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

Fredy Ortiz Miguel Pérez Francisco León

2025-05-05

Tabla de contenidos

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

Prefacio

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.

Fredy Alejandro Ortiz Meneses
Curso Microbiologia General y Microbiología II

Miguel Oswaldo Pérez Pulido
Curso Proyecto II – Microbiología Industrial
Maestría en Estadística Aplicada y Analítica de Datos

Francisco Javier León
Curso Proyecto I – Profesor de Microbiología Industrial
Maestría en Estadística Aplicada y Analítica de Datos

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

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

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.

2 Agradecimientos

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.

3 Introducción

En el ámbito de la microbiología industrial, donde los requerimientos analíticos varían según el tipo de experimento y los objetivos investigativos, R y RStudio® ofrecen una flexibilidad sobresaliente. La comunidad global de usuarios provee soporte constante y recursos actualizados, mientras que la amplia disponibilidad de paquetes especializados permite realizar análisis complejos con mayor precisión y eficiencia (Wickham & Grolemund, 2017). Esta combinación de potencia analítica, reproducibilidad y accesibilidad convierte a R en una herramienta idónea para el análisis de datos experimentales, el diseño de experimentos y la optimización de procesos biotecnológicos.

El presente libro está estructurado en tres partes principales.

- La Parte I aborda la instalación, configuración y manejo del entorno RStudio®, junto con un inventario de librerías esenciales para el análisis de datos en microbiología industrial.
- La Parte II desarrolla aplicaciones prácticas de R en el diseño experimental, con ejemplos reproducibles de diseños completamente al azar, en bloques y con mediciones repetidas en el tiempo.
- La Parte III introduce el uso de inteligencia artificial y simulación de datos, explorando cómo los modelos generativos y las herramientas computacionales pueden complementar la investigación microbiológica moderna.

De este modo, el libro ofrece una guía integral que combina fundamentos teóricos, práctica aplicada y perspectivas innovadoras, contribuyendo a fortalecer las competencias analíticas de los estudiantes y profesionales de la microbiología industrial.

4 Parte I: Preparación del Entorno y Herramientas

4.1 Introducción al software R y RStudio

4.1.1 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).

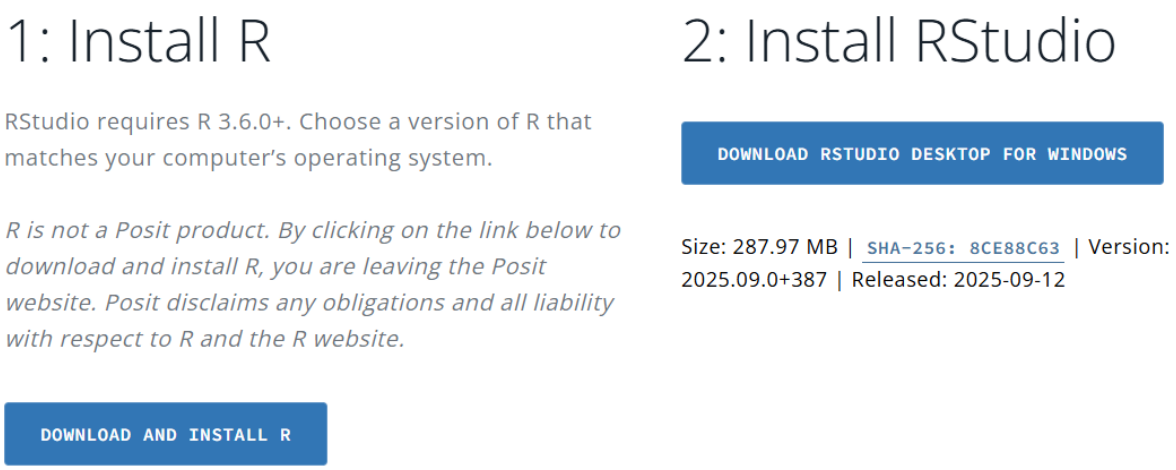


Figura 4.1: Figura 1.

Ambos programas están disponibles para múltiples sistemas operativos, incluyendo Windows, macOS y Linux (Figura 2).

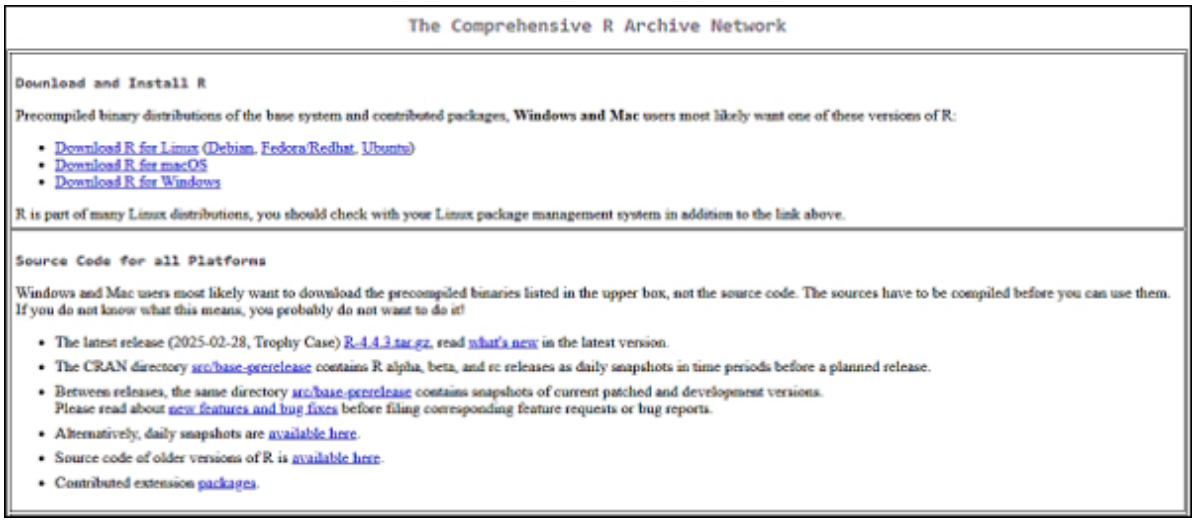


Figura 4.2: Figura 2.

Del mismo modo se deben descarga de diferentes directorios llamadas CRAN (Comprehensive R Archive Network o Red integral de archivo R) (Figura 3).

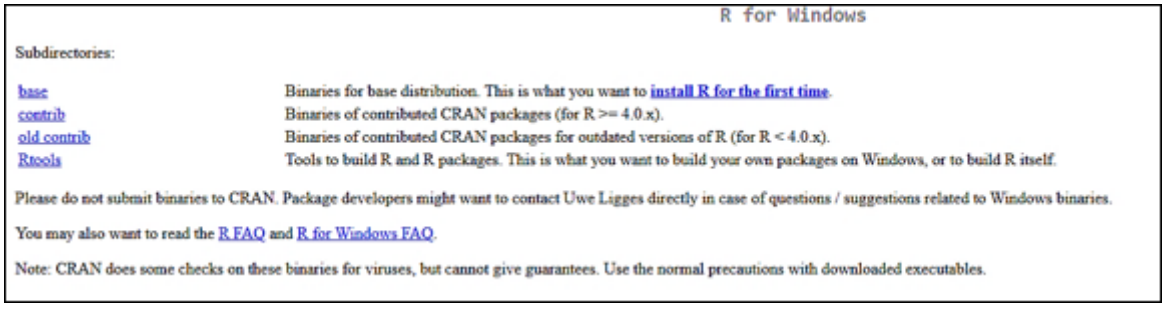


Figura 4.3: Figura 3.

Finalmente descargar la última versión de R para Windows (Figura 4) [@Rcore2021].

