R para el Análisis Estadístico de Datos Agrícolas

P. Agr. Ludwing Isaí Marroquín Jiménez

2 jul 2025

Tabla de contenidos

Introducción

En el ámbito de la investigación agronómica, la estadística se presenta como una herramienta esencial para la transformación de datos en conocimiento aplicable. Este manual, titulado 'R para el Análisis Estadístico de Datos Agrícolas', ha sido concebido como una introducción accesible y práctica al análisis estadístico moderno, con un enfoque particular en el lenguaje R.

Tradicionalmente, la estadística ha proporcionado los cimientos para la toma de decisiones informadas en la agricultura. Sin embargo, la creciente disponibilidad de datos y la necesidad de análisis más sofisticados exigen un enfoque actualizado y eficiente. R, un lenguaje de programación y entorno de software ampliamente adoptado en la ciencia de datos y la estadística aplicada, ofrece la flexibilidad y el poder necesarios para abordar estos desafíos (Ihaka & Gentleman, 1996; R Core Team, 2023).

Este manual está diseñado para guiar al lector a través de un proceso gradual y comprensible, desde los conceptos estadísticos fundamentales hasta las técnicas esenciales para el análisis estadístico de datos agrícolas. Se abordan temas clave como aspectos introductorios, clasificación de variables, notación sumatoria, medidas de tendencia central y dispersión (tanto para datos agrupados como no agrupados), introducción a probabilidades, distribuciones de probabilidad discretas, la distribución normal, intervalos de confianza, pruebas de hipótesis, y regresión lineal y correlación.

Cada capítulo combina la teoría con ejemplos prácticos y estudios de caso relevantes, facilitando la comprensión y la aplicación de los métodos en situaciones reales. El propósito central es proporcionar una base sólida que permita a los profesionales y estudiantes de agronomía utilizar R de manera efectiva en su trabajo diario. No se requiere experiencia previa en programación o estadística; el manual está estructurado para ser accesible a todos, independientemente de su nivel de conocimientos iniciales.

Organización del manual

El presente manual se estructura de manera progresiva, comenzando con los fundamentos esenciales y avanzando hacia técnicas estadísticas aplicadas, con el objetivo de facilitar una comprensión integral del análisis estadístico de datos agrícolas utilizando R. Cada capítulo incluye explicaciones detalladas, ejemplos prácticos y código R reproducible, diseñados para consolidar el aprendizaje y fomentar la aplicación efectiva de los conceptos.

A continuación, se presenta una tabla que resume la organización del manual, detallando los temas cubiertos en cada capítulo:

Capítulo	Título	Descripción
1	Aspectos introductorios	Introducción a la estadística, su importancia en agronomía y primeros pasos en R y RStudio.
2	Clasificación de variables	Tipos de variables, escalas de medición y ejemplos aplicados al ámbito agrícola.
3	Notación sumatoria	Fundamentos y aplicaciones de la notación sumatoria en el cálculo de estadísticos descriptivos.
4	Medidas de tendencia central y dispersión (datos no agrupados)	Cálculo e interpretación de media, mediana, moda, rango, varianza y desviación estándar para datos no agrupados.
5	Medidas de tendencia central y dispersión (datos agrupados)	Aplicación de medidas de tendencia central y dispersión en tablas de frecuencia utilizando R.
6	Introducción a probabilidades	Conceptos básicos de probabilidad, espacio muestral, eventos y reglas de probabilidad.
7	Distribuciones de probabilidad discretas	Estudio de las distribuciones binomial y Poisson, cálculo de probabilidades y representación gráfica en R.
8	Distribución normal	Propiedades y aplicaciones de la distribución normal, cálculo de probabilidades y gráficos en R.
9	Intervalos de confianza	Construcción e interpretación de intervalos de confianza para medias y proporciones con apoyo de R.
10	Pruebas de hipótesis	Formulación y evaluación de hipótesis estadísticas, cálculo de estadísticos de prueba y toma de decisiones.

Capítulo	Título	Descripción
11	Regresión lineal y correlación	Ajuste de modelos de regresión lineal, interpretación de coeficientes y análisis de correlación en R.

Cada capítulo está diseñado para ser independiente, permitiendo que los lectores avancen a su propio ritmo y consulten las secciones según sus necesidades. La tabla proporciona una visión general de la estructura del manual, facilitando la navegación y la comprensión de los temas abordados.

Requisitos Previos

El presente manual no exige conocimientos previos en programación ni en estadística. Está orientado a personas que se inician en el análisis estadístico de datos agrícolas, partiendo desde los conceptos más básicos y avanzando de manera progresiva. Cada tema se desarrolla con explicaciones claras y detalladas, acompañadas de ejemplos y ejercicios prácticos.

Para aprovechar al máximo el contenido, se recomienda contar con lo siguiente:

1. **Interés en aprender:** La disposición para explorar el análisis estadístico y el uso de nuevas herramientas facilita el proceso de aprendizaje.
2. **Acceso a una computadora:** Es necesario disponer de un equipo con capacidad para instalar R y RStudio, cuyas instrucciones de instalación y configuración se incluyen en el manual.
3. **Constancia y práctica:** El desarrollo de habilidades en estadística y en el uso de R requiere tiempo y dedicación. Los ejercicios propuestos están diseñados para acompañar y reforzar el aprendizaje.

Con este enfoque, cualquier persona interesada podrá utilizar el manual como guía para iniciarse en el análisis estadístico de datos agrícolas empleando R, sin importar su experiencia previa en el área.

Software y convenciones

La versión en línea de este manual está disponible en https://ludwing-mj.github.io/R-para-el-analisis-estadistico-de-datos-/, y la fuente en español se encuentra alojada en el siguiente repositorio de GitHub https://github.com/Ludwing-MJ/R-para-el-analisis-estadistico-de-datos-. El desarrollo del manual se realizó utilizando Quarto, una herramienta que permite transformar archivos con extensión .qmd en formatos publicables como HTML, PDF y EPUB, facilitando la integración de código, resultados y texto en un solo documento reproducible.

