R para Microbiología Industrial: análisis de datos y diseño experimental con un enfoque práctico

Fredy Ortiz, Miguel Pérez y Francisco León

2025-05-05

Table of contents

[1. R para Microbiología Industrial: Análisis de Datos y Diseño Experimental con un Enfoque Práctico 1](#_Toc206745226)

[Prefacio (resumir) 1](#_Toc206745227)

[Autores 1](#_Toc206745228)

[Fredy Alejandro Ortiz Meneses  1](#_Toc206745229)

[Miguel Oswaldo Pérez Pulido  1](#_Toc206745230)

[Francisco Javier León  1](#_Toc206745231)

[Agradecimientos 1](#_Toc206745232)

[2. Introducción (Modificar, reducir a tres parrafos y dar otro enfoque) 1](#_Toc206745233)

[3. Summary 1](#_Toc206745234)

[4. Parte I: Preparación del Entorno y Herramientas 1](#_Toc206745235)

[4.1 3. Introducción al software R y RStudio 1](#_Toc206745236)

[4.1.1 3.1 Instalación y configuración 1](#_Toc206745237)

[4.1.2 3.2 Paquetes Esenciales para el análisis de datos 1](#_Toc206745238)

[4.2 4. Análisis Bibliométrico con RStudio® (Un ejemplo concreto de análisis de datos con R) 1](#_Toc206745239)

[5. Parte II: Aplicaciones de R en Microbiología Industrial y Análisis de Datos 1](#_Toc206745240)

[5.1 2. Fundamentos del Diseño Experimental (Introducir los principios del diseño experimental OK) 1](#_Toc206745241)

[5.1.1 2.1 Tipos de diseños experimentales (Expandir sobre los diferentes tipos de diseños OK) 1](#_Toc206745242)

[5.1.2 2.2 Ejemplos prácticos de diseños experimentales en Microbiología Industrial OK 1](#_Toc206745243)

[5.1.3 Estructura de la base de datos 1](#_Toc206745244)

[6. Parte III: Uso de Inteligencia Artificial para la simulación de datos 1](#_Toc206745245)

[6.1 Uso de Inteligencia Artificial para la simulación de datos. 1](#_Toc206745246)

[6.1.1 1](#_Toc206745247)

[References 1](#_Toc206745248)

# 1. R para Microbiología Industrial: Análisis de Datos y Diseño Experimental con un Enfoque Práctico

# Prefacio (resumir)

En el campo de la Microbiología Industrial y el diseño de experimentos, la integración de herramientas estadísticas constituye un desafío pedagógico fundamental que requiere estrategias innovadoras de enseñanza-aprendizaje, y como profesores de estas áreas de aprendizaje hemos identificado que los estudiantes experimentan dificultades significativas al establecer conexiones entre los conceptos estadísticos y los resultados experimentales microbiológicos.

En respuesta a esta problemática, surge la propuesta del libro " Aplicaciones del Software RStudio® en la Microbiología Industrial ", diseñado específicamente para articular las áreas de: Diseño de Experimentos y la Microbiología Industrial con ayuda de Rstudio®, utilizando a lo largo de contenido ejemplos concretos derivados de trabajos de grado y proyectos académicos desarrollados en la Universidad de Santander - UDES.

La obra integra además temas relacionados a: Análisis Bibliométrico, integración de Inteligencia Artificial, reconociendo de este modo que la microbiología contemporánea demanda no solo competencias técnicas, sino adaptación de nuevas habilidades en una disciplina científica en constante evolución, contribuyendo de esta forma a la formación de profesionales capaces de afrontar los desafíos emergentes del campo de microbiológico industrial, tanto para el presente como su futuro profesional.

Microbiología General y Microbiología II

**Fredy Alejandro Ortiz Meneses**  
*Curso Microbiología General y Microbiología II – Profesor de Microbiología Industrial*

**Miguel Oswaldo Pérez Pulido**  
*Curso Proyecto II – Microbiología Industrial*

**Francisco Javier León**  
*Curso Proyecto I – Profesor de Microbiología Industrial*

|  |
| --- |
| Lo que significa este libro |
| La presente obra constituye nuestra contribución a la formación integral de los Microbiólogos Industriales en su desarrollo como científicos. Esperamos que su contenido no solo fortalezca su experiencia académica, sino que además les provea de las competencias prácticas indispensables para afrontar los desafíos contemporáneos y futuros del ámbito profesional. |

# Autores

|  |
| --- |
| Tip |
| Fredy Alejandro Ortiz Meneses ﷟HYPERLINK "https://orcid.org/0000-0002-9521-5033" Microbiólogo con énfasis en Alimentos, Especialista en Pedagogía y Didácticas Específicas, y Magíster en Fitopatología. |

|  |
| --- |
| Tip |
| Miguel Oswaldo Pérez Pulido  Director de Analítica Académica. Licenciado en Matemáticas y Magíster en Estadística. Actualmente se desempeña como Director de Analítica Académica, adscrito a la Vicerrectoría de Enseñanza. Está vinculado a la Universidad de Santander (UDES) desde 2011, donde ha sido docente en programas de pregrado y posgrado de la Facultad de Ciencias Exactas, Naturales y Agropecuarias. Es investigador Junior reconocido por Minciencias en la convocatoria 894 de 2021 y miembro del grupo de investigación CIBAS |

|  |
| --- |
| Tip |
| Francisco Javier León ﷟HYPERLINK "https://orcid.org/0000-0002-8904-5242" Bacteriólogo y laboratorista clínico, con formación avanzada como Magíster en Estadística Aplicada, Magíster en Ciencias Básicas Biomédicas y Especialista en Educación con Nuevas Tecnologías. Está vinculado a la Universidad de Santander (UDES) desde 2007, donde ha sido docente en la Facultad de Ciencias Exactas, Naturales y Agropecuarias. Actualmente, se desempeña como Coordinador de Analítica Académica, adscrito a la Vicerrectoría de Enseñanza. Es investigador Junior reconocido por Minciencias en la convocatoria 894 de 2021 y miembro del grupo de investigación CIBAS |

# Agradecimientos

En primer lugar, queremos expresar nuestro más sincero agradecimiento a todos los estudiantes y profesores del programa de Microbiología que, a lo largo del tiempo, han compartido con nosotros sus inquietudes y retos al intentar conectar el análisis estadístico con la microbiología industrial.