Figura 4.4: Figura 4.

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).

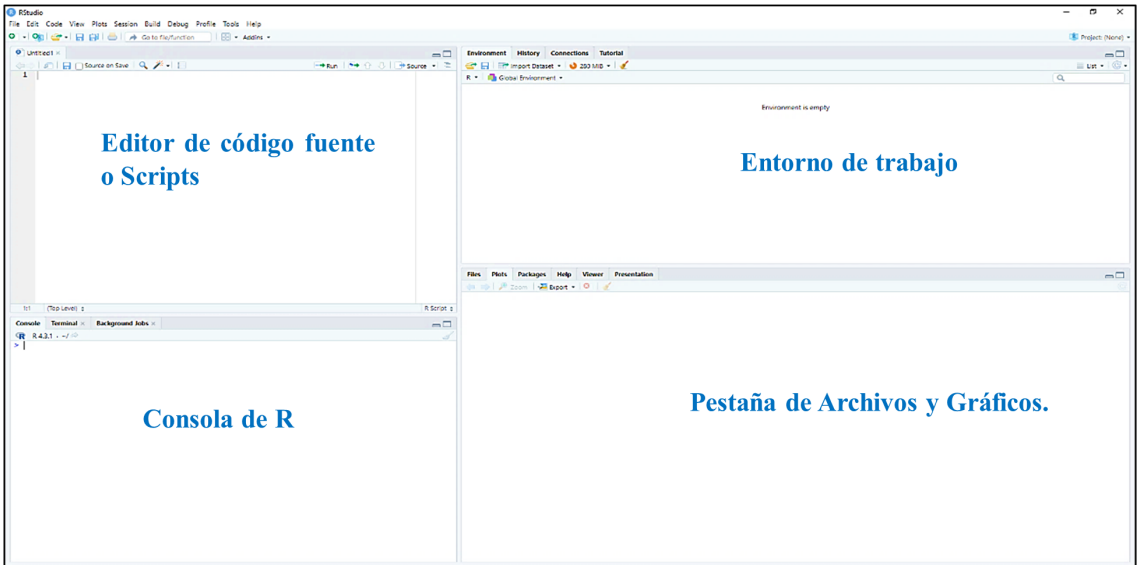


Figura 4.5: Figura 5

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).

> **!** Importante
>
> 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).

4.1.2 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 & 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 & 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.

Los analísis 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 & Cuccurullo, 2017) permite realizar análisis bibliométrico de publicaciones científicas, identificando tendencias emergentes y redes de colaboración que orientan nuevas investigaciones.

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.

4.1.2.2 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 & 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 & Shetty, 2017), 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 & 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 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:

(i) transcriptómica,
(ii) proteómica y
(iii) 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.

4.2 Análisis Bibliométrico para el Diseño de Experimentos y la Microbiología Industrial

El paquete **Bibliometrix**, desarrollado en R, constituye una herramienta clave para el análisis bibliométrico en distintas áreas del conocimiento, entre ellas la **microbiología industrial** y el **diseño experimental**. Su enfoque de código abierto permite recopilar, analizar y visualizar información científica de manera integral, ofreciendo una visión clara sobre las principales tendencias y la evolución de la investigación en cada campo (Aria & Cuccurullo, 2017).

En la **microbiología industrial**, **Bibliometrix** se ha convertido en un apoyo fundamental para reconocer **líneas emergentes de investigación**, **colaboraciones internacionales** y **autores influyentes** que marcan el desarrollo del área (Aria & Cuccurullo, 2017). A través de su interfaz visual **Biblioshiny**, los análisis complejos se vuelven accesibles incluso para quienes no tienen experiencia en programación. Esta accesibilidad favorece la creación de **mapas de conocimiento**, **redes de colaboración** y **agrupamientos temáticos** que ayudan a identificar oportunidades de trabajo conjunto, vacíos en la literatura o la evolución de determinadas técnicas experimentales.

El uso de Bibliometrix y Biblioshiny permite una comprensión argumentada del panorama científico, fomentando decisiones de investigación basadas en evidencia y fortaleciendo la planificación de proyectos dentro de la microbiología industrial.

4.2.1 Procedimiento para el Análisis Bibliométrico con Bibliometrix a partir de una Base de Datos Scopus

Como ejemplo de análisis bibliometrico se emplearán los datos de trabajo de grado titulado (sin publicar) “Evaluación del crecimiento de *Cordyceps militaris* en diferentes sustratos vegetales”(Chala, 2025).

4.2.2 Palabras clave, operadores Booleanos y búsqueda en bases de datos

El primer paso consiste en generar cinco palabras claves relacionadas con el tema de estudio, como son:

1. *Cultivation*: proceso de cultivar *Cordyceps militaris* en condiciones controladas paraestudiar su crecimiento y desarrollo.

2. *Mycelial growth*: crecimiento del micelio, la parte vegetativa del hongo, que escrucial para evaluar su desarrollo.

3. *Substrate optimization*: mejora de los sustratos utilizados para el cultivo del hongo, conel fin de maximizar su crecimiento y producción de compuestos bioactivos.

4. *Bioactive compounds*: compuestos producidos por Cordyceps militaris, como la cordicepina, quetienen propiedades medicinales y son un indicador de la calidad del crecimiento.

5. *Fermentation conditions*: condiciones de fermentación, como la temperatura, pH y nutrientes, que afectan el crecimiento y la producción de metabolitos del hongo.

El segundo paso es generar diez (10) ecuaciones de búsqueda ingresando a una de las siguientes base de revistas indexadas compatibles con Bibliometrix, como son: Web of Science (www.webofscience.com), Scopus (www.scopus.com); OpenAlex (www.openalex.org); Dimensions (www.dimensions.ai); The Lens (www.lens.org); PubMed (https://pubmed.ncbi.nlm.nih.gov/) y Cochrane Library (www.cochranelibrary.com/), desde donde se utilizaran operadores booleanos relacionados con el tema, es importante que cada una de las palabras clave estén encerradas entre comillas " " ; al tiempo que se utilicen los conectores: AND y/o OR y NOT para refinar la búsqueda, como se muestran a continuación:

- "*Cordyceps militaris*" AND "growth evaluation"
- "*Cordyceps militaris*" AND ("substrate optimization" OR "culture medium") AND "growth"
- "Cordyceps militaris" AND "cordycepin" AND "growth"
- "*Cordyceps militaris*" AND "temperature" AND "mycelial growth"
- "*Cordyceps militaris*" AND "fermentation" AND ("solid-state fermentation")
- "*Cordyceps militaris*" AND ("natural growth" OR "artificial cultivation") AND ("yield" OR "biomass production")
- "*Cordyceps militaris*" AND ("metabolite profiling" OR "secondary metabolites") AND "growth conditions"
- "*Cordyceps militaris*" AND ("carbon source" OR "nitrogen source") AND "mycelial biomass"
- "*Cordyceps militaris*" AND ("rice medium" OR "wheat medium" OR "synthetic medium") AND "growth performance"
- "*Cordyceps militaris*" AND "commercial production" AND ("growth enhancement" OR "yield improvement"

En la Tabla 1, se aprecian el número de publicaciones científicas relacionadas con algunas de las ecuanciones de búsqueda, lo que permitirá la elección del resultado más promisorio, para seguidamente proceder con el proceso de descarga de los metadatos en formato .csv.

Tabla 4.1: Tabla 1. Salida de resultados para cada uno de los operadores booleanos, introducidos dentro de la plataforma de Scopus®, y que están con el tema de investigación de Cordyceps militaris.

