ClustMC: comparaciones múltiples basadas en conglomerados

Santiago García Sánchez

Palabras clave: comparaciones múltiples - análisis de conglomerados - dendrogramas

Abstract

Se presenta un paquete de R que implementa distintas pruebas de comparaciones múltiples aplicando técnicas de *clustering* para agrupar las medias de los distintos tratamientos en conglomerados no superpuestos, considerándose los tratamientos estadísticamente diferentes si se encuentran en grupos separados. Asimismo, se incluye la creación de representaciones gráficas de las pruebas en la forma de dendrogramas, brindando al usuario la capacidad de personalizarlos.

Introducción

Las técnicas de comparaciones múltiples se suelen utilizar luego de rechazar la hipótesis nula de un test F en un ANOVA. Como alternativa a los métodos clásicos, es posible aplicar un análisis de conglomerados o *clusters* para agrupar las medias de distintos tratamientos en grupos no superpuestos. Los tratamientos que pertenezcan a distintos conglomerados son considerados estadísticamente diferentes.

Uno de los primeros algoritmos de comparaciones múltiples basadas en conglomerados fue el de Scott y Knott (1974), enfoque que ha sido implementado en R mediante los paquetes *ScottKnott* (Jelihovschi et al., 2014) y *ScottKnottESD* (Tantithamthavorn, 2018). Posteriormente, diversos autores han propuesto nuevas técnicas y se han realizado estudios analizándolas y comparándolas con métodos clásicos (Di Rienzo et al., 2002). Entre sus ventajas, la aplicación de análisis de *clusters* para comparaciones múltiples permite obtener grupos de tratamientos no superpuestos, los cuales pueden ser visualizados de manera efectiva mediante la construcción de un dendrograma (Jolliffe, 1975). Además, muchas de estas técnicas requieren supuestos más débiles que aquellos necesarios para utilizar métodos tradicionales.

Funciones

El paquete ClustMC implementa diversas metodologías que se han propuesto para aplicar técnicas de análisis de conglomerados a la tarea de realizar pruebas de comparaciones múltiples. A la fecha, se incluyen las siguientes funciones:

- *bss_test()*: Prueba de Bautista, Smith y Steiner (1997).
- *dgc_test()*: Prueba de Di Rienzo, Guzmán y Casanoves (2002).
- *jolliffe_test()*: Prueba de Jolliffe (1975).

La documentación de cada una de ellas puede consultarse en el sitio web del paquete (https://sgs2000.github.io/ClustMC). Estas técnicas, especialmente la segunda, han sido

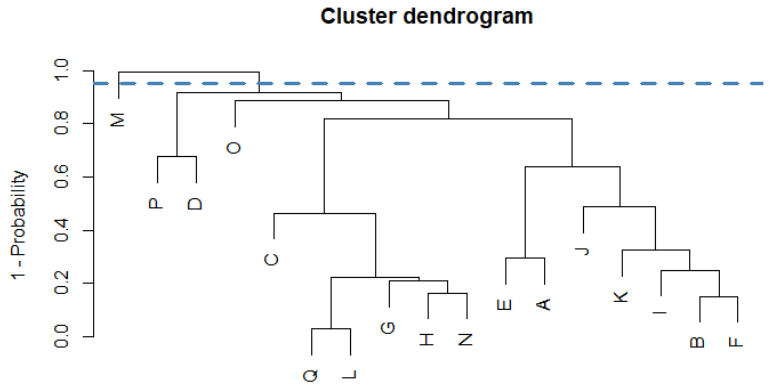
ampliamente utilizadas en publicaciones científicas y, previo al desarrollo de este paquete, solo estaban disponibles mediante determinados softwares estadísticos tales como Infostat (2008).

Similar al paquete *agricolae* (de Mendibiru, 2023), todas las funciones permiten trabajar con modelos (creados con *lm()* o *aov()* e indicando el nombre del factor a evaluar) o bien utilizar un vector con las observaciones de la variable respuesta y otro con los tratamientos. Tras aplicar la metodología correspondiente, se imprime en consola una tabla indicando los distintos tratamientos y los grupos a los que han sido asignados.