Durante la elaboración del manual se emplearon diversos paquetes del ecosistema de R, entre los que destacan knitr y bookdown, los cuales permiten combinar las ventajas de LaTeX y R para la generación de documentos dinámicos y reproducibles (Xie et al., 2018). Esta integración posibilita que los ejemplos de código y los resultados presentados sean fácilmente replicables por el

A lo largo del manual, se presentan fragmentos de código que pueden ser copiados y ejecutados directamente en la consola de R para obtener los mismos resultados que se muestran en el texto. Los bloques de código se destacan en recuadros similares al siguiente:

```
4 + 6
a <- c(1, 5, 6)
5 * a
1:10
```

Los resultados generados por la ejecución de estos códigos se identifican con el numero uno encerrado entre cochetes (`[1]`) al inicio de cada línea, indicando que corresponden a la salida producida por R. Todo lo que comience con `[1]` representa resultados y no debe ser copiado como parte del código. Por ejemplo, al ejecutar el bloque anterior, se obtendrían los siguientes resultados:

```
[1] 10
```

```
[1]  5 25 30
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

Para garantizar la reproducibilidad y transparencia, se recomienda que el lector utilice versiones actualizadas de R y de los paquetes mencionados. La información sobre el entorno de desarrollo y las versiones de los paquetes utilizados en la construcción de este manual puede consultarse ejecutando el siguiente comando en R:

```
devtools::session_info()
```

```
Warning in system2("quarto", "-V", stdout = TRUE, env = paste0("TMPDIR=", : el
comando ejecutado '"quarto"
TMPDIR=C:/Users/FAUSAC/AppData/Local/Temp/Rtmpsl98Ok/file141079f65976 -V' tiene
el estatus 1
```

```
- Session info -------------------------------------------------------------
 setting  value
 version  R version 4.4.3 (2025-02-28 ucrt)
 os       Windows 11 x64 (build 26100)
 system   x86_64, mingw32
 ui       RTerm
 language (EN)
```

```
 collate  Spanish_Guatemala.utf8
 ctype    Spanish_Guatemala.utf8
 tz       America/Guatemala
 date     2025-07-01
 pandoc   3.4 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/ (via rmarkdo
 quarto   NA @ C:\\PROGRA~1\\Quarto\\bin\\quarto.exe

- Packages -------------------------------------------------------------------
 package     * version date (UTC) lib source
 cachem        1.1.0   2024-05-16 [1] CRAN (R 4.4.3)
 cli           3.6.5   2025-04-23 [1] CRAN (R 4.4.3)
 devtools      2.4.5   2022-10-11 [1] CRAN (R 4.4.3)
 digest        0.6.37  2024-08-19 [1] CRAN (R 4.4.3)
 ellipsis      0.3.2   2021-04-29 [1] CRAN (R 4.4.3)
 evaluate      1.0.3   2025-01-10 [1] CRAN (R 4.4.3)
 fastmap       1.2.0   2024-05-15 [1] CRAN (R 4.4.3)
 fs            1.6.6   2025-04-12 [1] CRAN (R 4.4.3)
 glue          1.8.0   2024-09-30 [1] CRAN (R 4.4.3)
 htmltools     0.5.8.1 2024-04-04 [1] CRAN (R 4.4.3)
 htmlwidgets   1.6.4   2023-12-06 [1] CRAN (R 4.4.3)
 httpuv        1.6.16  2025-04-16 [1] CRAN (R 4.4.3)
 jsonlite      2.0.0   2025-03-27 [1] CRAN (R 4.4.3)
 knitr         1.50    2025-03-16 [1] CRAN (R 4.4.3)
 later         1.4.2   2025-04-08 [1] CRAN (R 4.4.3)
 lifecycle     1.0.4   2023-11-07 [1] CRAN (R 4.4.3)
 magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.4.3)
 memoise       2.0.1   2021-11-26 [1] CRAN (R 4.4.3)
 mime          0.13    2025-03-17 [1] CRAN (R 4.4.3)
 miniUI        0.1.2   2025-04-17 [1] CRAN (R 4.4.3)
 pkgbuild      1.4.7   2025-03-24 [1] CRAN (R 4.4.3)
 pkgload       1.4.0   2024-06-28 [1] CRAN (R 4.4.3)
 profvis       0.4.0   2024-09-20 [1] CRAN (R 4.4.3)
 promises      1.3.2   2024-11-28 [1] CRAN (R 4.4.3)
 purrr         1.0.4   2025-02-05 [1] CRAN (R 4.4.3)
 R6            2.6.1   2025-02-15 [1] CRAN (R 4.4.3)
 Rcpp          1.0.14  2025-01-12 [1] CRAN (R 4.4.3)
 remotes       2.5.0   2024-03-17 [1] CRAN (R 4.4.3)
 rlang         1.1.6   2025-04-11 [1] CRAN (R 4.4.3)
 rmarkdown     2.29    2024-11-04 [1] CRAN (R 4.4.3)
 rstudioapi    0.17.1  2024-10-22 [1] CRAN (R 4.4.3)
 sessioninfo   1.2.3   2025-02-05 [1] CRAN (R 4.4.3)
 shiny         1.10.0  2024-12-14 [1] CRAN (R 4.4.3)
 urlchecker    1.0.1   2021-11-30 [1] CRAN (R 4.4.3)
 usethis       3.1.0   2024-11-26 [1] CRAN (R 4.4.3)
 vctrs         0.6.5   2023-12-01 [1] CRAN (R 4.4.3)
 xfun          0.52    2025-04-02 [1] CRAN (R 4.4.3)
 xtable        1.8-4   2019-04-21 [1] CRAN (R 4.4.3)