Ecuación de búsqueda	Documentos encontrados
"Cordyceps militaris" AND ("metabolite" OR "growth conditions"	225
"Cordyceps militaris" AND ("cordycepin") AND "growth"	191
"Cordyceps militaris" AND "medium" AND "growth"	99
"Cordyceps militaris" AND ("substrate optimization" OR "culture medium") AND "growth"	40

4.2.3 Analísis con bibliometrix

Para iniciar el análisis bibliométrico, se procede a ingresar al programa RStudio®, al tiempo que se realiza la instalación de los paquetes: *bibliometrix* y *bibliometrixData*, desde la pestaña de Archivos y Gráficos en la sección de Packages (Figura 8).

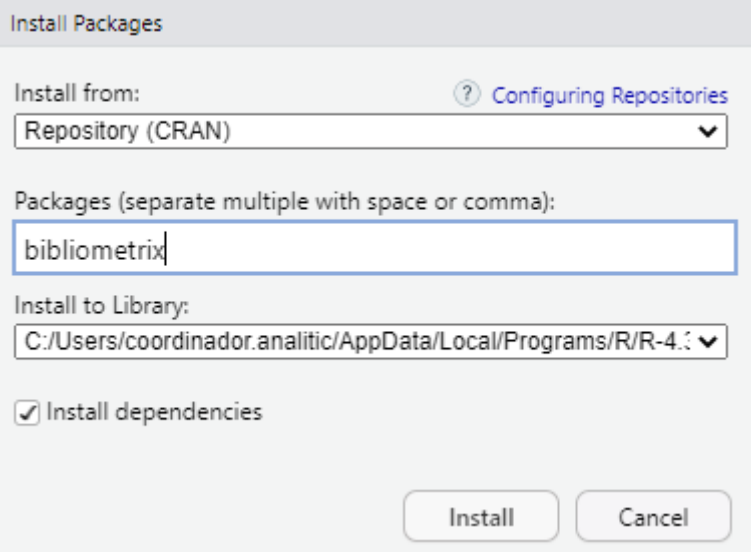


Figura 4.6: Figura 8.

Posteriormente, desde la Consola de RStudio se digitan y ejecutan los comandos:

library(bibliometrix) biblioshiny()

Tambien se puede habilitar manualmente en la pestaña de Archivos y Graficos de la interfaz de RStudio, pestaña de Packages (figura9) seleccionado las librerias a usar.

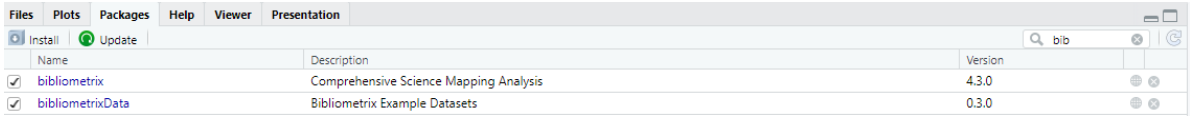


Figura 4.7: Figura 9.

Se abrirá el servidor de Bibliometrix (Figura 9) en el navegador web (Chrome, Mozilla, Edge, entre otros). Es importante aclarar que la interfaz de Bibliometrix únicamente pude ser ejecutada desde RStudio®.

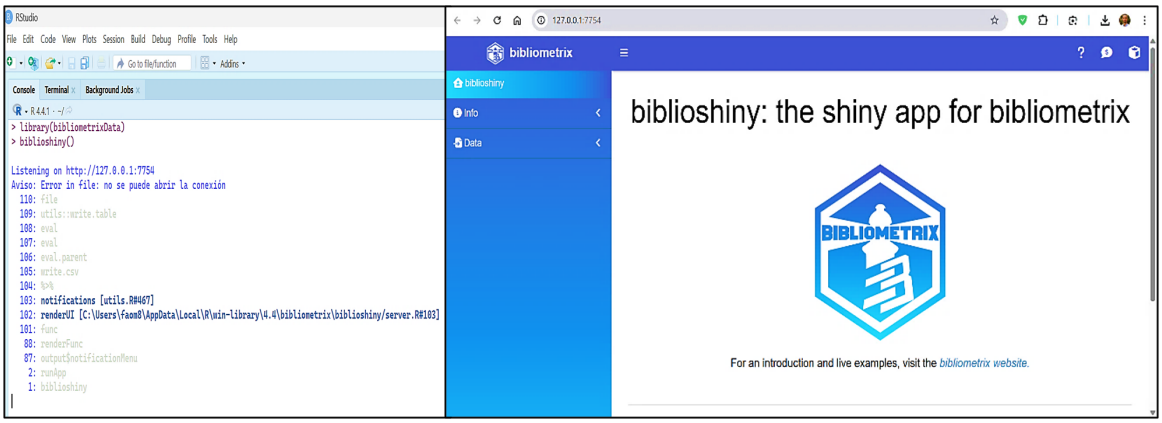


Figura 4.8: Figura 9.

Siguiendo las indicaciones de (Aria & Cuccurullo, 2017), en la parte izquierda se despliega el menú de biblioshiny, se procede con la importación en la seción "*Import or Load*" del archivo CSV, al tiempo que se seleccionan las casillas: *Import raw file(s)*, la procedencia de la base de datos consultada (*Scopus,* en nuestro caso) , junto la opción: *Surname and Initials*, y finalmente damos click en el botón de *Start* (Figura 10).

Figura 4.9: Figura 10.

Después se despliega una nueva ventana que muestra el estado de los componentes de los metadatos importados desde el archivo CSV (Figura 11), allí se muestra una tabla que resume la completitud de los metadatos (para nuestro ejemplo: 191 documentos de Scopus),.

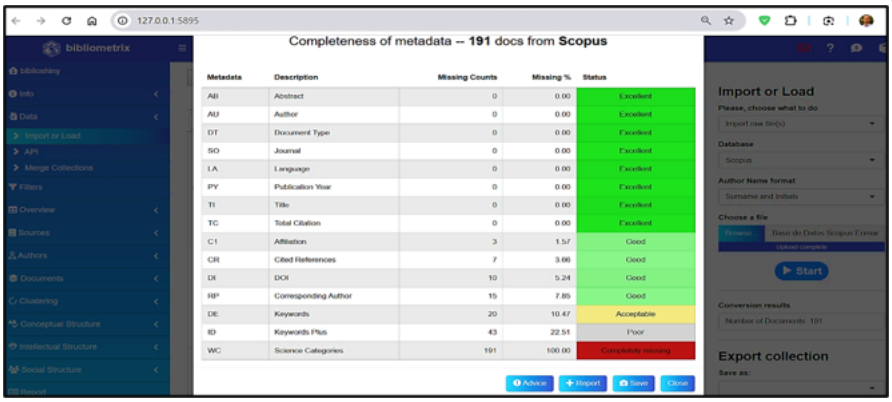


Figura 4.10: Figura 11.

Así mismo Bibliometrix evalúa cada uno de los diferentes campos de metadatos (Abstract, afiliación, autor, tipo de documento, etc.) utilizando *junto a* diferentes criterios de clasificación como son: “Excelente”, “Bueno”, “Aceptable” “Pobre” y “Completamente perdido” (Figura 11), y que al ser cerrado “Close” muestra una tabla dinámica con todos los datos importados incluyendo el DOI, desde el cual se puede acceder y leer directamente el articulo científico (Figura 12).

