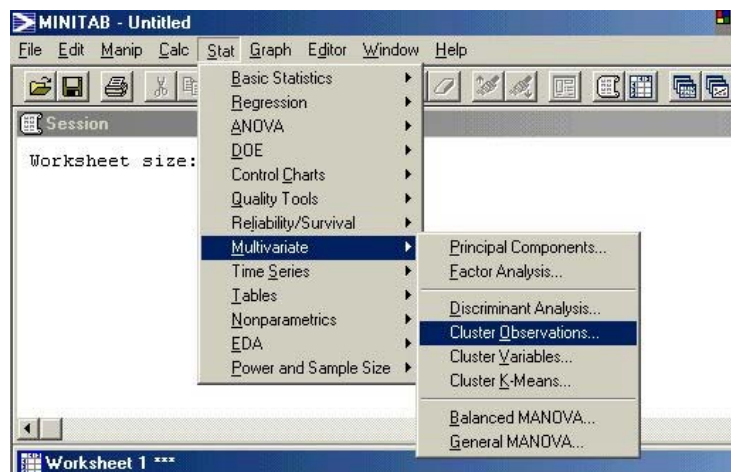
Es recomendable trabajar con datos estandarizados, para eliminar el efecto de la escala de medida, y así poder aplicar el análisis sobre variables que presentan similares valores medios y desviaciones estándar, lo cual facilita la interpretación.