 [1] C:/Users/FAUSAC/AppData/Local/R/win-library/4.4
```

```
 [2] C:/Program Files/R/R-4.4.3/library

--------------------------------------------------------------------------------
```

1 Instalación y configuración

Antes de iniciar el trabajo con análisis estadístico en R, es fundamental realizar la instalación y configuración tanto del lenguaje R como del entorno de desarrollo RStudio. R es un lenguaje de programación y entorno computacional ampliamente utilizado en el análisis estadístico, la visualización de datos y la investigación reproducible (Ihaka & Gentleman, 1996). Por su parte, RStudio constituye un Entorno de Desarrollo Integrado (IDE) diseñado específicamente para facilitar el uso de R, proporcionando una interfaz intuitiva y herramientas avanzadas que optimizan el flujo de trabajo (Xie et al., 2018). Esta sección describe los pasos necesarios para descargar, instalar y configurar ambos programas, asegurando un entorno de trabajo funcional y eficiente.

1.1 Descarga de R y RStudio

Para utilizar R y RStudio, es necesario descargar ambos programas desde sus sitios oficiales. R proporciona el núcleo del lenguaje y las herramientas computacionales fundamentales, mientras que RStudio actúa como una interfaz que simplifica el uso de R y mejora la experiencia del usuario, integrando funciones para la gestión de proyectos, edición de scripts y visualización de resultados (Xie et al., 2018).

1.1.1 Descarga de R

Se recomienda descargar una versión estable de R para evitar posibles incompatibilidades con paquetes que aún no han sido actualizados para las versiones más recientes. Por ejemplo, la versión R 4.4.3 es reconocida por su estabilidad y amplio soporte dentro de la comunidad de usuarios (R Core Team, 2023). El repositorio oficial de R se encuentra disponible en https://cran.r-project.org/bin/windows/base/old/, donde es posible acceder a todas las versiones publicadas. Para descargar una versión específica, se debe seleccionar el nombre de la versión deseada y hacer clic en el archivo con terminación -**win.exe**, lo que iniciará la descarga del instalador correspondiente (R Core Team, 2023).

1.1.2 Descarga de RStudio

La descarga de RStudio se realiza desde su página oficial, donde se encuentra disponible la versión más reciente para los principales sistemas operativos. Para usuarios de Windows, se debe seleccionar la opción "Download RStudio Desktop for Windows", mientras que para quienes utilizan macOS o Linux, la misma página ofrece las versiones correspondientes para estos sistemas (Xie et al., 2018). Es importante asegurarse de descargar la versión adecuada según el sistema operativo del equipo para garantizar la compatibilidad y el correcto funcionamiento del entorno.

1.2 Instalación de R y RStudio

La instalación de R y RStudio debe realizarse siguiendo un orden específico para evitar conflictos y asegurar que ambos programas funcionen correctamente. A continuación, se describen los pasos detallados para cada uno:

1. **Instalación de R:** Una vez descargado el instalador de R, se debe ejecutar el archivo .exe y seguir las instrucciones proporcionadas por el asistente de instalación. En la mayoría de los casos, es suficiente con aceptar las configuraciones predeterminadas, a menos que se requiera una configuración personalizada para necesidades específicas del usuario o del proyecto (R Core Team, 2023).
2. **Instalación de RStudio:** Después de instalar R, se procede a ejecutar el instalador de RStudio previamente descargado. Al igual que en el caso de R, se pueden aceptar las opciones predeterminadas durante la instalación. Es relevante destacar que RStudio permite gestionar múltiples versiones de R en un mismo dispositivo, lo que resulta especialmente útil para trabajar en proyectos que requieren versiones específicas del lenguaje. Esta selección puede realizarse desde la configuración de RStudio, facilitando así la administración de entornos de trabajo diferenciados (Xie et al., 2018; R Core Team, 2023).

1.3 Configuración inicial de RStudio

Tras completar la instalación de R y RStudio, es recomendable realizar una configuración inicial que permita personalizar el entorno de trabajo, mejorar la organización y facilitar el desarrollo de análisis estadísticos. Estas configuraciones contribuyen a optimizar la experiencia del usuario y a establecer un flujo de trabajo más eficiente y productivo (Xie et al., 2018). A continuación, se describen los pasos esenciales para configurar RStudio de manera adecuada.

1.3.1 Seleccionar la versión de R

RStudio permite elegir la versión de R que se utilizará, lo cual es especialmente útil si se tienen múltiples versiones instaladas en el mismo dispositivo. Esta funcionalidad garantiza la compatibilidad con proyectos que requieren versiones específicas del lenguaje (R Core Team, 2023). Para seleccionar la versión de R en RStudio, se deben seguir estos pasos:

1. Ir al menú **Tools** y seleccionar **Global Options**.
2. En la ventana emergente, dirigirse a la pestaña **General**.
3. En el apartado **R version**, elegir la versión deseada de R.

1.3.2 Configurar la apariencia de RStudio

RStudio ofrece opciones de personalización para adaptar su apariencia a las preferencias del usuario, lo que puede mejorar la experiencia de trabajo y reducir la fatiga visual durante sesiones prolongadas (Xie et al., 2018). Para cambiar el tema de la interfaz y ajustar la fuente, se deben seguir los siguientes pasos:

1. Acceder al menú **Tools** y seleccionar **Global Options**.
2. En la ventana emergente, ir a la pestaña **Appearance**.
3. Elegir el tema preferido, ya sea claro u oscuro (por ejemplo, el tema Cobalt para reducir la fatiga visual).
4. Ajustar el tamaño y el tipo de fuente según las preferencias personales.

1.3.3 Configurar el panel de trabajo

La interfaz de RStudio está organizada en cuatro paneles principales: editor de scripts, consola, entorno/archivos y gráficos/ayuda. Estos paneles pueden reorganizarse para optimizar el flujo de trabajo. Para modificar la disposición de los paneles, se deben seguir estos pasos:

1. Ir al menú **Tools** y seleccionar **Global Options**.
2. Acceder a la sección **Pane Layout**.
3. Ajustar la ubicación de los paneles según las necesidades, por ejemplo, colocando el editor de scripts en la parte superior izquierda y la consola en la parte inferior.
4. Guardar los cambios para aplicar la nueva disposición.

1.3.4 Habilitar el número de líneas en el editor de scripts

La numeración de líneas en el editor de scripts facilita la navegación y depuración del código. Para habilitar esta opción, se deben seguir los siguientes pasos:

1. Acceder al menú **Tools** y seleccionar **Global Options**.
2. Ir a la pestaña **Code** y luego a **Display**.
3. Marcar la casilla **Show line numbers** para activar la numeración de líneas.

1.4 Organización de proyectos

La organización adecuada de proyectos en RStudio es esencial para establecer un flujo de trabajo eficiente, reproducible y estructurado. Una gestión ordenada de archivos y scripts no solo facilita el desarrollo de los análisis, sino que también mejora la colaboración y la reproducibilidad de los resultados (Xie et al., 2018).

1.4.1 Crear un proyecto en RStudio

Para organizar los archivos, datos y scripts de un análisis específico, RStudio permite crear proyectos siguiendo estos pasos:

1. En la barra de menú, seleccionar **File > New Project**.
2. Elegir una de las siguientes opciones:

   New Directory: para crear un proyecto desde cero en una nueva carpeta.

   Existing Directory: para convertir una carpeta existente en un proyecto de RStudio.

   Version Control: para clonar un repositorio de Git y trabajar en un proyecto con control de versiones.
3. Configurar el nombre y la ubicación del proyecto según las necesidades del análisis.
4. Hacer clic en **Create Project** para finalizar la configuración.

El uso de proyectos en RStudio permite mantener una estructura clara y organizada, facilitando la gestión de los recursos necesarios para el análisis y promoviendo la reproducibilidad (Xie et al., 2018).

1.5 Directorio de trabajo o *Working Directory*

El directorio de trabajo es la carpeta donde R buscará los archivos y guardará los resultados generados durante el análisis. Para establecerlo manualmente, se puede utilizar la función `setwd()`, como se muestra a continuación:

```
# Establecer directorio de trabajo
setwd("ruta/del/directorio")
```

Sin embargo, al trabajar con proyectos en RStudio, el directorio de trabajo se configura automáticamente al abrir el archivo del proyecto, lo que elimina la necesidad de establecerlo manualmente y reduce errores relacionados con rutas incorrectas (R Core Team, 2023).

1.5.1 Uso de archivos .Rproj

El archivo `.Rproj` es el elemento central de cada proyecto en RStudio. Este archivo almacena las configuraciones específicas del proyecto, como el directorio de trabajo, las opciones de visualización y otros ajustes personalizados. Al abrir un archivo `.Rproj`, se carga automáticamente el entorno de trabajo asociado, lo que facilita la continuidad y la gestión del análisis (Xie et al., 2018).

Capítulo I

Aspectos Introductorios

2 Aspectos introductorios

2.1 Definición de estadística

La estadística es una ciencia derivada de la matemática que se ocupa de la extracción de información contenida en datos provenientes de muestras y de su uso para hacer inferencias acerca de la población de donde fueron extraídos estos datos. Además, la estadística estudia los métodos científicos para recolectar, organizar, resumir y analizar datos, así como para extraer conclusiones válidas y tomar decisiones razonables basadas en tal análisis (López & González, 2018).

El término "estadística" tiene su origen en la palabra alemana Statistik, utilizada por el profesor Gottfried Achenwall en el siglo XVIII, y proviene del término latino status, que significa estado o situación. Históricamente, la estadística ha estado vinculada a la recolección de datos por parte de los gobiernos, especialmente en relación con información demográfica, como los censos (López & González, 2018).

2.2 División de la estadística

La estadística, como disciplina científica, se divide tradicionalmente en dos grandes ramas: la estadística descriptiva y la estadística inferencial. Esta división responde tanto a los objetivos como a los métodos empleados en el análisis de datos.

La **estadística descriptiva** comprende el conjunto de métodos y técnicas destinados a recolectar, organizar, presentar y resumir datos de manera cuantitativa o gráfica. Su propósito principal es describir las características principales de un conjunto de datos, facilitando su interpretación y permitiendo identificar patrones, tendencias o comportamientos dentro de la información analizada (López & González, 2018; Montgomery & Runger, 2018).

Por otro lado, la **estadística inferencial** se ocupa de realizar generalizaciones, predicciones o inferencias sobre una población a partir de la información obtenida en una muestra representativa. Esta rama utiliza herramientas de probabilidad para estimar parámetros poblacionales, probar hipótesis y cuantificar el grado de incertidumbre asociado a las conclusiones (López & González, 2018; Walpole et al., 2012).

Además de estas dos ramas fundamentales, la estadística moderna reconoce otras áreas especializadas que amplían su campo de aplicación:

1. La **estadística paramétrica** se centra en el análisis de datos bajo el supuesto de que estos provienen de poblaciones que siguen distribuciones conocidas, generalmente la normal. Permite realizar estimaciones y pruebas de hipótesis sobre parámetros poblacionales (López & González, 2018; Montgomery & Runger, 2018).

2. La **estadística no paramétrica** incluye métodos que no requieren suposiciones estrictas sobre la distribución de los datos, siendo especialmente útil cuando no se puede asumir normalidad o cuando los datos son de nivel ordinal o nominal (Conover, 1999).
3. La **geoestadística** aplica técnicas estadísticas al análisis de variables distribuidas en el espacio o el tiempo, siendo fundamental en disciplinas como la agronomía, la geografía y la gestión ambiental (López & González, 2018; Webster & Oliver, 2007).
4. La **inferencia bayesiana** utiliza el teorema de Bayes para actualizar la probabilidad de una hipótesis a medida que se dispone de nueva información, incorporando el conocimiento previo en el análisis estadístico (Gelman et al., 2013).
5. La **estadística multivariada** estudia simultáneamente múltiples variables, permitiendo analizar relaciones complejas y estructuras latentes en grandes conjuntos de datos (Johnson & Wichern, 2014).

2.3 Definiciones importantes

En el estudio de la estadística, es fundamental comprender ciertos conceptos básicos que constituyen la base para el análisis y la interpretación de datos. Entre los más relevantes se encuentran: individuo o unidad estadística, población, muestra, parámetro, estimador e indicador.

2.3.1 Individuo o unidad estadística

El individuo o unidad estadística es el elemento básico sobre el cual se realiza la observación o medición en un estudio estadístico. Puede tratarse de una persona, animal, planta, objeto o cualquier entidad sobre la que se recolectan datos. Por ejemplo, en un estudio agronómico, una unidad estadística puede ser una planta de maíz, una parcela de terreno o un saco de fertilizante (López & González, 2018; Walpole et al., 2012).

2.3.2 Población

La población se define como el conjunto total de unidades estadísticas que comparten al menos una característica observable y relevante para el estudio. El análisis de toda la población se denomina censo. En agronomía, la población puede estar formada por todos los árboles de una plantación, todos los lotes de un cultivo o todos los animales de una granja (López & González, 2018; Montgomery & Runger, 2018).

2.3.3 Muestra

La muestra es un subconjunto representativo de la población, seleccionado con el propósito de inferir características o parámetros de la población total. El muestreo permite realizar estudios más eficientes y menos costosos que el censo, siempre que la muestra sea seleccionada adecuadamente (López & González, 2018; Walpole et al., 2012).

2.3.4 Parámetro