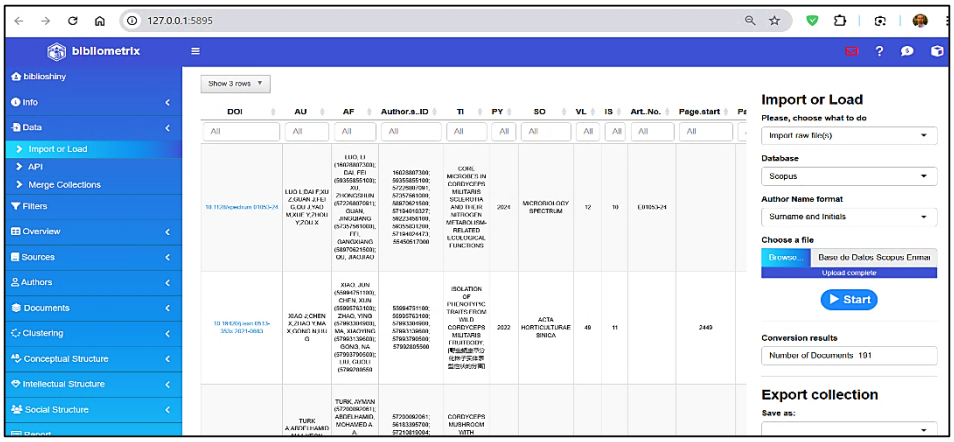
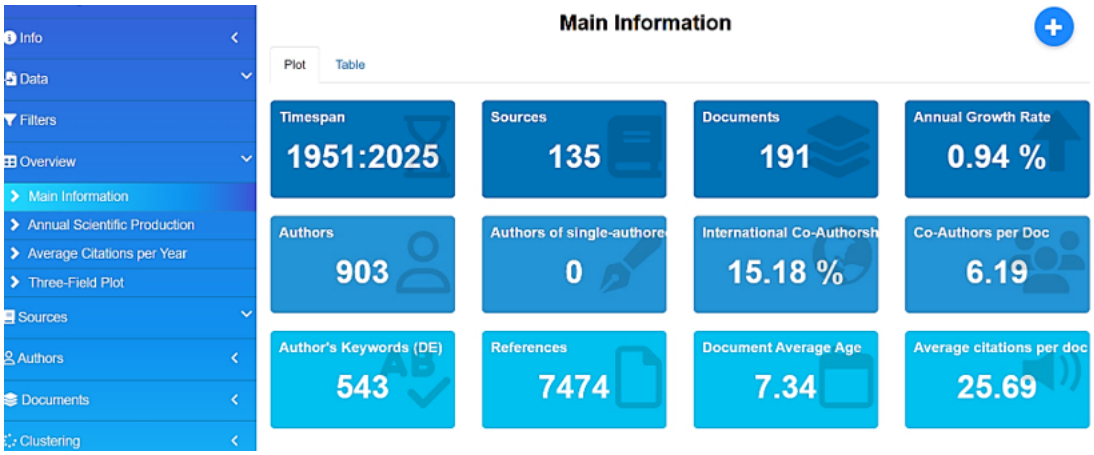


Figura 4.11: Figura 12

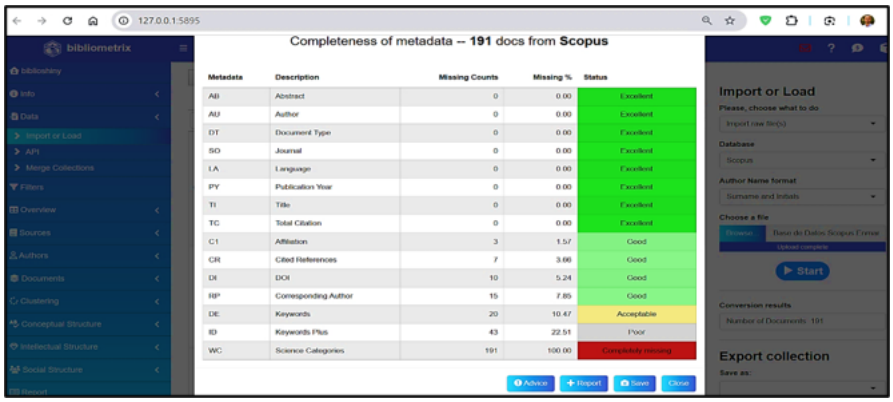
En el menú ***Overview*** se despliega la opción ***Main information*** (Traducido: Información Principal) la cual muestra de forma dinámica y visual los valores de los metadatos descargas, para nuestro ejemplo didáctico tenemos que : El análisis bibliométrico abarca el periodo de 1951 a 2025 con un total de 135 fuentes y 191 documentos, la tasa de crecimiento anual es de 0,94%, se identificaron 903 autores, sin registros un solo autor; la coautoría internacional alcanza un valor de 15,18% y el promedio de coautores por documento es de 6,19. Se encontraron 543 palabras clave de autor y 7474 referencias, la edad promedio de los documentos es de 7,34 años y la cantidad media de citas por documento es de 25,69 (Figura 13) ***Falta***.



En la sección ***Annual Scientific Production*** (traducido: Producción Científica Anual) y continuando con ejemplo didáctico se muestra que: la evolución de la cantidad de artículos científicos relacionados con Cordyceps militaris publicados por año, van desde 1951 hasta aproximadamente el año 2000, la producción científica es muy baja, con solo unos pocos artículos publicados anualmente; a partir del 2000, se observa un ligero aumento en la cantidad de publicaciones, pero es a partir de 2010 cuando el crecimiento se acelera considerablemente,

reflejando una tendencia ascendente más marcada. Entre 2015 y 2024, la producción científica alcanza sus niveles más altos, con un notable incremento en la cantidad de artículos publicados cada año. Sin embargo, en 2025 se observa una caída significativa, lo que podría deberse a datos incompletos, retrasos en las publicaciones o factores externos como cambios en políticas de financiamiento o publicación (Figura 14).

Para la sección ***Average Citations Per Year*** (Traducido: Citas Promedio por año) se muestra la evolución del número promedio de citas recibidas por los artículos a lo largo del tiempo. Para nuestro ejemplo didáctico en la temática de Cordyceps militaris, específicamente en 1951 y 1961, se observan valores relativamente altos, pero con pocos artículos publicados en esos períodos. Posteriormente, durante varias décadas, el número promedio de citas se mantiene bajo y estable. A partir del año 2000 aproximadamente , hay un incremento notable en la cantidad de citas por artículo, alcanzando picos significativos en el 2006 y 2008 (Figura 15).



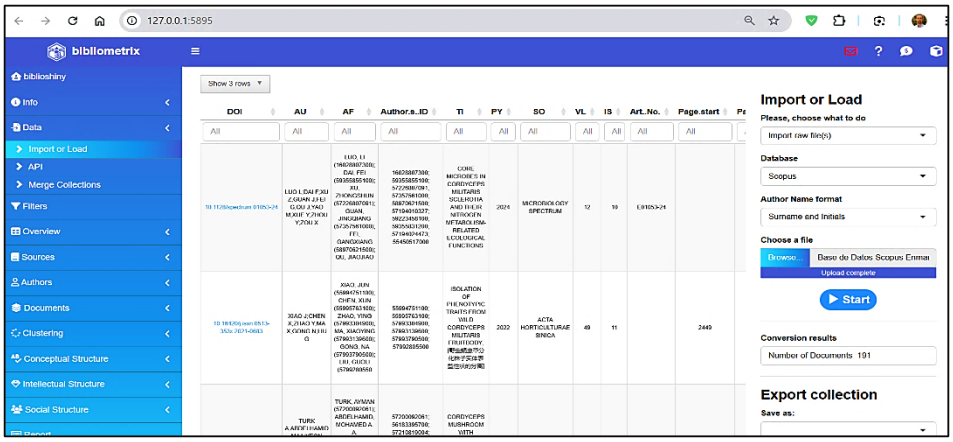
No corresponde

Para el menú de *Overview* específicamente en la sección ***Three-Field Plot*** (Traducido: Trazo de tres campos) se proporciona una visión general de la colaboración en investigación y los

temas principales dentro de un cuerpo específico de literatura. *AU_CO (Países)*: Muestra la participación de China, Tailandia, Reino Unido, Singapur, estados Unidos , entre otros; en los artículos analizados. China lidera la investigación del tema, sugiriendo una fuerte presencia investigadora en el área del estudio; *AU (Autores)*: destaca a: Li X, Li Y, Vongsangnak W y Zhang J, como los autores más productivos, al tiempo que muestra posibles colaboraciones o equipos de investigación.; *DE (Descriptores)*: Los términos que se asocian el estudio del hongo *Cordyceps militaris*, la biotecnología del hongo, compuestos bioactivos y análisis transcriptómico. Este gráfico reafirma el liderazgo asiático en esta línea de investigación, así como la concentración temática en los compuestos bioactivos de Cordyceps militaris y su aplicación en salud y biotecnología (Figura 16). Importante destacar que Bibliometrix permite cambiar los parametros de análisis y el número de ítems.