Nuestro agradecimiento se extiende a los colegas académicos, especialmente a los profesores: Christian Andrey Chacín Zambrano y Daniel Adyro Martinez; y a los estudiantes graduados que generosamente compartieron sus experiencias y bases de datos, provenientes de importantes experimentos académicos. Sus aportes han sido fundamentales para dar vida a este manual y hacerlo relevante y aplicable a situaciones reales dentro del contexto de la microbiología industrial.

Asimismo, manifestamos gratitud a Robert Gentleman y Ross Ihaka, creadores del software R, así como a todos los colaboradores de la comunidad de R y RStudio®. Gracias a su compromiso y dedicación, estas herramientas se han mantenido accesibles para la comunidad científica.

A la Universidad de Santander (UDES) y a su Departamento de Desarrollo Profesoral, por la apertura de la Convocatoria Interna “Producción de Material Profesoral (2025), gracias a esta iniciativa, hemos encontrado un espacio de apoyo institucional que valora la producción material educativo de calidad, gracias a ello, nos sentimos motivados a seguir desarrollando herramientas que fortalezcan una enseñanza efectiva en los campos de la Microbiología Industrial y la Estadística Aplicada.

# 2. Introducción (Modificar, reducir a tres parrafos y dar otro enfoque)

El software R® representa un entorno y lenguaje de programación fundamental para el análisis estadístico y la visualización de datos en la investigación científica contemporánea, su capacidad para procesar grandes volúmenes de información y flexibilidad excepcional lo han posicionado como una herramienta indispensable a ser usada en la microbiología industrial, donde la optimización de procesos y la toma de decisiones informadas dependen crucialmente del análisis riguroso de datos experimentales (R Core Team, 2023).

Complementariamente, RStudio® constituye una Interfaz de Desarrollo Integrada (IDE) que transforma significativamente la experiencia del usuario al proporcionar un entorno accesible y amigable para la programación, integrando múltiples herramientas esenciales como: un editor de código avanzado, una consola interactiva y sistemas de visualización de datos, lo que facilita considerablemente el aprendizaje y la aplicación práctica de R en contextos académicos e industriales (RStudio Team, 2023).

En el contexto específico de la microbiología industrial, donde los requisitos analíticos varían considerablemente según el tipo de experimento y los objetivos investigativos, R y RStudio® ofrecen una flexibilidad excepcional, la comunidad activa de usuarios proporciona recursos y soporte continuo, mientras que la amplia disponibilidad de paquetes especializados permite realizar análisis complejos con mayor precisión y eficiencia (Wickham & Grolemund, 2017).

Referencias

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org

RStudio Team. (2023). RStudio: Integrated Development for R. RStudio, PBC. https://posit.co/download/rstudio-desktop/

Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. https://r4ds.had.co.nz

# 3. Summary

Posible Reorganización:

• Prefacio • Autores • Agradecimientos • Introducción (Mantener como introducción general al tema del libro)

**Parte I: Preparación del Entorno y Herramientas** • Instalación y configuración de R y RStudio® (Lógica para comenzar con la configuración del software) • Inventario de librerías básicas de R aplicadas para análisis de datos (Después de la instalación, presentar las herramientas fundamentales) • Paquetes Esenciales para el Análisis de Datos (Extender las librerías básicas con paquetes clave)

**Parte II: Aplicaciones de R en Microbiología Industrial y Análisis de Datos** • Uso de R en Microbiología Industrial (Presentar la aplicación específica de R en tu campo) • Análisis Bibliométrico con RStudio® (Un ejemplo concreto de análisis de datos con R) • Fundamentos del Diseño Experimental (Introducir los principios del diseño experimental) • Tipos de diseños experimentales (Expandir sobre los diferentes tipos de diseños) • Diseño de experimentos en RStudio® (Aplicar los conceptos de diseño experimental en R)

**Parte III: Gestión y Presentación de la Información** • Gestión de la información con informes dinámicos (Conectar el análisis de datos con la creación de informes)

References

|  |
| --- |
| Note |
| Este manual representa nuestra contribución a su formación integral como científicos. Esperamos que no solo enriquezca su experiencia académica, sino que también les proporcione las habilidades prácticas necesarias para enfrentar los desafíos actuales y futuros del campo. |

|  |
| --- |
| Using the data without downloading it |
| This same dataset is also available in the package palmerpenguins. |

|  |
| --- |
| Título del Tip |
| Texto del contenido o recomendación. |

Tipos comunes de callouts:

callout-note → Para notas informativas.  
  
callout-tip → Para consejos o buenas prácticas.  
  
callout-warning → Para advertencias.  
  
callout-important → Para contenido crítico o destacado.  
  
callout-caution → Para precauciones.

|  |
| --- |
| Cuidado |
| Este análisis puede producir resultados erróneos si no se limpian los datos previamente. |

To learn more about Quarto books visit <https://quarto.org/docs/books>.