Un parámetro es un valor numérico que resume o describe una característica de la población, como la media, la varianza o la proporción. Los parámetros son generalmente desconocidos y se estiman a partir de los datos muestrales (Montgomery & Runger, 2018).

2.3.5 Estimador

El estimador es una función o estadístico calculado a partir de los datos de la muestra, utilizado para aproximar el valor de un parámetro poblacional. Por ejemplo, la media muestral es un estimador de la media poblacional (Walpole et al., 2012).

2.3.6 Indicador

Un indicador es un elemento extraído de la realidad que permite cuantificar características medibles de un fenómeno o sistema. Su principal función es servir como base para la construcción de índices relativos, facilitando la comparación y el análisis entre diferentes situaciones o periodos. Los indicadores transforman observaciones concretas en valores numéricos que reflejan la presencia, magnitud o evolución de una característica específica. De este modo, permiten señalar o evidenciar que una variable está ocurriendo y proporcionan información útil para la toma de decisiones. La objetividad de los indicadores puede variar según la naturaleza de la característica que representan, siendo algunos más fácilmente cuantificables que otros (López & González, 2018).

Capítulo II

Clasificación de variables

3 Clasificación de Variables

3.1 Definición de Variable

Una variable en estadística se define como aquello que se observa o mide sobre las unidades estadísticas (López & González, 2018). Constituye una característica que varía de un individuo a otro dentro de la población o muestra bajo estudio. Las variables representan los atributos o propiedades que pueden ser medidos, observados o categorizados en las unidades de análisis.

Desde el punto de vista de la notación estadística, las variables se representan mediante letras mayúsculas del alfabeto (X, Y, Z), mientras que los valores específicos que estas asumen se denotan con letras minúsculas correspondientes (x, y, z) (López & González, 2018). Esta distinción notacional permite diferenciar claramente entre el concepto abstracto de la variable y sus manifestaciones concretas en los datos.

3.2 Tipos de Variables

Dependiendo de su naturaleza, las variables estadísticas se clasifican en dos categorías principales que determinan tanto los métodos de análisis apropiados como las técnicas de presentación de datos más adecuadas (López & González, 2018).

3.2.1 Variables Cuantitativas

Las variables cuantitativas son aquellas que expresan cantidades y cuyos resultados son de naturaleza numérica (López & González, 2018). Estas variables permiten realizar operaciones matemáticas y estadísticas avanzadas debido a su carácter numérico. Se subdividen en dos tipos fundamentales:

3.2.1.1 Variables Cuantitativas Discretas

También denominadas variables de conteo, son aquellas que no aceptan valores decimales y típicamente resultan de procesos de enumeración (López & González, 2018). Ejemplos relevantes en el contexto agronómico incluyen: número de plantas de café por metro cuadrado, cantidad de áfidos por planta, número de brotes por planta, número de racimos de banano por hectárea, y número de ausencias de un trabajador por mes.

Matemáticamente, estas variables pueden representarse mediante conjuntos discretos. Por ejemplo, si X representa el número de árboles con cáncer en una muestra de 10 árboles, entonces X {0, 1, 2, 3, ..., 9, 10} (López & González, 2018).

3.2.1.2 Variables Cuantitativas Continuas

Este tipo de variables pueden asumir cualquier valor dentro de un rango determinado, incluyendo valores decimales, y resultan típicamente de procesos de medición (López & González, 2018). En el ámbito agronómico, ejemplos representativos incluyen: altura de plantas, peso de semillas, temperatura de almacenamiento, diámetro de árboles, caudal de ríos, y precipitación pluvial.

La representación matemática de estas variables utiliza intervalos continuos. Por ejemplo, si D representa el diámetro de árboles de *Pinus maximinoii* en una plantación, entonces D [10, 50] centímetros (López & González, 2018).

3.2.2 Variables Cualitativas

Las variables cualitativas presentan como posibles resultados una cualidad o atributo del individuo investigado (López & González, 2018). Las posibles cualidades que puede presentar una variable cualitativa se denominan modalidades, categorías o atributos de la variable.

Según el número de categorías, estas variables se clasifican en:

1. **Dicotómicas**: presentan únicamente dos modalidades, como sexo (masculino, femenino) o resultado de evaluación (aprobado, reprobado)
2. **Politómicas**: presentan más de dos categorías, como estado civil, color de ojos, lugar de origen, o susceptibilidad de plantas a enfermedades (López & González, 2018)

3.3 Mapa mental de la clasificación de variables estadísticas

A continuación, se presenta un mapa mental que sintetiza la información esencial sobre la clasificación de variables estadísticas, facilitando la comprensión visual de los conceptos abordados. Para explorar el mapa mental de manera interactiva y detallada, se recomienda acceder al siguiente enlace: https://ma-variables.vercel.app/.

3.4 Escalas de Medición

Las escalas de medición constituyen un sistema de clasificación que determina el nivel de información que proporcionan las variables y, consecuentemente, los tipos de análisis estadísticos que pueden aplicarse (López & González, 2018).

3.4.1 Escalas para Información Cualitativa

3.4.1.1 Escala Nominal

La escala nominal representa el nivel más básico de medición y consiste en asignar nombres o etiquetas a las observaciones para distinguir diferentes agrupamientos (López & González, 2018). Cuando se emplean números en esta escala, estos tienen únicamente carácter simbólico, no numérico.

Ejemplos en el contexto agronómico incluyen: especies arbóreas presentes en una cuenca, tipos de uso del suelo (agrícola, forestal, pecuario), y municipio de procedencia de estudiantes (López & González, 2018).

3.4.1.2 Escala Ordinal