Figura 4.12: Figura 14.

Para el menú de *Sources* en la sección ***Most Relevant Sources*** (Traducido: Fuentes más Relevantes) la producción científica procedente de Scopus está lideradas por: International Journal of Medicinal Mushrooms con 15 artículos publicados, seguido de Mycosystema con 8, le sigue: Applied Microbiology and Biotechnology con 5. Dichas revistas destacan por su enfoque en microbiología, biotecnología y farmacología, áreas clave dentro del estudio abordado. La distribución sugiere que la investigación en este campo se encuentra bien representada en revistas especializadas,las demás revistas como: Biology, Bioresource Technolgy, y Nutrients, que cuentan con 3 artículos cada una (Figura 17).



No corresponde

En la sección de **Bradford's Law** (Traducido: La Ley de Bradford) y continuando con nuestro ejemplo didáctico de *Cordyceps mil*itaris ,se observa que: "International Journal of Medicinal Mushrooms", junto con "Mycosystema" y "Applied Microbiology and Biotechnology", conforman el núcleo de fuentes indexadas más relevantes, aportando el mayor número de publicaciones, estas tres revistas están dentro del área sombreada, lo que confirma su papel central en la diseminación del conocimiento sobre *C. militari*s y compuestos bioactivoss. A medida que se avanza hacia la derecha del gráfico, el número de artículos por revista disminuye, lo que representa publicaciones de interés más disperso (Figura 18).

Para el menú de *Authors* específicamente en: ***Authors' Production over Time*** (traducido: La producción de los autores a lo largo del tiempo) se muestra que investigadores: Wang Y., Li X., y Li Y., han mantenido una producción constante en los últimos años, con picos significativos en 2020 y 2022. La visualización indica que hay una concentración de publicaciones en los últimos cinco años, lo que sugiere un crecimiento en la investigación dentro de este campo. Otros autores, como Vongsangnak W. y Zhang J., han contribuido de manera más esporádica, pero siguen participando activamente en la investigación del hongo *Cordyceps militaris* (Figura 19).

Para el menú de *Authors* específicamente en: ***Countries' Scientific Production*** (traducido: Producción científica de los países), el mapa muestra la distribución geográfica de la producción científica. China es el país con mayor producción científica (azul oscuro); otros países con destacada producción científica entre los que se incluyen: Estados Unidos, Corea del Sur, Tailandia, Japón, India y varios países europeos y asiáticos (azul celeste). Algunos países no presentan producción registrada y aparecen coloreados en gris. En la tabla se observa que China lidera con 702 publicaciones, seguida por Corea del Sur con 212, Tailandia con 83, Japón con 50 e India con 39. Otros países con menor producción incluyen Estados Unidos con 9, Reino Unido con 8, Alemania con 7, Italia con 6 y Colombia con 5 (Figura 20).

Para el menú de *Document* específicamente en: ***Most Frequent Words*** (traducido: Palabras más frecuentes), la visualización muestra las palabras de uso frecuente con el tema de

investigación de Cordyceps militaris son: Cordyceps con 196 apariciones y Cordycepin con 187. Otras palabras destacadas incluyen article con 109, Cordyceps militaris con 108, nonhuman con 92, metabolism con 79, deoxyadenosines con 77, controlled study con 76, deoxyadenosine derivative con 65 y adenosine con 64 (figura 21).

Para el menú de *Document* específicamente en: ***Reference Spectroscopy*** (traducido: Espectroscopia de referencias), la visualización muestra la evolución de las referencias citadas en espectroscopia a lo largo del tiempo; antes de 1990, el número de referencias citadas se mantuvo prácticamente nulo y a partir de 1995, se observa una pendiente creciente sostenida, que se acelera alrededor del año 2005, alcanzando su punto máximo entre 2015 y 2020 con más de 400 referencias citadas por año, después de 2018, se evidencia una disminución en el número de referencias citadas. También se puede observar una caída abrupta después del 2020 que puede explicarse por efectos de rezago en la citación, ya que las publicaciones recientes aún no han tenido suficiente tiempo para acumular cita (Figura 22).

Para el menú de *Conceptual Structure* concretamente en: ***Co-occurrence Network*** (traducido: Red de Coocurrencias), la red se encuentra claramente dividida en dos comunidades principales, identificadas por los colores rojo y azul. La comunidad roja, dominada por términos como cordycepin, Cordyceps militaris, metabolism y article, se orienta al estudio bioquímico y farmacológico del compuesto, mientras que la comunidad azul está asociada a modelos experimentales, destacando términos como animal experiment, human, mouse y cell line. Esta segmentación temática sugiere una dualidad en la línea de investigación: una centrada en la caracterización química y otra en los efectos biológicos en modelos preclínicos. El análisis de centralidad (como grado y betweenness) permitiría identificar términos puente como: nonhuman o controlled study, que conectan ambas comunidades (Figura 23).

Para el menú de ***Conceptual Structure*** concretamente en: ***Factorial Analysis*** (traducido: Análisis factorial), se muestra una representación bidimensional de los términos más relevantes dentro del corpus bibliográfico analizado, en el eje X (Dim 1), que explica el 50,18% de la variabilidad, se observan términos fuertemente relacionados con estudios experimentales: in vivo e in vitro, como in vitro study, mouse, animal tissue y protein expression, agrupados en el cuadrante inferior derecho. Esto sugiere una fuerte carga temática asociada a investigaciones biomédicas y farmacológicas; por otro lado, en el eje Y (Dim 2), que explica un 9,68% adicional, aparecen términos como: transcriptome y carbon, más vinculados a estudios genéticos y metabólicos, separados del resto de la nube léxica; dicha segmentación espacial revela la existencia de subdominios temáticos diferenciados dentro del campo de estudio de los cordyceps y sus derivados, evidenciando un enfoque dual: uno centrado en la bioquímica y biotecnología del hongo, y otro enfocado en los ensayos experimentales en organismos modelo (Figura 24).

Ingresando en menú de ***Social Structure*** concretamente en: ***Collaboration Network*** (traducido: Red de colaboración), según Aria & Cuccurullo (2017) el análisis se realiza mediante redes de coautoría, las cuales revelan patrones de colaboración y productividad; continuando con el tema de *Cordyceps militaris*, se muestra una red de colaboración entre autores en la que los nodos representan investigadores y las conexiones indican coautoría en

publicaciones científicas, se observan varios grupos de colaboración con diferentes colores lo que sugiere comunidades de investigadores que trabajan juntos frecuentemente Algunos nodos como Li X y Vongsangnak W tienen un tamaño mayor lo que indica que son autores con un alto número de colaboraciones mientras que otros aparecen más aislados mostrando menos conexiones dentro de la red (Figura 25).