{r} 1 + 1}

# 4. Parte I: Preparación del Entorno y Herramientas

## 4.1 3. Introducción al software R y RStudio

### 4.1.1 3.1 Instalación y configuración

La instalación de R y RStudio® es un proceso sencillo que puede completarse en unos pocos pasos; primero, se debe descargar e instalar R desde el sitio web oficial del Proyecto R (<https://www.r-project.org>); una vez instalado R, se puede proceder a descargar e instalar RStudio® desde su sitio web (<https://posit.co/download/rstudio-desktop/>) (Figura 1); ambos programas están disponibles para múltiples sistemas operativos, incluyendo Windows, macOS y Linux (Figura 2), del mismo modo se deben descarga de diferentes directorios llamadas CRAN (Comprehensive R Archive Network o Red integral de archivo R) (Figura 3); y finalmente descargar la última compilación de R para Windows (Figura 4) [@Rcore2021](https://universidaddesantnader-my.sharepoint.com/personal/analitica_academica_udes_edu_co/Documents/Material%20Docente%202025/LibroQuarto/Disen_Exp_MI/docx/R%20Core%20Team,%202023).

Una vez instalados R y RStudio®, es importante familiarizarse con la interfaz de RStudio® (Figura 5); esta interfaz está dividida en varias secciones, incluyendo: (i) el editor de código o Script, el cual permite escribir y editar las instrucciones o Scripts; (ii) la consola: se utiliza para ejecutar comandos interactivos;  (iii) el entorno de trabajo: muestra los objetos y datos cargados en la sesión actual y las (iv) pestañas de archivos y gráficos: permiten gestionar archivos y visualizar gráficos generados por R (R Core Team 2021).

Además de la interfaz básica, RStudio® permite la instalación y gestión de paquetes adicionales que amplían sus funcionalidades; para instalar un paquete, se puede utilizar la función install.packages (“nombre\_del\_paquete”) en la consola de RStudio®; una vez instalado, el paquete se puede cargar en la sesión actual utilizando la función library (nombre\_del\_paquete).

|  |
| --- |
| Important |
| Mantener R y RStudio® actualizados es clave para aprovechar las últimas mejoras, nuevas funcionalidades y correcciones de errores. Ambos programas notifican automáticamente cuando hay versiones más recientes disponibles, por lo que se recomienda estar atento a estos avisos y actualizar oportunamente. |

Para actualizar R, se debe descargar e instalar la nueva versión desde el sitio web del Proyecto R; para actualizar RStudio®, se puede utilizar la opción de actualización en el menú de ayuda de RStudio®; mantener el software actualizado garantiza un rendimiento óptimo y acceso a las últimas funcionalidades (R Core Team 2021) (R Core Team, 2023; RStudio Team, 2023).

**Referencias**

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.r-project.org>

RStudio Team. (2023). RStudio: Integrated Development for R. RStudio, PBC. <https://posit.co/download/rstudio-desktop/>

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer. <https://doi.org/10.1007/978-3-319-24277-4>

Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O’Reilly Media. [https://r4ds.had.co.nz](https://r4ds.had.co.nz/)

### 4.1.2 3.2 Paquetes Esenciales para el análisis de datos

#### 4.1.2.1 Inventario de Librerías y Paquetes de R aplicados para el análisis de datos en Microbiología Industrial.

El ecosistema de librerías y paquetes de R constituye una herramienta fundamental para el análisis de datos en microbiología industrial, proporcionando soluciones específicas para cada etapa del proceso investigativo, y en este contexto, las librerías básicas como: (i) *readxl*, desarrollada por Wickham y Bryan (2015), facilitan la importación de datos desde hojas de cálculo Excel®, donde tradicionalmente los microbiólogos registran sus resultados experimentales; (ii) complementariamente *car* (Companion to Applied Regression), creada por Fox y Weisberg (2019), ofrece herramientas esenciales para la verificación de supuestos estadísticos mediante gráficos QQ-plot, permitiendo evaluar la normalidad de los datos antes de aplicar pruebas paramétricas en experimentos de optimización de medios de cultivo y comparación de cepas microbianas.

La revolución en el análisis de datos microbiológicos se materializa principalmente a través del paquete *tidyverse*, desarrollado por Wickham et al. (2019), que integra múltiples librerías bajo una lógica común de programación, Este conjunto incluye: (i) *dplyr* para manipulación de datos, (ii) *ggplot2* para visualización avanzada, (iii) *tidyr* para ordenamiento de información y readr para importación eficiente, facilitando significativamente el flujo de trabajo en análisis complejos típicos de estudios de cinética enzimática y crecimiento microbiano. Así mimso, la librería *apaTables*, creada por Stanley (2021), permite la generación automática de tablas en el estilo APA directamente desde los análisis estadísticos realizados, con el fin de ser presentados en publicaciones científicas.

El arsenal analítico se enriquece considerablemente con librerías especializadas que abordan necesidades específicas de la investigación microbiológica industrial, y tal es el cso de: La librería *gridExtra*, desarrollada por Auguie (2017), la cual facilita la organización de múltiples gráficos en una sola visualización, permitiendo comparaciones efectivas entre diferentes condiciones experimentales. Por otra parte *lsr* (Learning Statistics with R), creada por Navarro (2015) proporciona funciones accesibles para análisis estadísticos fundamentales como pruebas t, ANOVA y cálculos de tamaño del efecto; adicionalmente, bibliometrix, desarrollado por Aria y Cuccurullo (2017) permite realizar análisis bibliométrico de publicaciones científicas, identificando tendencias emergentes y redes de colaboración que orientan nuevas investigaciones.