En este nivel de medición, las unidades mantienen una relación jerárquica que permite establecer ordenamientos del tipo "mayor que" o "menor que" (López & González, 2018). Las categorías poseen un orden lógico, pero las distancias entre ellas no son necesariamente iguales.

Ejemplos representativos incluyen: nivel de estudios (primaria, secundaria, diversificado, universitaria), grado de aceptación de productos (buena, regular, mala), y escalas de severidad de enfermedades en plantas. La escala de Likert constituye un ejemplo paradigmático de medición ordinal, típicamente empleando cinco niveles de respuesta desde "totalmente en desacuerdo" hasta "totalmente de acuerdo" (López & González, 2018).

3.4.2 Escalas para Información Cuantitativa

3.4.2.1 Escala de Intervalo

Esta escala proporciona información más precisa y permite mediciones sofisticadas al informar tanto sobre el orden de los objetos como sobre las distancias numéricas entre ellos (López & González, 2018). Los intervalos de igual tamaño en la escala representan diferencias equivalentes, independientemente de su ubicación en la escala.

Sin embargo, la escala de intervalo carece de un punto cero absoluto, siendo este arbitrario y no representando la ausencia total de la característica medida (López & González, 2018). Ejemplos incluyen: temperatura, coordenadas geográficas, y resultados de exámenes académicos.

3.4.2.2 Escala de Razón

Los atributos cuantitativos organizados en escala de razón poseen tanto intervalos significativos como un punto cero real que indica ausencia absoluta del valor medido (López & González, 2018). Esta escala permite realizar todas las operaciones matemáticas, incluyendo la determinación de razones o proporciones entre medidas.

Variables agronómicas medidas en esta escala incluyen: peso, longitud, diámetro, volumen, estatura, y densidad. Por ejemplo, un individuo de 190 cm es exactamente dos veces más alto que uno de 95 cm, relación que se mantiene independientemente de la unidad de medida empleada (López & González, 2018).

3.5 Mapa mental de las escalas de medición

A continuación, se presenta un mapa mental que sintetiza la información esencial sobre las escalas de medición de variables estadísticas, facilitando la comprensión visual de los conceptos abordados. Para explorar el mapa mental de manera interactiva y detallada, se recomienda acceder al siguiente enlace: https://ma-escalas.vercel.app/.

Capítulo III

Notación sumatoria

4 Introducción a la Notación Sumatoria

La notación sumatoria, representada por la letra griega sigma mayúscula (Σ), es una herramienta matemática que permite expresar de manera concisa la suma de una serie de términos. En lugar de escribir largas sumas de forma explícita, la notación sumatoria ofrece una manera compacta y generalizable de representar estas operaciones, facilitando el análisis y la manipulación de datos en diversos campos, incluyendo la agronomía (López & González, 2018).

4.1 Elementos de la Notación Sumatoria

La notación sumatoria se compone de los siguientes elementos clave:

1. **Índice de Sumación:** Es una variable, comúnmente denotada por i, j o k, que actúa como un contador, indicando el término específico que se está sumando en cada iteración.
2. **Límite Inferior:** Es el valor inicial del índice de sumación, situado debajo del símbolo Σ. Indica el punto de partida de la suma.
3. **Límite Superior:** Es el valor final del índice de sumación, situado encima del símbolo Σ. Indica el punto de finalización de la suma.
4. **Sumando:** Es la expresión matemática que se va a sumar, y generalmente depende del índice de sumación. Esta expresión define cómo se calcula cada término de la suma.

La expresión general de la notación sumatoria se presenta de la siguiente manera:

$$\sum_{i=m}^{n} xi$$

Donde:

1. i es el índice de sumación.
2. m es el límite inferior.
3. n es el límite superior.
4. x_i es el sumando.

Por ejemplo, la suma de los primeros nnn números se expresa como:

$$\sum_{i=m}^{n} i = 1 + 2 + 3 + ... + n$$

4.2 Propiedades de la Notación Sumatoria

La notación sumatoria es fundamental en el análisis estadístico, ya que permite expresar de manera compacta la suma de una serie de términos. Sus propiedades facilitan la manipulación algebraica de expresiones y el desarrollo de fórmulas estadísticas (López & González, 2018). A continuación se detallan las propiedades de la notación sumatoria:

4.2.1 Suma de una constante

Cuando se suma una constante k un número n de veces, el resultado es igual al producto de la constante por el número de sumandos:

$$\sum_{i=1}^{n} k = n \cdot k$$

Por ejemplo, si $k = 3$ y $n = 5$:

$$\sum_{i=1}^{5} 3 = 3 + 3 + 3 + 3 + 3 = 5 \times 3 = 15$$

Esta propiedad es útil para simplificar sumas donde el sumando no depende del índice de sumación.

4.2.2 Factor constante

Si cada término de la suma es el producto de una constante k y una variable x_i, la constante puede factorizarse fuera de la sumatoria:

$$\sum_{i=1}^{n} k \cdot x_i = k \sum_{i=1}^{n} x_i$$

Esto permite simplificar cálculos y es especialmente útil en operaciones como el cálculo de medias ponderadas.

4.2.3 Suma de variables

La suma de la suma de dos variables es igual a la suma de las sumas de cada variable por separado:

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Esta propiedad refleja la linealidad de la suma y es fundamental en la manipulación de expresiones estadísticas.

4.2.4 Diferencia de variables

De manera análoga, la suma de la diferencia entre dos variables es igual a la diferencia entre las sumas de cada variable:

$$\sum_{i=1}^{n}(x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i$$

Esta propiedad también se deriva de la linealidad de la suma.

4.2.5 Producto de dos variables

La suma del producto de dos variables se expresa como:

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + ... + x_n y_n$$

Es importante destacar que, en general,