CASOS PRÁCTICOS CON SOFTWARE

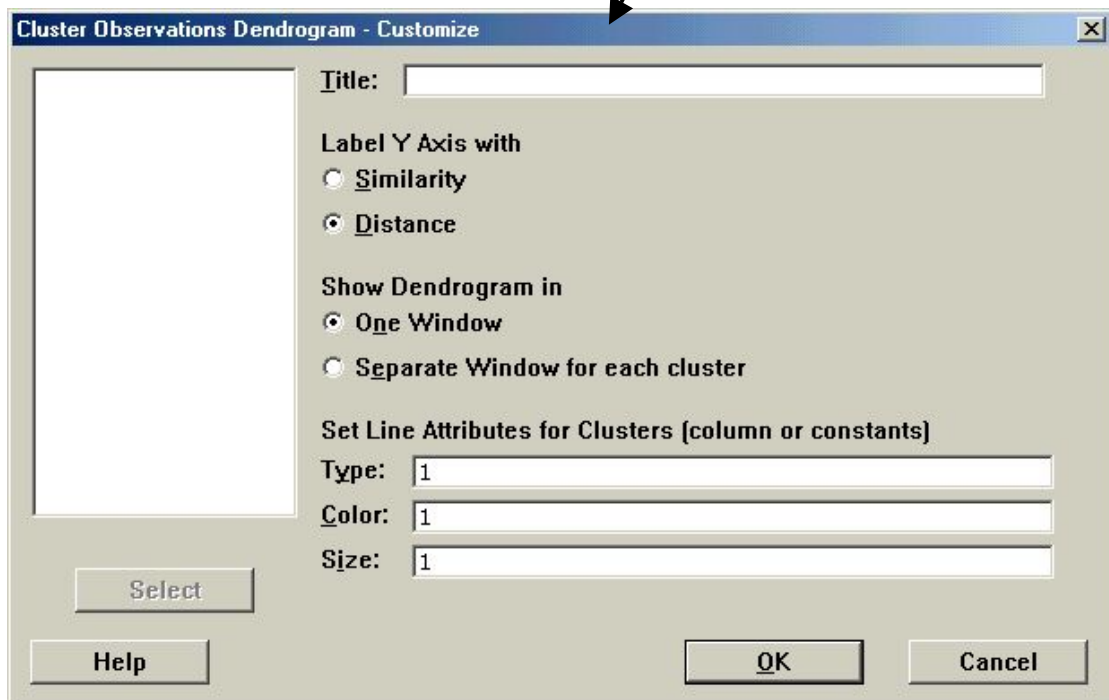
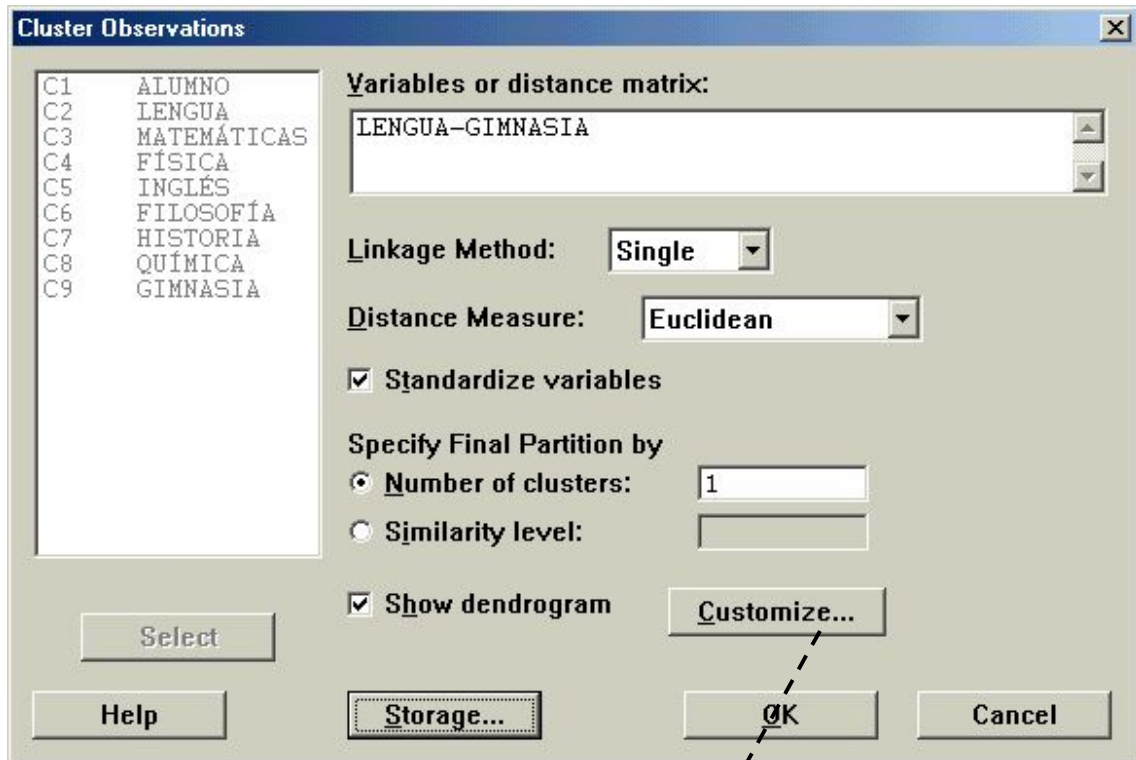
□ Calificaciones escolares

Vamos a utilizar los datos del archivo **asignaturas.mtw**, que recogen las calificaciones de los 15 alumnos de una clase en diversas asignaturas

Stat → Multivariate → Cluster Observations...



Tal y como podemos apreciar en los gráficos siguientes, solicitaremos el análisis con las variables estandarizadas, así como el dendograma (representado en función de las distancias).

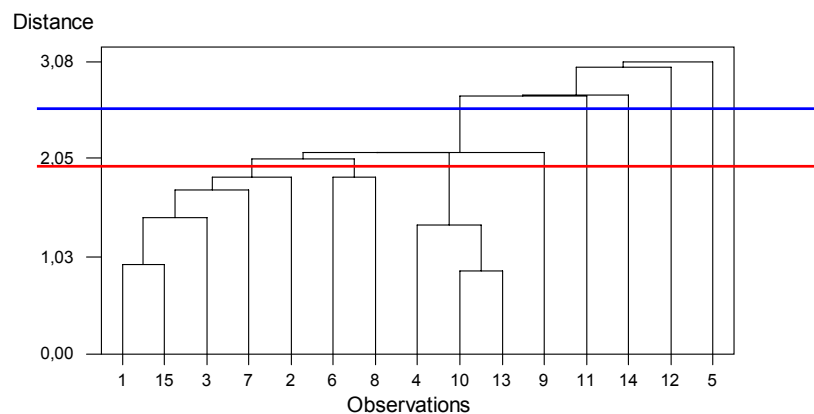


La salida que ofrece Minitab es la siguiente:

Hierarchical Cluster Analysis of Observations							
Standardized Variables, Euclidean Distance, Single Linkage							
Amalgamation Steps							
Step	Number of clusters	Similarity level	Distance level	Clusters joined	New cluster	Number of obs. in new cluster	Number of obs. cluster
1	14	88,47	0,871	10	13	10	2
2	13	87,54	0,941	1	15	1	2
3	12	82,03	1,357	4	10	4	3
4	11	80,93	1,441	1	3	1	3
5	10	77,12	1,728	1	7	1	4
6	9	75,35	1,862	1	2	1	5
7	8	75,34	1,862	6	8	6	2
8	7	72,74	2,059	1	6	1	7
9	6	71,89	2,123	1	4	1	10
10	5	71,85	2,126	1	9	1	11
11	4	64,00	2,720	1	11	1	12
12	3	63,87	2,729	1	14	1	13
13	2	59,97	3,024	1	12	1	14
14	1	59,21	3,081	1	5	1	15

Aquí se nos muestra el proceso de creación de cada *cluster*, pero no entraremos a analizarlo con detalle, ya que excede el nivel de esta asignatura.

Donde sí nos detendremos es en la interpretación del dendograma:



En el dendograma queda reflejada la formación de los conglomerados, así como las distancias entre ellos.

Se puede comprobar, por ejemplo, que la observación más distante al resto es la del alumno número 5, ya que es la última (mayor distancia) en incorporarse al cluster final, seguida de la 12 y la 14.

Por el contrario, las observaciones más cercanas entre sí son la 10 y la 13, que forman el primer grupo (distancia más próxima a 0), y la 1 y la 15, que forman el segundo.

El dendograma también nos sirve para saber la composición de cada cluster en cada paso: por ejemplo, si quisiéramos hacer una división en 5 conglomerados bastaría con trazar la línea azul y comprobaríamos que las observaciones 5, 11, 12 y 14 quedarían aisladas (formando cada una de ellas un cluster de tamaño 1), y el resto de observaciones formarían otro grupo.

Sin embargo, si deseáramos conocer la división en 8 conglomerados trazaríamos la línea roja, y obtendríamos la siguiente distribución:

CLUSTER	OBSERVACIONES
1	1, 2, 3, 7, 15
2	6, 8
3	4, 10, 13
4	9
5	11
6	14
7	12
8	5

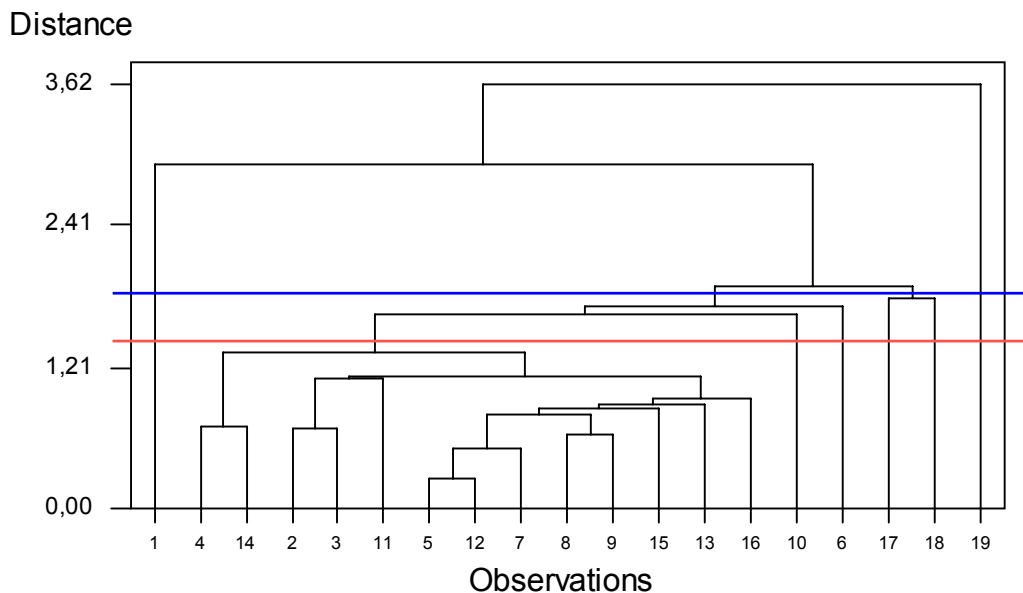
□ División en distritos de una ciudad

Procedemos análogamente con el archivo **entidades.mtw**, que recoge datos relativos a los distritos de la ciudad de Valencia (Fuente: Anuario Estadístico de Valencia 1999).

Las variables son las siguientes: NOMBRE (Nombre abreviado del distrito), SUPERFICIE (Superficie del distrito en m²), HABITANTES (Número de habitantes), TURISMOS (Número de turistas), VIVIENDAS (Número de viviendas), A E Industriales (Número de actividades económicas industriales), ENTIDADES BANCARIAS y TIPO (1: Centro, 2: Pericentro, 3: Periferia).

Obtenemos el dendograma que aparece más abajo, y nos interesa responder a las siguientes preguntas:

- ¿Cuáles son las dos observaciones más similares entre sí?
- ¿Cuáles son las dos observaciones más distintas al resto?
- Si realizamos una división en 4 grupos, ¿qué observaciones contendría cada grupo? ¿Y si la división fuera en 7 grupos?
- ¿Qué se podría decir sobre la homogeneidad de los datos?



- Las observaciones más similares entre sí son las que menor distancia presentan: en este caso, la 5 y la 12.
- La observación más distinta al resto es claramente la 19, ya que es la última que se incorpora al grupo, siendo su distancia a él la mayor; la siguiente es la 1.
- Realizando 4 conglomerados (línea azul), uno de ellos contendría a la observación 19, otro a la 1, otro a la 17 y la 18, y el resto de observaciones (2-16) formarían un grupo. Con 7 grupos (línea roja), seis de ellos serían individuales (observaciones 1, 6, 10, 17, 18, 19) y todas las demás observaciones formarían el grupo restante.
- Podemos considerar que en general los datos son bastante homogéneos, ya que la mayoría de observaciones quedan a una distancia inferior a 2 del resto; sin embargo, hay algunas que se alejan mucho de las demás, como es el caso de la 1 y la 19.

BIBLIOGRAFÍA

- [1] Baró, J. y Alemany, R. (2000): "Estadística II". Ed. Fundació per a la Universitat Oberta de Catalunya. Barcelona.
- [2] Peña Sánchez de Rivera, D. (1987): "Estadística. Modelos y Métodos. Volumen 2". Alianza Editorial. Madrid. ISBN: 84-206-8110-5
- [3] Johnson, R. R. (1996): "Elementary statistics". Belmont, etc. : Duxbury, cop
- [4] Martín-Guzmán, P. (1991): "Curso básico de estadística económica". AC, DL. Madrid. ISBN: 84-7288-142-3

ENLACES

<http://www.5campus.org/leccion/cluster>

Lección sobre Análisis Cluster (Universidad de Zaragoza)

www.ual.es/~freche/practicas/practica7/practica7.html

Práctica sobre Análisis Cluster (Universidad de Almería)

<http://home-3.tiscali.nl/~xp117079/mtad/>

Modelos y técnicas de análisis de datos (Universidad de Vigo)