Finalmente, las aplicaciones especializadas en análisis multivariado y modelado avanzado complementan este inventario tecnológico, y es donde converge la librería *vegan*, desarrollada por Oksanen et al. (2020), la cual proporciona proporciona herramientas para análisis de diversidad ecológica mediante técnicas como PCA (Análisis de Componentes Principales), NMDS (Escalamiento Multidimensional No Métric) permitiendo visualizar relaciones complejas entre comunidades microbianas y variables ambientales en procesos industriales. Paralelamente, *nlme* desarrollado por Pinheiro et al., (2025) ofrece capacidades para modelar datos longitudinales con estructura jerárquica, típicos de estudios de cinética microbiana. Asimismo, paquetes como *agricolae* (Mendiburu, 2020) facilitan el diseño experimental, mientras que *shiny* (Chang et al., 2021) permite desarrollar aplicaciones web interactivas para visualización dinámica de resultados, mejorando la colaboración y transparencia en la investigación microbiológica industrial.

Referencias

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression (3rd ed.). Sage Publications. https://uk.sagepub.com/en-gb/eur/an-r-companion-to-applied-regression/book246125

Mendiburu, F. (2020). agricolae: Statistical Procedures for Agricultural Research. R package version 1.3-3. https://CRAN.R-project.org/package=agricolae

Navarro, D. J. (2015). Learning Statistics with R: A tutorial for psychology students and other beginners. (Version 0.5). University of Adelaide. https://learningstatisticswithr.com/

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). vegan: Community ecology package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2025). nlme: Linear and nonlinear mixed effects models. R package version 3.1-157. https://CRAN.R-project.org/package=nlme

Stanley, D. (2021). apaTables: Create American Psychological Association (APA) Style Tables. R package version 2.0.8. https://CRAN.R-project.org/package=apaTables

Wickham, H., & Bryan, J. (2015). readxl: Read Excel Files. R package version 1.4.2. https://CRAN.R-project.org/package=readxl

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

#### 4.1.2.3 Uso de R en Microbiología Industrial

El uso de R como herramienta de análisis estadístico en la microbiología industrial ha experimentado un crecimiento exponencial en la última década, según Mohammadi et al. (2019) R proporciona una plataforma versátil que permite analizar datos complejos derivados de experimentos microbiológicos, facilitando la identificación de patrones de crecimiento microbiano, optimización de condiciones de cultivo y evaluación de la producción de metabolitos secundarios, lo que resulta crucial para el desarrollo y mejora de procesos biotecnológicos en entornos industriales.

McMurdie y Holmes (2013) desarrollaron el paquete *phyloseq*, el cual ha transformado el análisis de datos de secuenciación en estudios de comunidades microbianas, permitiendo la integración de información taxonómica, filogenética y de abundancia en un solo entorno analítico; este avance ha sido fundamental para comprender la dinámica de poblaciones microbianas en procesos industriales como: el tratamiento de aguas residuales, la producción de biocombustibles y la fermentación alimentaria.

Por otra parte, el paquete microbiome, descrito por Lahti y Shetty (2017) (Reeder et al. 2021), proporciona herramientas especializadas para el análisis de datos metagenómicos, facilitando la caracterización de comunidades microbianas y sus funciones metabólicas en entornos industriales, lo que resulta esencial para la optimización de bioprocesos y el control de calidad en la industria alimentaria.

El diseño experimental en microbiología industrial se ha beneficiado significativamente de la aplicabilidad de R, permitiendo planificar y analizar experimentos de manera más rigurosa y eficiente, el paquete *agricolae*, desarrollado por de Mendiburu (2021) (Zhou et al. 2012) es utilizado para la implementación de diseños experimentales complejos como: bloques aleatorizados y diseños factoriales entre otros, al tiempo que frecuentemente son utilizados en estudios de optimización de medios de cultivo, condiciones de fermentación y producción de enzimas microbianas.

Complementariamente, Ritz y Streibig (2005) presentaron el paquete *drc* (Dose-Response Curves), que ha facilitado el análisis de Curvas dosis-respuesta en estudios de inhibición microbiana, pruebas de susceptibilidad a antimicrobianos y evaluación de compuestos bioactivos producidos por microorganismos, proporcionando herramientas estadísticas robustas para cuantificar y modelar respuestas biológicas a diferentes tratamientos, lo cual es fundamental en el desarrollo de nuevos productos biotecnológicos.

Gracias al paquete *ggplot2* desarrollado por Wickham (2016), el cual ha permitido la creación de gráficos elegantes y altamente informativos que facilitan la interpretación de resultados experimentales; en particular, la representación gráfica de cinéticas de crecimiento microbiano, producción de metabolitos y análisis multivariantes se ha vuelto más accesible e intuitiva para investigadores en el campo; de manera similar el paquete *ggtree*, creado por Yu et al. (2017), ha revolucionado la visualización de datos filogenéticos en estudios de diversidad microbiana industrial, permitiendo representar relaciones evolutivas entre microorganismos de interés biotecnológico y correlacionarlas con características fenotípicas relevantes para procesos industriales, lo que facilita la selección de cepas microbianas con potencial biotecnológico.

|  |
| --- |
| Expandir para aprender son el analis de datos ómicos |
| El análisis de datos ómicos en microbiología industrial se ha visto significativamente potenciado gracias al aporte de Love et al. (2014) quienes introdujeron *DESeq2*, un paquete que ha transformado el análisis de datos de RNA-seq en estudios transcriptómicos de microorganismos industriales, permitiendo identificar genes diferencialmente expresados bajo diversas condiciones de cultivo o modificaciones genéticas; lo que contribuye a la mejora de cepas microbianas industriales y a optimizar rutas metabólicas de interés comercial; paralelamente Rohart et al. (2017) desarrollaron el paquete mixOmics, el cual facilita la integración de múltiples conjuntos de datos ómicos, como:   1. transcriptómica, 2. proteómica y 3. metabolómica,   proporcionando una visión holística de los sistemas microbianos en contextos industriales, lo que permite desentrañar complejas redes regulatorias y metabólicas que subyacen a procesos biotecnológicos importantes como de compuestos bioactivos.. |

|  |
| --- |
| Expand to learn more about where the data was collected |
| Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. |

**Referencias**

Zhou et al. (2012)

González-Peña, D., Dudzik, D., García, A., Legido-Quigley, C., & Barbas, C. (2021). R in LC-MS-based metabolomics: A review of packages, tutorials, and case studies. Metabolites, 11(10), 672. <https://doi.org/10.3390/metabo11100672> ***Esta referencia esta mal***

***Zhou, B., Xiao, J. F., Tuli, L., & Ressom, H. W. (2012). LC-MS-based metabolomics. Molecular BioSystems, 8(2), 470–481. https://doi.org/10.1039/c1mb05350g*** ( es esta ???)