$$\sum_{i=1}^{n} x_i y_i \neq \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Esta distinción es crucial en el cálculo de covarianzas y otros estadísticos.

4.2.6 Suma de cuadrados vs. cuadrado de la suma

Se debe diferenciar entre la suma de los cuadrados y el cuadrado de la suma:

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2$$

La suma de cuadrados es:

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + ... + x_n^2$$

El cuadrado de la suma es:

$$\left(\sum_{i=1}^{n} x_i \right)^2 = (x_1 + x_2 + ... + x_n)^2$$

Ambas expresiones son diferentes y tienen aplicaciones distintas en estadística, por ejemplo, en el cálculo de la varianza.

4.2.7 Constante multiplicada por el cuadrado

Si se multiplica una constante k por el cuadrado de cada término y se suman los resultados, se puede factorizar la constante fuera de la sumatoria:

$$\sum_{i=1}^{n} k \cdot x_i^2 = k \sum_{i=1}^{n} x_i^2$$

Esta propiedad es útil en el desarrollo de fórmulas para momentos y otras medidas estadísticas.

4.3 Aplicación de la notación sumatoria en estadística

La notación sumatoria es una herramienta esencial en estadística, permitiendo expresar de manera concisa y eficiente operaciones de suma que son fundamentales en el cálculo de diversos estadísticos. En el contexto de la estadística y su aplicación en agronomía, la notación sumatoria facilita la comprensión y el cálculo de estadísticos clave. Estos estadísticos son fundamentales para resumir y analizar datos relacionados con el rendimiento de cultivos, características de plantas y otros parámetros relevantes en la investigación y la práctica agrícola (López & González, 2018). A continuación, se presentan ejemplos de cómo la notación sumatoria se aplica en el cálculo de estos estadísticos.

4.3.1 Media Aritmética (Promedio)

La media aritmética es el valor representativo de un conjunto de datos y se utiliza para resumir el rendimiento promedio de cultivos, alturas de plantas u otras variables de interés en agronomía (López & González, 2018).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

4.3.2 Varianza Muestral

La varianza muestral cuantifica la dispersión de los datos respecto a la media, permitiendo evaluar la uniformidad de características como el peso de frutos o el rendimiento entre parcelas (López & González, 2018).

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

4.3.3 Desviación Estándar Muestral

La desviación estándar, al ser la raíz cuadrada de la varianza, expresa la dispersión de los datos en las mismas unidades que la variable original, facilitando la interpretación de la variabilidad en experimentos agrícolas (López & González, 2018).

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

4.3.4 Covarianza Muestral

La covarianza permite analizar la relación lineal entre dos variables, como la asociación entre la cantidad de fertilizante aplicado y el rendimiento del cultivo, siendo fundamental en estudios de correlación y regresión (López & González, 2018).

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

4.3.5 Coeficiente de Correlación de Pearson

El coeficiente de correlación de Pearson mide la fuerza y dirección de la relación lineal entre dos variables, lo que resulta útil para evaluar la asociación entre factores ambientales y respuestas agronómicas (López & González, 2018).

$$r = \frac{Cov(x, y)}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4.3.6 Ecuación de la Recta de Regresión Lineal Simple

La regresión lineal simple modela la relación entre una variable dependiente y una independiente, permitiendo predecir valores y analizar el efecto de factores como la fertilización sobre el rendimiento (López & González, 2018).

$$y = \beta_0 + \beta_1 x$$

4.3.7 Estimador del Intercepto ( )

El intercepto de la recta de regresión representa el valor esperado de la variable dependiente cuando la independiente es cero, lo que puede interpretarse como el rendimiento base en ausencia de tratamiento (López & González, 2018).

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

4.3.8 Estimador de la Pendiente ( )

La pendiente de la recta de regresión indica el cambio promedio en la variable dependiente por cada unidad de cambio en la independiente, siendo clave para interpretar el efecto de tratamientos en ensayos agrícolas (López & González, 2018).

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

4.3.9 Coeficiente de Determinación (R^2)

El coeficiente de determinación expresa la proporción de la variabilidad de la variable dependiente explicada por el modelo de regresión, siendo un indicador de la calidad del ajuste en estudios agronómicos (López & González, 2018).

$$R^2 = \frac{[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2]}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

5 Ejemplo Aplicado Sumatoria Simple

Para ilustrar el uso de la notación sumatoria, se simulan los rendimientos de diez parcelas experimentales de maíz, expresados en kilogramos por planta. Los valores observados son los siguientes:

i	y_i
1	23
2	25
3	18
4	27
5	22
6	20
7	24
8	26
9	19
10	21

Total de observaciones: $n = 10$

5.1 Cálculo de la Suma Total de los Valores Observados

El primer paso consiste en sumar todos los valores observados, utilizando la notación sumatoria:

$$\sum_{i=1}^{10} y_i = 23 + 25 + 18 + 27 + 22 + 20 + 24 + 26 + 19 + 21 = 225$$

5.2 Cálculo de la Media Muestral

La media muestral se obtiene dividiendo la suma total entre el número de observaciones (n = 10):

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{225}{10} = 22.5$$

Esto significa que, en promedio, cada parcela produjo 22.5 kg por planta.

5.3 Cálculo de la Varianza Muestral

La varianza muestral mide la dispersión de los datos respecto a la media. Para calcularla, se sigue el siguiente procedimiento:

1. Se resta la media a cada valor observado (($y_i - \bar{y}$).
2. Se eleva al cuadrado cada una de estas diferencias ($(y_i - \bar{y})^2$).
3. Se suman todos los cuadrados de las diferencias.
4. Finalmente, se divide esta suma entre $n - 1$ (en este caso, 9).