Al ingresar al menú de ***Social Structure*** en la sub sección: ***Countries' Collaboration World* Mak** (traducido: Mapa mundial de colaboración entre países) se representa visualmente las redes de colaboración científica entre países en torno a investigaciones relacionadas con el género *Cordyceps* y el compuesto activo Cordycepina; La intensidad del color azul en el mapa refleja el volumen de publicaciones: a mayor intensidad, mayor producción científica: en este caso, se observa que China destaca significativamente como el nodo más activo, lo cual es consistente con su liderazgo en investigaciones sobre hongos medicinales. La Figura 26 muestra un patrón de colaboración transcontinental, con conexiones entre China y países como Estados Unidos, Alemania, Corea del Sur, y Australia, lo cual sugiere una red científica relativamente globalizada, esta interacción internacional favorece la transferencia de conocimiento, fortalece la calidad metodológica de los estudios y facilita el acceso a recursos técnicos avanzados, dicho mapa es útil para identificar núcleos de producción científica, barreras geográficas o idiomáticas, y oportunidades de cooperación estratégica entre países.

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

5.1 2. Fundamentos del Diseño Experimental quarto

El diseño experimental corresponde a una metodología científica y estadística destinada a planear, ejecutar y analizar pruebas controladas, con el propósito de obtener evidencia objetiva que responda a interrogantes sobre procesos o fenómenos específicos. El **diseño de experimentos (DOE)** se diferencia de la práctica empírica de prueba y error porque estructura el proceso investigativo bajo principios formales que permiten generar información confiable, optimizar recursos y reducir incertidumbre (Gutiérrez Pulido & Vara Salazar, 2012).

El DOE se da en ámbitos industriales y de investigación aplicada, los experimentos suelen realizarse para resolver problemas de calidad, mejorar procesos o comprobar hipótesis sobre materiales, condiciones de operación o métodos de trabajo. Sin embargo, cuando estas pruebas carecen de planeación rigurosa, se corre el riesgo de interpretar datos de manera subjetiva y desaprovechar el potencial de la variabilidad natural del sistema. Por ello, el DOE proporciona un marco que asegura resultados válidos y generalizables (Gutiérrez Pulido & Vara Salazar, 2012).

En cuanto a la terminología básica, conceptos como unidad experimental, tratamiento, factor controlable y no controlable, niveles de los factores, variable de respuesta, repetición y matriz de diseño, es requerido manejarlos. Estos términos constituyen la gramática operativa del diseño experimental, permitiendo estructurar adecuadamente las hipótesis y la recolección de datos. Además, se distingue entre error aleatorio y error experimental, resaltando la necesidad de minimizar y cuantificar ambos para garantizar validez estadística. Entre las etapas del diseño experimental, es incluyen:

- **Planeación:** formulación del problema, identificación de factores y niveles, selección de variables de respuesta y definición de objetivos.
- **Ejecución:** implementación del plan experimental bajo condiciones de control y aleatorización.
- **Análisis:** aplicación de métodos estadísticos, principalmente análisis de varianza (ANOVA), para estimar efectos principales e interacciones.
- **Interpretación:** extracción de conclusiones técnicas y toma de decisiones basadas en la evidencia.

Un aporte central son los principios básicos del DOE:

- **Aleatorización**, que asegura independencia de los errores y evita sesgos sistemáticos.
- **Replicación,** que incrementa la precisión de las estimaciones al cuantificar la variabilidad experimental.
- **Bloqueo**, que controla fuentes de variación no deseadas (turno, lote, operador), incrementando la potencia estadística del experimento.

Estos principios permiten estructurar experimentos que sean eficientes en costo y tiempo, pero robustos en cuanto a la validez de sus conclusiones.

La clasificación de diseños va desde los más simples (completamente al azar, bloques completos, cuadrados latinos) hasta los más complejos (factoriales, fraccionados, superficies de respuesta, diseños robustos). Se subraya que la selección depende de los objetivos, el número de factores, las restricciones prácticas y el tipo de información buscada. También se enfatiza que la decisión debe considerar tanto la significancia estadística como la significancia práctica, es decir, el impacto real de los resultados sobre el proceso o fenómeno bajo estudio (Gutiérrez Pulido & Vara Salazar, 2012).

5.1.1 2.1 Tipos de diseños experimentales

La selección de un diseño experimental depende de distintos factores que condicionan su pertinencia y aplicabilidad en cada situación. Entre los aspectos determinantes se encuentran: los objetivos que se persiguen con el estudio, la cantidad de factores que se desea analizar, el número de niveles que adoptará cada factor, los efectos que se pretende identificar en la relación causa-efecto y, finalmente, las restricciones de costo, tiempo y precisión que impone la investigación (Gutiérrez Pulido & Vara Salazar, 2012).

Estos elementos no actúan de forma aislada, ya que la modificación de cualquiera de ellos obliga generalmente a replantear el diseño a utilizar. En consecuencia, resultan fundamentales para guiar la clasificación de los diseños experimentales.

El **objetivo del experimento** constituye el criterio principal para diferenciar entre tipos de diseño, mientras que los demás factores funcionan como subcriterios de clasificación. Bajo esta perspectiva, los diseños pueden agruparse en varias categorías: aquellos orientados a la comparación de dos o más tratamientos; los que examinan el efecto de diversos factores sobre una o varias variables de respuesta; los que buscan establecer el punto óptimo de operación de un proceso; los que se enfocan en la optimización de mezclas; y finalmente, los dirigidos a lograr que un producto o proceso se mantenga estable frente a factores no controlables (Gutiérrez Pulido & Vara Salazar, 2012).

Así, la clasificación general de los diseños experimentales responde al objetivo central del estudio, y dentro de cada categoría se consideran elementos adicionales como el número de factores, los tipos de efectos a investigar y las restricciones prácticas que condicionan la ejecución.

5.1.2 Clasificación de los diseños experimentales

La siguente clasificación es tomada el libro de (Gutiérrez Pulido & Vara Salazar, 2012).

1. **Diseños para comparar dos o más tratamientos**
   Diseño completamente al azar
   Diseño de bloques completos al azar
   Diseño de cuadros latino y grecolatino

2. **Diseños para estudiar efectos de varios factores sobre una o más variables de respuesta**
   Diseños factoriales 2
   Diseños factoriales 3
   Diseños fraccionados 2
   Diseños anidados
   Diseños en parcelas divididas

3. **Diseños para la optimización de procesos**
   Modelo de primer orden
   Diseños factoriales 2 y 2
   Diseño de Plackett-Burman
   Diseño simplex
   Modelo de segundo orden
   Diseño de composición central
   Diseño de Box-Behnken
   Diseños factoriales 3 y 3

4. **Diseños robustos**
   Arreglos ortogonales (factoriales)
   Diseño con arreglos interno y externo

5. **Diseños de mezclas**
   Diseño simplex-reticular
   Diseño simplex con centroide
   Diseño sin restricciones
   Diseño axial

5.1.3 2.2 Ejemplos prácticos de diseños experimentales en Microbiología Industrial

5.1.3.1 2.2.1 Diseño Completamente al Azar OK

5.1.3.2 Problema

Introducción: La **antracnosis del banano** , causada por *Colletotrichum musae* (Berk. y M.A. Curtis) Arx, representa una problemática fitosanitaria de considerable relevancia económica en la industria bananera mundial, puesto que genera pérdidas postcosecha que oscilan entre el 10 y 80% debido al deterioro de la calidad visual del fruto, dicho patógeno desarrolla lesiones (formación de acérvulos) de coloración marrón oscuro a negro en el epicarpio del fruto, las cuales afectan la calidad visual del fruto (Vásquez-Castillo et al., 2019).