Lahti, L., & Shetty, S. (2017). microbiome R package. Tools Microbiome Anal. R. <http://microbiome.github.io/microbiome> ***Cambiarla mejor por la que sigue???***

***Reeder, J., Huang, M., Kaminker, J. S., & Paulson, J. N. (2021). MicrobiomeExplorer: an R package for the analysis and visualization of microbial communities. Bioinformatics (Oxford, England), 37(9), 1317–1318. https://doi.org/10.1093/bioinformatics/btaa838*** ( es esta ???)

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLOS ONE, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>

Mendiburu, F. (2021). agricolae: Statistical procedures for agricultural research. R package version 1.4-0. <https://CRAN.R-project.org/package=agricolae>

Mohammadi, R., Ghomi, S. M. T. F., & Nazari, F. (2019). The application of R software for the assessment of microbial fermentation processes. Journal of Microbiological Methods, 156, 54-58. <https://doi.org/10.1016/j.mimet.2018.12.003>

Ritz, C., & Streibig, J. C. (2005). Bioassay analysis using R. Journal of Statistical Software, 12(5), 1-22. <https://doi.org/10.18637/jss.v012.i05>

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for ’omics feature selection and multiple data integration. PLOS Computational Biology, 13(11), e1005752. <https://doi.org/10.1371/journal.pcbi.1005752>

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-24277-4>

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8(1), 28-36. <https://doi.org/10.1111/2041-210X.12628>

## 4.2 4. Análisis Bibliométrico con RStudio® (Un ejemplo concreto de análisis de datos con R)

# 5. Parte II: Aplicaciones de R en Microbiología Industrial y Análisis de Datos

## 5.1 2. Fundamentos del Diseño Experimental (Introducir los principios del diseño experimental OK)

### 5.1.1 2.1 Tipos de diseños experimentales (Expandir sobre los diferentes tipos de diseños OK)

### 5.1.2 2.2 Ejemplos prácticos de diseños experimentales en Microbiología Industrial OK

#### 5.1.2.1 2.2.1 Diseño completamente al azar OK

|  |
| --- |
| Problema |
| Escibrir los datos del problema y la bibliografci de donde se saco la información. |

### 5.1.3 Estructura de la base de datos

La base de datos utilizada en este análisis corresponde a los resultados de un experimento agrícola que evalúa el comportamiento de cuatro cultivos diferentes bajo condiciones similares de manejo. La tabla contiene tres columnas principales:

| Variable | Descripción |
| --- | --- |
| Tratamiento | Tipo de cultivo evaluado. Incluye cuatro niveles: Arroz, Avena, Cebada y Maíz. |
| Repetición | Número de repetición del tratamiento (del 1 al 4). Permite el análisis estadístico con replicación. |
| Resultado | Valor numérico correspondiente a la variable respuesta medida (por ejemplo, rendimiento en kg/ha). |

**Pasos para trabajar con R o RStudio:**

Especificar el directorio que me interesa donde se encuentra la base de datos.

|  |
| --- |
| Antes e inciar |
| R lee / (slash o division) y no el de Windows \  En **R**, setwd() es una función que significa **“set working directory”** o “establecer el directorio de trabajo”. Se utiliza para **definir la carpeta predeterminada** en la que R buscará archivos para leer y donde guardará archivos por defecto.  Por ejemplo: setwd (“D:/OneDrive - Universidad de Santander/Material Docente 2025/CodigoR”“) |

**Lectura de datos**

library(readxl)

Warning: package 'readxl' was built under R version 4.3.3

DCA <- read\_excel("C:/Users/coordinador.analitic/OneDrive - Universidad de Santander/Material\_Profesoral\_2025/CodigoR/DCA2.xlsx")  
  
View(DCA)  
attach(DCA)  
names(DCA)

[1] "Tratamiento" "Repeticion" "Resultado"

str(DCA)

tibble [16 × 3] (S3: tbl\_df/tbl/data.frame)  
 $ Tratamiento: chr [1:16] "Arroz" "Arroz" "Arroz" "Arroz" ...  
 $ Repeticion : num [1:16] 1 2 3 4 1 2 3 4 1 2 ...  
 $ Resultado : num [1:16] 8.76 8.74 8.72 8.72 8.39 ...

summary(DCA$Resultado)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 8.341 8.635 8.792 8.775 8.954 9.141

**Análisis de la Varianza - ANOVA**

Cuando se desea saber si varios grupos (Ej. tratamientos) presentan diferencias reales en sus promedios, una de las herramientas estadísticas más utilizadas es el Análisis de la Varianza, conocido como ANOVA. Esta técnica permite examinar si los valores medios de tres o más grupos son lo suficientemente distintos como para concluir que no se trata de simples fluctuaciones aleatorias.

El enfoque de ANOVA se basa en comparar dos tipos de variación: por un lado, l**a variabilidad que se observa entre los distintos grupos**, y por otro, **la variabilidad que existe dentro de cada grupo individual**.

Si al analizar los datos se encuentra que la variación entre los grupos supera notablemente la que ocurre dentro de ellos, es razonable pensar que las diferencias en los promedios reflejan algo más que el azar. En cambio, si la variabilidad interna es más pronunciada, entonces es posible que las diferencias observadas no sean significativas y respondan a variaciones normales del comportamiento de los datos.

