Análisis cluster

- El **análisis cluster** **o** **análisis de conglomerados** engloba un conjunto de técnicas multivariantes que tiene como objetivo:
- Agrupar a un conjunto de individuos o casos en **grupos**, también denominados **conglomerados** o **cluster**, con el siguiente criterio:
- Los individuos dentro del cluster serán lo más parecidos posible (*máxima homogeneidad interna*)
- Los individuos de grupos diferentes serán lo más diferentes posible (*máxima heterogeneidad entre grupos*)
- También se pueden agrupar las **variables**.

Análisis cluster

- La agrupación se realiza teniendo en cuenta un conjunto de variables observadas sobre dichos individuos.
- **Las variables**:
- El investigador debe seleccionar, previo al análisis, aquellas variables que crea que aportan información para clasificar a los individuos en grupos. Pueden ser cuantitativas o cualitativas.

Análisis cluster

- **Los datos** se estructuran en una matriz **X** de orden Nxp, constituida por N casos (filas) y p variables (columnas).

- Algunos métodos trabajan directamente con esta información (X), y otros derivan de ella una nueva ***matriz de similaridades/distancias*** de orden NxN, que establece la similitud o distancia entre pares de casos (por ejemplo, distancia euclídea)

- **Distancia o similitud** entre los objetos:

  Medida utilizada para evaluar la distancia o similitud entre los objetos que se han de clasificar en grupos. Estas medidas vienen determinadas por el tipo de datos recopilados.

- Cuando lo que interesa es **agrupar variables**, la matriz de similitudes de orden pxp está formada por coeficientes de asociación o correlación entre variables

Análisis cluster

- **Transformación de los datos**
- Las variables cuantitativas requerirán, en muchos casos, una transformación previa al tratamiento. Si las variables presentan muy diversas unidades de medida se recomienda esta operación.
- La transformación más popular es la estandarización.
- Las variables binarias no suelen transformarse. Las variables categóricas se convierten en binarias (presencia/ausencia) para las distintas modalidades.

Análisis cluster

- **Métodos:** Una vez seleccionada la matriz de distancia a usar en el análisis, es preciso elegir un algoritmo de agrupación y establecer el número de grupos a formar.
- Uno de los procedimientos de agrupación más importante es el jerárquico. Este a su vez se clasifica en jerárquico por división o por aglomeración. Este último se caracteriza por partir de tantos grupos como objetos y en fases sucesivas ir formando conglomerados o grupos cada vez más voluminosos uniendo en cada etapa los dos grupos más similares, hasta finalizar el proceso con un solo grupo que aglutina a todos los elementos.
- Los métodos jerárquicos usan una matriz de proximidades (distancias o similaridades) como base para el agrupamiento. Comienzan con tantos cluster como objetos y terminan con un solo cluster.
  - Primero obtiene la distancia entre cluster iniciales
  - Segundo une los dos cluster mas cercanos
  - Tercero recalcula la distancia entre los clusters formados y vuelve al paso segundo

Análisis cluster

- **Algoritmos usados**
- El método se diversifica según las reglas usadas para el agrupamiento. Uno de los más usados es el método AVERAGE.

- Gráficos para conocer el modo en que se forman los grupos: el ***dendrograma***.
- El dendrograma es un gráfico usado en el procedimiento jerárquico que permite *visualizar el proceso de agrupamiento* de los cluster en los distintos pasos, formando un diagrama en árbol. Da una idea visual de la proximidad entre cluster y ayuda a decidir cuántos grupos formar.

- Los resultados del agrupamiento de un número determinado de grupos pueden guardarse como información adicional a los datos, generando una nueva variable que describe la pertenencia del caso al grupo