Tradicionalmente, el manejo de esta epifítia se ha fundamentado en la aplicación de fungicidas sintéticos como: tiabendazol, azoxystrobin y trifloxystrobin; no obstante, estas sustancias generan impactos ambientales adversos y residualidad ((Arias B., 2007), por ello, la búsqueda de alternativas de biocontrol sostenibles ha cobrado especial relevancia, particularmente mediante el uso de extractos fúngicos con propiedades antagónicas.

Metodología: El estudio se estructuró a partir de dos diseños experimentales: un **Diseño Completamente al Azar** para la evaluación de sustratos, y un **Diseño de Medidas Repetidas en el Tiempo** para la evaluación de la actividad inhibitoria.

5.1.3.3 Diseño 1: Sustratos de cultivo para *Penicillum* sp.

Se empleó un **Diseño Completamente al Azar** con los siguientes tratamientos: avena en hojuelas, maíz partido, semillas de cebada y arroz blanco. Se prepararon bolsas de polipropileno con cada sustrato, se inocularon con cinco discos de micelio de *Penicillum* sp. (0.5 mm de diámetro) y se incubaron de forma aleatorizada a 22 ± 2 °C durante ocho días. El experimento se realizó por quintuplicado, considerando cada bolsa como una repetición.

5.1.3.4 Diseño 2: Evaluación de la actividad inhibitoria

Se implementó un **Diseño de Medidas Repetidas en el Tiempo** para analizar el efecto de las concentraciones del extracto sobre dos variables de respuesta clave:

- **Porcentaje de Inhibición del Área de la Lesión (PIAL)**: Para evaluar la eficacia `in vivo`.
- **Porcentaje de Inhibición del Crecimiento Micelial (PICM)**: Para evaluar la eficacia `in vitro`.

Las variables independientes fueron las diferentes concentraciones del extracto y los testigos correspondientes, mientras que las variables de respuesta se midieron a lo largo del tiempo para observar la evolución de la inhibición.

Resultados: El maíz partido constituyó el sustrato óptimo para la producción conidial de *Penicillium digitatum*, alcanzando valores de Log_{10} 9,13 conidios/mL, seguido de la cebada Log_{10} 8,88 conidios/mL (**Figura 1**).

Figura 1.
Sustratos con Conidios de *Penicillium* sp.

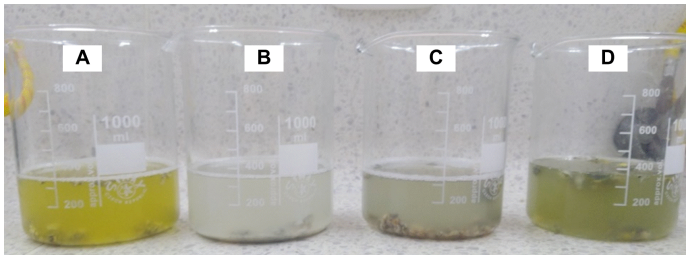


Figura 5.1: Nota: Dilución de conidios y sustrato, en solución tween80® 0,01%: Avena (A); Arroz (B); Cebada (C); Maíz Partido (D).

La evaluación in vitro reveló que las concentraciones de extracto crudo de 4,0 al 6,0% generaron Porcentajes de Inhibición del Crecimiento micelial (PICM) del 40 al 50 % respectivamente al quinto día después de la inoculación (ddi) **(Figura 2)**.

Figura 2.
Efecto de los tratamientos in vitro frente al crecmiento de *Colletotrichum musae.*

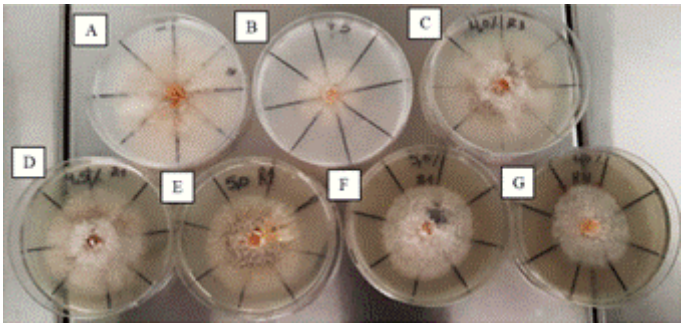


Figura 5.2: Nota: Prueba de inhibición in vitro de Colletotrichum musae, frente a diferentes tratamientos. (A) Testigo negativo; (B) Testigo positivo (Amistar a 60mg/100mL); (C) Extracto de *Penicillium* sp., al 4%; (D) Extracto de *Penicillium* sp., al 4,5%; (E) Extracto de *Penicillium* sp., al 5%; (F) Extracto de *Penicillium* sp., al 5,5%; (G) Extracto de *Penicillium* sp., al 6%.

Por otro lado, los ensayos in vivo evidenciaron una mayor eficacia del extracto crudo, donde las concentraciones de 8, 9, 10, 11, 12 y 13% generaron porcentajes de inhibición del área de la lesión (PIAL) de 60, 55, 70, 72, 77 y 80% respectivamente **(Figura 3)**, sugiriendo que Penicillium digitatum podría representar una alternativa viable para el manejo preventivo de la antracnosis del banano.

Figura 3.
Efecto in vivo de bananos infectados con *Colletotrichum musae* en los tratamientos.

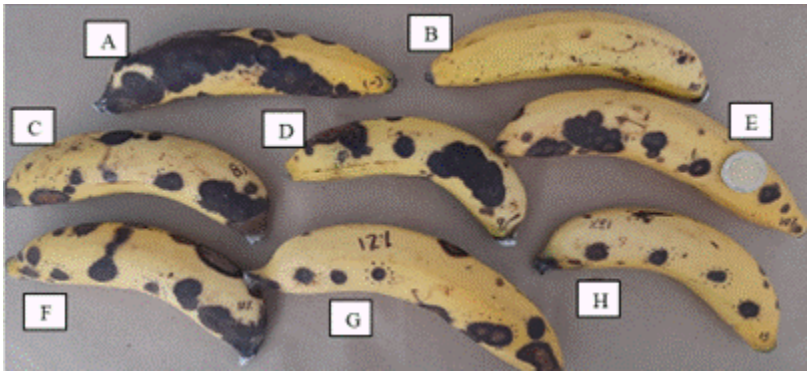


Figura 5.3: Nota: Experimento in vivo de los bananos infectados con 107 conidios de Colletotrichum musae, frente a tratamientos (A los 7 días de la inoculación). (A) Testigo negativo; (B) Azoxystrobin (Testigo positivo); Extractos de Penicillium sp. a (C) 8%; (D) 9%; (E) 10%; (F) 11%; (G) 12%; (H) 13%.

Para mayor información puede consultar: Mejía-Sarmiento, J. S. (2022). Evaluación de Extracto Crudo de Penicillium sp. para la Inhibición del Crecimiento in vitro e in vivo de Colletotrichum musae (Berk. y M. A. Curtis) Arx. Agente Causal de Antracnosis en Banano [Tesis de pregrado, Universidad de Santander UDES]. Repositorio Institucional UDES. https://repositorio.udes.edu.co/handle/001/8674

5.1.4 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:/R-Proyectos/r-para-mi/data/dca.xlsx")

View(DCA)
attach(DCA)
names(DCA)
```

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