**Código de R para ANOVA**

Anova<-aov(Resultado~Tratamiento, data=DCA)  
summary(Anova)

Df Sum Sq Mean Sq F value Pr(>F)   
Tratamiento 3 1.1794 0.3931 660.4 1.39e-13 \*\*\*  
Residuals 12 0.0071 0.0006   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Interpretación:** La prueba ANOVA muestra diferencias significativas entre los tratamientos (p < 0.001). El valor de F (660.4) indica que la variación entre tratamientos es mucho mayor que la variación dentro de los grupos, lo que sugiere que al menos uno de los tratamientos afecta significativamente el resultado.

**Modelo Lineal**

modelo=lm(Resultado~(Tratamiento))  
summary(modelo)

Call:  
lm(formula = Resultado ~ (Tratamiento))  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.038397 -0.016205 0.001983 0.012013 0.040116   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 8.73389 0.01220 715.921 < 2e-16 \*\*\*  
TratamientoAvena -0.35848 0.01725 -20.778 8.93e-11 \*\*\*  
TratamientoCebada 0.12669 0.01725 7.343 8.94e-06 \*\*\*  
TratamientoMaiz 0.39630 0.01725 22.970 2.75e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.0244 on 12 degrees of freedom  
Multiple R-squared: 0.994, Adjusted R-squared: 0.9925   
F-statistic: 660.4 on 3 and 12 DF, p-value: 1.393e-13

**Interpretación:** El modelo lineal confirma que el tratamiento influye significativamente en los resultados (p < 0.001). El tratamiento “Arroz” actúa como referencia, con una media estimada de 8.73. Comparado con este:

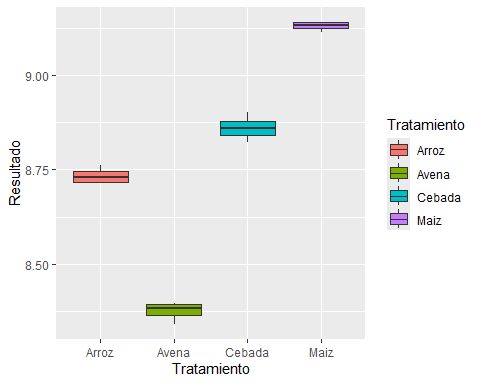
Avena presenta una media significativamente menor (–0.36, p < 0.001).  
  
Cebada muestra un aumento moderado (+0.13, p < 0.001).  
  
Maíz tiene el mayor incremento (+0.40, p < 0.001).

El modelo explica el 99.4% de la variabilidad en los datos (R² = 0.994), y el error estándar residual es bajo (0.0244), lo que indica un ajuste excelente.

**Gráfico Boxplot**

Se toma el Tratamiento para hacer un boxplot utilizando la variable “Resultado”, pero primero se transformar en factor la variable Tratamiento:

library(ggplot2)  
  
DCA$Treatamiento<-factor(DCA$Tratamiento) #transformamos una variable númerica en un factor categórico  
ggplot(DCA, aes(x = Tratamiento, y = Resultado, fill=Tratamiento)) +   
 geom\_boxplot()



**Interpretación:** Las diferencias en las medianas entre tratamientos son claras y consistentes con los resultados del ANOVA y del modelo lineal, lo que sugiere un efecto significativo del tipo de cultivo sobre la variable resultado.

**Supuestos del diseño**

**Normalidad:** Para verificar la normalidad de los residuos utilizaremos la prueba de Shapiro-Wilks cuyo script es el siguiente:

shapiro.test(residuals(Anova))

Shapiro-Wilk normality test  
  
data: residuals(Anova)  
W = 0.97944, p-value = 0.959

**Interpretación:** El test de Shapiro-Wilk aplicado a los residuos del modelo ANOVA devuelve un valor de p = 0.959, que es mucho mayor que 0.05. Esto indica que no hay evidencia estadística para rechazar la hipótesis nula de normalidad. Por lo tanto, se concluye que los residuos del modelo siguen una distribución normal, cumpliendo así uno de los supuestos fundamentales del análisis de varianza.

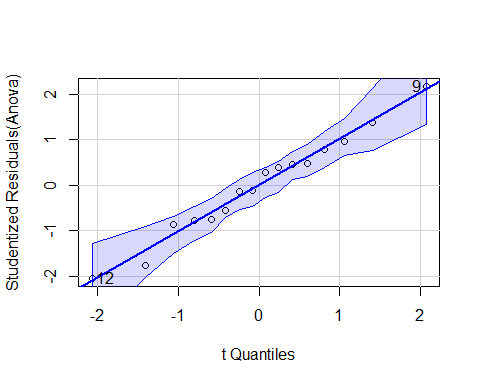
**Gráficos para evaluar la normalidad**

Para construir el gráfico QQ (QQ plot) y evaluar la normalidad de los datos, se utiliza la función correspondiente del paquete car. Si no está instalado previamente, es necesario instalar también el paquete auxiliar carData.

Instalación (si es necesario) install.packages (“car”) install.packages (“carData”) install.packages (“dplyr”) install.packages (“purrr”)

**Cargar los paquetes (librerias)**

library(car) #Grafico de QQ plot  
library(carData)  
library(dplyr)  
library(purrr)  
  
qqPlot(Anova)



[1] 9 12

**Interpretación:** El gráfico QQ muestra que los residuos estandarizados del modelo ANOVA se alinean adecuadamente con la línea diagonal, lo que indica que su distribución es aproximadamente normal. La mayoría de los puntos se ubican dentro de la banda de confianza, y no se observan desviaciones sistemáticas. Esta gráfica complementa el resultado del test de Shapiro-Wilk (p = 0.959), confirmando que se cumple el supuesto de normalidad de los residuos en el modelo.