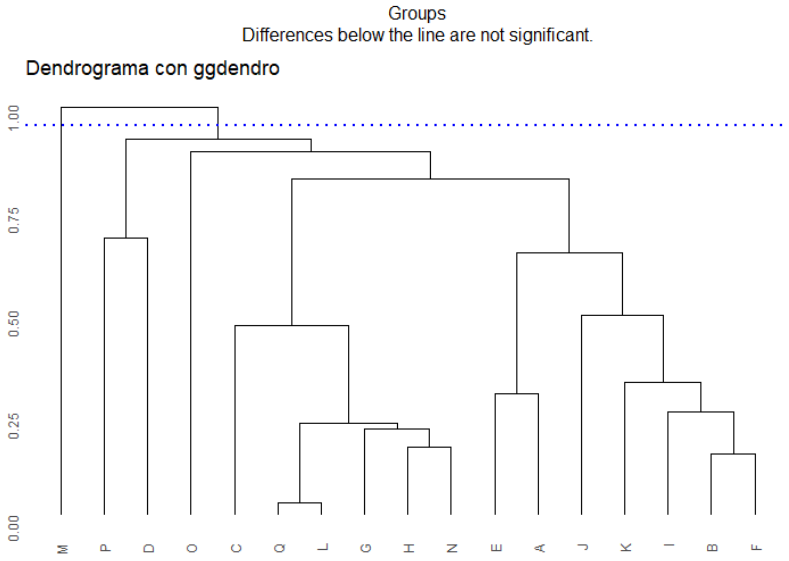
Finalmente, las funciones también devuelven por defecto un dendrograma, figura que permite examinar gráficamente los distintos conglomerados que se forman según el criterio seleccionado. Dicho gráfico puede ser personalizado por el usuario, como se explica en la siguiente sección.

Gráficos

Por defecto, se crean dendrogramas aplicando la función *plot()* al objeto de clase *hclust* que utiliza cada una de las técnicas. Estos gráficos pueden ser modificados pasando cualquier argumento disponible para *plot()* al llamar a una función del paquete.



Para darle al usuario aún más control sobre la representación de los dendrogramas, las funciones retornan el objeto *hclust* mencionado anteriormente (llamado *dendrogram_data*). Esto permite el uso de otras librerías, tales como *ggdendro* (de Vries y Ripley, 2024).



Una guía más detallada, incluyendo ejemplos y código, se presenta en el siguiente artículo: https://sgs2000.github.io/ClustMC/articles/CustomPlots.html

Referencias

- Bautista, M. G., Smith, D. W., y Steiner, R. L. (1997). A Cluster-Based Approach to Means Separation. *Journal of Agricultural, Biological, and Environmental Statistics, 2*(2), 179-197. https://doi.org/10.2307/1400402
- de Mendiburu, F. (2023). agricolae: Statistical Procedures for Agricultural Research (Version 1.3-7) [Software de computador]. The Comprehensive R Archive Network. https://myaseen208.github.io/agricolae/
- de Vries, A., Ripley, B. D. (2024). ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. (Version 0.2.0) [Software de computador]. The Comprehensive R Archive Network. https://andrie.github.io/ggdendro/.
- Di Rienzo, J. A., Guzmán, A. W., y Casanoves, F. (2002). A Multiple-Comparisons Method Based on the Distribution of the Root Node Distance of a Binary Tree. *Journal of Agricultural, Biological, and Environmental Statistics, 7*(2), 129-142. http://www.jstor.org/stable/1400690
- InfoStat (2008). *InfoStat versión 2008.* Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina.
- Jelihovschi, E. G., Faria, J. C., y Allaman, I. B. (2014). ScottKnott: A package for performing the Scott-Knott clustering algorithm in R. *Trends in Applied and Computational Mathematics, 15*(1), 3-17. https://tema.sbmac.org.br/tema/article/view/646/643
- Jolliffe, I. T. (1975). Cluster analysis as a multiple comparison method. *Applied Statistics: Proceedings of Conference at Dalhousie University, Halifax*, 159-168.
- Tantithamthavorn, C. (2018). ScottKnottESD: The Scott-Knott Effect Size Difference (ESD) Test. (Version 2.0.3) [Software de computador]. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=ScottKnottESD.