```
str(DCA)
```

```
tibble [16 x 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).
```

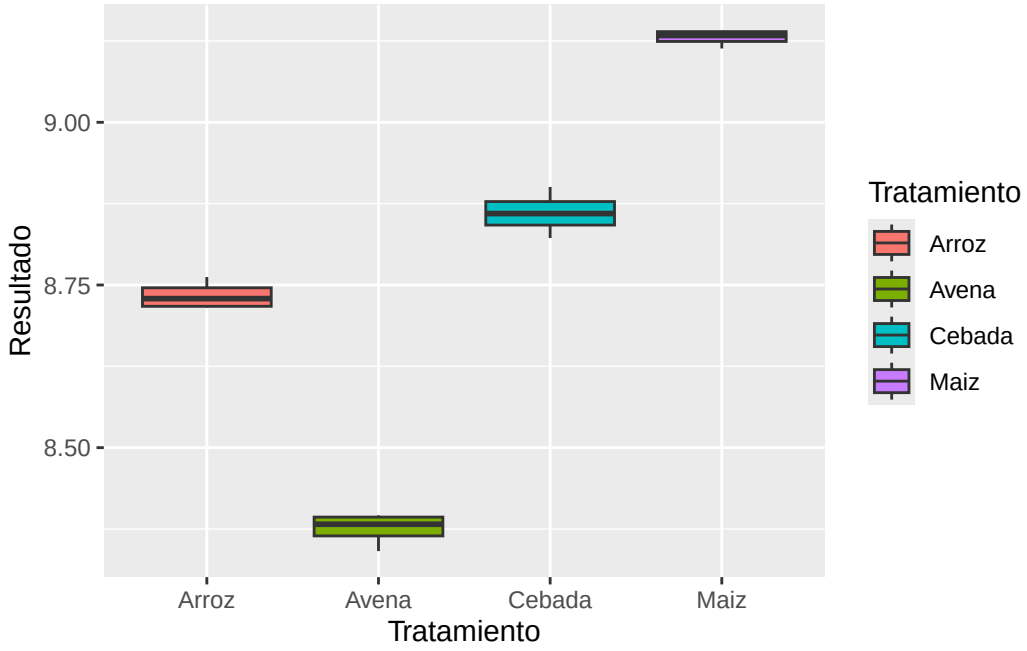
El modelo explica el 99.4% de la variabilidad en los datos ($R^2 = 0.994$), y el error estándar residual es bajo (0.0244), lo que indica un ajuste excelente.

Gráfico Boxplot

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

```r
library(ggplot2)

DCA$Treatamiento<-factor(DCA$Tratamiento) #transformamos una variable númerica en un factor c
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:

```r
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

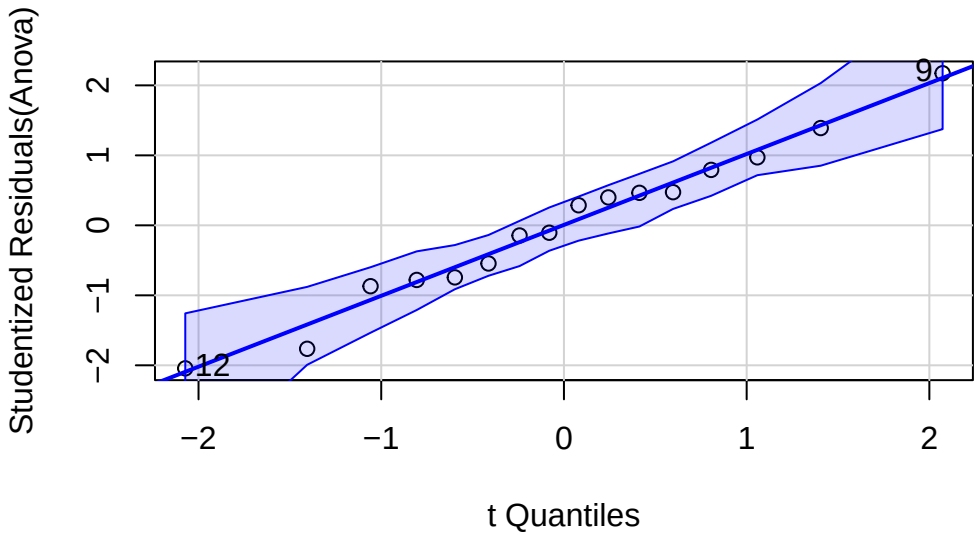
Para construir el gráfico QQ (QQ plot) y evaluar la normalidad de los datos, se utiliza la función correspondiente del paquete car. Si no está instalado previamente, es necesario instalar también el paquete auxiliar carData.

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:

```r
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)
```

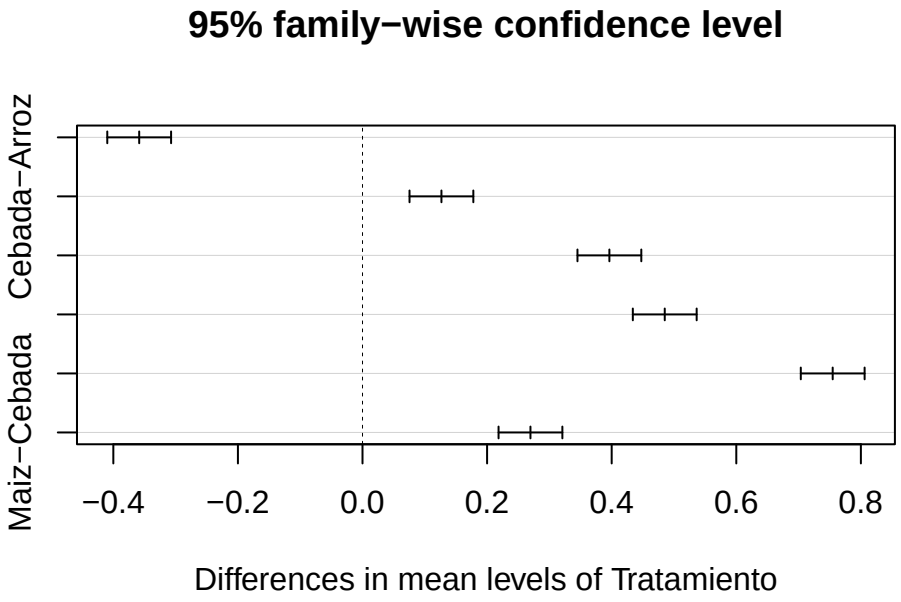
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

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.

5.1.4.1 2.2.2 Diseño de bloques completamente al azar OK

5.1.4.2 2.2.3 Diseño longitudinal (ANOVA de medidas repetidas) OK

5.1.5 Problema

> **!** Importante
>
> **Metodología:** Se realizó un diseño de medidas repetidas en el tiempo, donde la variable independiente fue cada una de las concentraciones del extracto y los testigos; y la variable respuesta fueron: Porcentaje de Inhibición del Área de la Lesió (PIAL) y Porcentaje de Inhibición del Crecimiento Micelial (PICM).

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

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

Arias B., C. L. (2007). Control Químico de la Antracnosis del Mango (Mangifera indica L.) en pre y postcosecha. *Bioagro*, *19*(1), 19-25. http://www.ucla.edu.ve/bioagro/Rev19(1)/3.%20Control%20qu%C3%ADmico%20de%20la%20antracnosis.pdf

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

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

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

Gutiérrez Pulido, H., & Vara Salazar, R. de la. (2012). *Análisis y diseño de experimentos* (3.ª ed.). McGraw-Hill/Interamericana Editores, S.A. de C.V.

Lahti, L., & Shetty, S. (2017). *microbiome R package*. http://microbiome.github.io/microbiome

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. (2020). *agricolae: Statistical Procedures for Agricultural Research*. 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

Navarro, D. J. (2015). *Learning Statistics with R: A tutorial for psychology students and other beginners* (Versión 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*. https://CRAN.R-project.org/package=vegan

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

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

Vásquez-Castillo, W., Racines-Oliva, M., Moncayo, P., Viera, W., & Seraquive, M. (2019). Calidad del fruto y pérdidas postcosecha de banano orgánico (Musa acuminata) en el Ecuador. *Enfoque UTE*, *10*(4), 57-66. https://doi.org/10.29019/enfoque.v10n4.545

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

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

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

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

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

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