**Homocedasticidad:** Para evaluar el supuesto de homogeneidad de varianzas entre los grupos (homocedasticidad), se aplicará la prueba de Bartlett, la cual es apropiada cuando los datos provienen de poblaciones aproximadamente normales. Esta prueba contrasta la hipótesis nula de igualdad de varianzas frente a la alternativa de varianzas diferentes. El procedimiento se implementa mediante el siguiente script:

bartlett.test(Resultado~Tratamiento, data=DCA)

Bartlett test of homogeneity of variances  
  
data: Resultado by Tratamiento  
Bartlett's K-squared = 2.2722, df = 3, p-value = 0.5179

**Interpretación:** Dado que el valor de p es mayor que 0.05 (p = 0.5179), no se rechaza la hipótesis nula. Por tanto, se asume que las varianzas entre los tratamientos son homogéneas, cumpliéndose este supuesto clave para el análisis de varianza y para la aplicación de pruebas a posteriori como LSD.

**Pruebas aposteriori** Para identificar diferencias específicas entre las medias de los tratamientos, una vez detectada significancia en el análisis de varianza, se aplicará una prueba de comparaciones múltiples a posteriori. En este caso, se empleará la técnica LSD (Least Significant Difference), que permite realizar comparaciones pareadas entre tratamientos asumiendo homogeneidad de varianzas.

La implementación de esta prueba requiere la carga del paquete agricolae, utilizando el siguiente script. Instalación si es necesario: install.packages(“agricolae”). Carga del paquete: library(agricolae).

library(agricolae)  
Grupos <- LSD.test(y = Anova, trt = "Tratamiento", group = T, console = T)

Study: Anova ~ "Tratamiento"  
  
LSD t Test for Resultado   
  
Mean Square Error: 0.0005953124   
  
Tratamiento, means and individual ( 95 %) CI  
  
 Resultado std r se LCL UCL Min Max  
Arroz 8.733890 0.02192214 4 0.01219951 8.707310 8.760471 8.715318 8.762183  
Avena 8.375414 0.02519485 4 0.01219951 8.348834 8.401995 8.341039 8.395990  
Cebada 8.860578 0.03330518 4 0.01219951 8.833998 8.887159 8.822181 8.900695  
Maiz 9.130190 0.01251613 4 0.01219951 9.103609 9.156770 9.113429 9.140539  
 Q25 Q50 Q75  
Arroz 8.717232 8.729030 8.745688  
Avena 8.364419 8.382314 8.393309  
Cebada 8.842075 8.859719 8.878222  
Maiz 9.124249 9.133395 9.139335  
  
Alpha: 0.05 ; DF Error: 12  
Critical Value of t: 2.178813   
  
least Significant Difference: 0.03759044   
  
Treatments with the same letter are not significantly different.  
  
 Resultado groups  
Maiz 9.130190 a  
Cebada 8.860578 b  
Arroz 8.733890 c  
Avena 8.375414 d

**Intrepretación:** La prueba LSD reveló que los cuatro tratamientos presentan diferencias estadísticamente significativas entre sus medias. El tratamiento Maíz obtuvo el mayor rendimiento promedio, seguido por Cebada, Arroz y Avena, en ese orden descendente.

Otra opcion cuando cambiamos el argumento “group” a F(false), se interpreta a mi parecer de forma mas sencilla la diferencia entre las medias.A continuación, se presentan las pruebas de comparaciones múltiples a posteriori aplicadas al modelo de ANOVA ajustado. Se incluyen la prueba LSD, la prueba de Tukey y el test de Scheffé, las cuales permiten identificar diferencias estadísticamente significativas entre los tratamientos evaluados:

Grupos<- LSD.test(y = Anova, trt = "Tratamiento", group = F, console = T)

Study: Anova ~ "Tratamiento"  
  
LSD t Test for Resultado   
  
Mean Square Error: 0.0005953124   
  
Tratamiento, means and individual ( 95 %) CI  
  
 Resultado std r se LCL UCL Min Max  
Arroz 8.733890 0.02192214 4 0.01219951 8.707310 8.760471 8.715318 8.762183  
Avena 8.375414 0.02519485 4 0.01219951 8.348834 8.401995 8.341039 8.395990  
Cebada 8.860578 0.03330518 4 0.01219951 8.833998 8.887159 8.822181 8.900695  
Maiz 9.130190 0.01251613 4 0.01219951 9.103609 9.156770 9.113429 9.140539  
 Q25 Q50 Q75  
Arroz 8.717232 8.729030 8.745688  
Avena 8.364419 8.382314 8.393309  
Cebada 8.842075 8.859719 8.878222  
Maiz 9.124249 9.133395 9.139335  
  
Alpha: 0.05 ; DF Error: 12  
Critical Value of t: 2.178813   
  
Comparison between treatments means  
  
 difference pvalue signif. LCL UCL  
Arroz - Avena 0.3584760 0 \*\*\* 0.3208855 0.39606642  
Arroz - Cebada -0.1266884 0 \*\*\* -0.1642788 -0.08909794  
Arroz - Maiz -0.3962994 0 \*\*\* -0.4338899 -0.35870901  
Avena - Cebada -0.4851644 0 \*\*\* -0.5227548 -0.44757392  
Avena - Maiz -0.7547754 0 \*\*\* -0.7923659 -0.71718499  
Cebada - Maiz -0.2696111 0 \*\*\* -0.3072015 -0.23202064

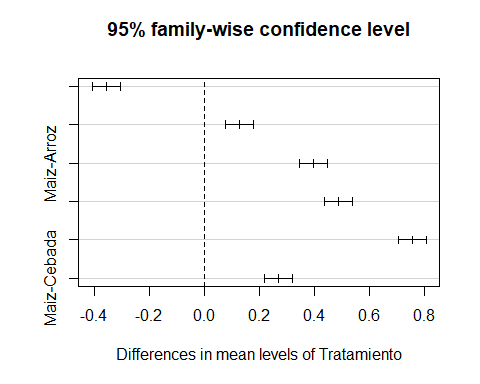
**Interpretación:** todas las diferencias entre tratamientos son altamente significativas (p < 0.001). Esto confirma que ninguno de los tratamientos comparte una media similar.

TukeyHSD(Anova)

Tukey multiple comparisons of means  
 95% family-wise confidence level  
  
Fit: aov(formula = Resultado ~ Tratamiento, data = DCA)  
  
$Tratamiento  
 diff lwr upr p adj  
Avena-Arroz -0.3584760 -0.40969759 -0.3072544 0.0e+00  
Cebada-Arroz 0.1266884 0.07546677 0.1779100 4.6e-05  
Maiz-Arroz 0.3962994 0.34507784 0.4475211 0.0e+00  
Cebada-Avena 0.4851644 0.43394275 0.5363860 0.0e+00  
Maiz-Avena 0.7547754 0.70355383 0.8059970 0.0e+00  
Maiz-Cebada 0.2696111 0.21838947 0.3208327 0.0e+00

**Interpretación:** La prueba de Tukey también confirma diferencias estadísticamente significativas en todas las comparaciones, manteniendo control del error familiar. El gráfico generado muestra intervalos de confianza del 95% que no se solapan, lo que respalda visualmente los resultados.

plot(TukeyHSD(Anova))



**Interpretación:** El gráfico muestra los intervalos de confianza del 95 % para las diferencias de medias entre los tratamientos, ajustados por comparaciones múltiples (family-wise). Ninguno de los intervalos cruza la línea vertical en cero, lo cual indica que todas las comparaciones entre pares de tratamientos son estadísticamente significativas. La diferencia más grande se observa entre Maíz y Avena, mientras que la más pequeña, aunque significativa, es entre Cebada y Arroz. Este resultado es coherente con los análisis previos (ANOVA, LSD y Scheffé), y respalda que cada tratamiento tiene un efecto significativamente distinto sobre la variable “Resultado”.

scheffe.test(Anova, "Tratamiento",console=TRUE)

Study: Anova ~ "Tratamiento"  
  
Scheffe Test for Resultado   
  
Mean Square Error : 0.0005953124   
  
Tratamiento, means  
  
 Resultado std r se Min Max Q25 Q50  
Arroz 8.733890 0.02192214 4 0.01219951 8.715318 8.762183 8.717232 8.729030  
Avena 8.375414 0.02519485 4 0.01219951 8.341039 8.395990 8.364419 8.382314  
Cebada 8.860578 0.03330518 4 0.01219951 8.822181 8.900695 8.842075 8.859719  
Maiz 9.130190 0.01251613 4 0.01219951 9.113429 9.140539 9.124249 9.133395  
 Q75  
Arroz 8.745688  
Avena 8.393309  
Cebada 8.878222  
Maiz 9.139335  
  
Alpha: 0.05 ; DF Error: 12   
Critical Value of F: 3.490295   
  
Minimum Significant Difference: 0.05582762   
  
Means with the same letter are not significantly different.  
  
 Resultado groups  
Maiz 9.130190 a  
Cebada 8.860578 b  
Arroz 8.733890 c  
Avena 8.375414 d

**Interpretación:** A pesar de ser una prueba más conservadora, el test de Scheffé también encontró diferencias significativas entre todos los tratamientos. El análisis agrupó los tratamientos en distintos niveles.Mínima diferencia significativa (Scheffé): 0.0558. Valor crítico de F: 3.4903

**Conclusión general** Las tres pruebas aplicadas (LSD, Tukey y Scheffé) coinciden en que todos los tratamientos difieren significativamente entre sí. El tratamiento con mayor rendimiento fue Maíz, seguido por Cebada, Arroz y Avena, en orden descendente. Esto respalda la conclusión de que el tipo de tratamiento influye de manera significativa sobre la variable respuesta.

#### 5.1.3.1 2.2.2 Diseño de bloques completamente al azar OK

#### 5.1.3.2 2.2.3 Diseño longitudinal (ANOVA de medidas repetidas) OK

# 6. Parte III: Uso de Inteligencia Artificial para la simulación de datos

## 6.1 Uso de Inteligencia Artificial para la simulación de datos.

### 6.1.1

<https://www.researchgate.net/publication/391573622_Caso_de_estudio_sobre_simulacion_de_datos_para_investigaciones_academicas_mediante_Inteligencia_Artificial_Generativa_y_Google_Colab>

# References

Grolemund, Garrett, and Hadley Wickham. 2017. *R for Data Science*. O’Reilly Media.

Peng, Roger D. 2011. “Reproducible Research in Computational Science.” *Science* 334 (6060): 1226–27. <https://doi.org/10.1126/science.1213847>.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Reeder, Janina, Mo Huang, Joshua S Kaminker, and Joseph N Paulson. 2021. “MicrobiomeExplorer: an R package for the analysis and visualization of microbial communities.” *Bioinformatics (Oxford, England)* 37 (9): 1317–18. <https://doi.org/10.1093/bioinformatics/btaa838>.

Wickham, Hadley. 2016. *Ggplot2*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-24277-4>.

Wickham, Hadley, and Jennifer Bryan. 2015. “Readxl: Read Excel Files.” The R Foundation. <https://doi.org/10.32614/cran.package.readxl>.

Zhou, Bin, Jun Feng Xiao, Leepika Tuli, and Habtom W Ressom. 2012. “LC-MS-based metabolomics.” *Molecular bioSystems* 8 (2): 470–81. <https://doi.org/10.1039/c1mb05350g>.