La tabla siguiente resume estos cálculos:

i	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	23	0.5	0.25
2	25	2.5	6.25
3	18	-4.5	20.25
4	27	4.5	20.25
5	22	-0.5	0.25
6	20	-2.5	6.25
7	24	1.5	2.25
8	26	3.5	12.25
9	19	-3.5	12.25
10	21	-1.5	2.25

Sumando la última columna:

$$\sum_{i=1}^{10}(y_i - \bar{y})^2 = 82.5$$

La varianza muestral se calcula así:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{82.5}{9} = 9.17$$

Por lo tanto, la varianza muestral es 9.17 kg^2 por $planta^2$.

6 Sumatorias Dobles y Múltiples

La notación sumatoria puede extenderse para representar sumas sobre dos o más índices, lo que resulta útil en el análisis de datos organizados en tablas o matrices, como ocurre frecuentemente en experimentos agrícolas con varios tratamientos y repeticiones (López & González, 2018). La sumatoria doble se expresa de la siguiente manera:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}$$

En esta expresión, x_{ij} representa el elemento ubicado en la fila i y la columna j de una matriz de datos. El primer índice (i) recorre las filas y el segundo (j) las columnas. Este tipo de sumatoria es fundamental para calcular totales generales, promedios por tratamiento o por repetición, y para el análisis de varianza en diseños experimentales.

Cuando los datos se organizan en más dimensiones (por ejemplo, en matrices, cubos o hipercubos) la notación sumatoria se extiende añadiendo un índice por dimensión (Wackerly et al., 2014).

Sumatoria triple (o múltiple) sobre un arreglo de orden 3:

$\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p} x_{ijk}$

Para sumas finitas, el orden de los signos sigma es intercambiable, debido a la conmutatividad y asociatividad de la adición (Montgomery, 2017). Para los fines del curso solamente se profundizará en las sumatorias dobles.

6.1 Propiedades básicas de las sumatorias dobles

Las sumatorias dobles conservan las propiedades fundamentales de las sumatorias simples, pero su aplicación requiere consideraciones adicionales debido a la presencia de múltiples índices. Estas propiedades son esenciales para la manipulación algebraica de expresiones estadísticas en el análisis de datos multidimensionales (Wackerly et al., 2014).

6.1.1 Propiedad del factor constante

Cuando una constante c multiplica a cada término de una sumatoria doble, esta constante puede factorizarse fuera de ambos signos de sumatoria. Esta propiedad se deriva directamente de la distributividad de la multiplicación sobre la adición (Ross, 2014).

Enunciado formal:

$\sum_{i=1}^{n}\sum_{j=1}^{m} c\, x_{ij} = c\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}$

Demostración:

La demostración se basa en la aplicación repetida de la propiedad del factor constante para sumatorias simples:

$\sum_{i=1}^{n} \sum_{j=1}^{m} c\, x_{ij} = \sum_{i=1}^{n} \left(c \sum_{j=1}^{m} x_{ij} \right) = c \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}$

Aplicación en estadística agrícola:

Esta propiedad resulta fundamental cuando se requiere convertir unidades de medida. Por ejemplo, si los rendimientos están expresados en kg/ha y se desea convertirlos a t/ha, se multiplica por la constante $c = 0.001$:

$\sum_{i=1}^{n} \sum_{j=1}^{m} 0.001 \cdot x_{ij} = 0.001 \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}$

6.1.2 Propiedad de linealidad (aditividad)

La sumatoria doble de una suma de términos es igual a la suma de las sumatorias dobles de cada término por separado. Esta propiedad refleja la linealidad inherente del operador suma y es fundamental en el desarrollo de fórmulas estadísticas complejas (Montgomery, 2017).

Enunciado formal:

$\sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} + y_{ij} \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}$

6.1.3 Propiedad de descomposición en sumas parciales

Esta propiedad establece que una sumatoria doble puede descomponerse en una sumatoria simple de sumatorias simples, lo que facilita el cálculo de totales por filas, columnas o grupos específicos. La conmutatividad de la adición garantiza que el orden de evaluación no afecte el resultado final (Hogg et al., 2019).

Enunciado formal:

$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_{ij} \right) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} x_{ij} \right)$

Interpretación estadística:

1. $\sum_{j=1}^{m} x_{ij}$ representa el total de la fila i
2. $\sum_{i=1}^{n} x_{ij}$ representa el total de la columna j
3. La suma total puede calcularse como la suma de totales por filas o por columnas

6.2 Aplicaciones en Estadística Agrícola

Para ilustrar la aplicación práctica de las sumatorias dobles, se presenta un ejemplo donde se analiza el rendimiento de tres variedades de maíz en tres localidades diferentes, diferenciando entre producción de grano y producción de rastrojo.

6.2.1 Planteamiento del problema

Un investigador evalúa tres variedades de maíz (V1, V2, V3) en tres localidades (L1, L2, L3), registrando tanto la producción de grano como la de rastrojo en kilogramos por parcela. Los datos se organizan en dos matrices de 3×3:

Matriz de producción de grano (G_{ij}):

	L1	L2	L3
V1	4500	4700	4400
V2	4200	4300	4100
V3	4800	4900	4700

Matriz de producción de rastrojo (R_{ij}):

	L1	L2	L3
V1	3000	3200	3100
V2	2800	2900	2700
V3	3500	3600	3400

Donde i representa la variedad ($i = 1, 2, 3$) y j la localidad ($j = 1, 2, 3$).

6.2.2 Análisis paso a paso de la producción de grano

Cálculo de totales por variedad (sumas por filas)

Para cada variedad i , se calcula el total de producción de grano mediante:

Variedad	Notación sumatoria	Cálculo detallado	Total (kg)
V1	$\sum_{j=1}^{n} G_{i1} = G_{.1}$	$4500 + 4700 + 4400$	**13,600**
V2	$\sum_{j=1}^{n} G_{i2} = G_{.2}$	$4200 + 4300 + 4100$	**12,600**
V3	$\sum_{i=1}^{n} G_{i3} = G_{.3}$	$4800 + 4900 + 4700$	**14,400**

Cálculo de totales por localidad (sumas por columnas)

Para cada localidad j, se obtiene el total mediante:

Localidad	Notación sumatoria	Cálculo detallado	Total (kg)
L1	$\sum_{j=1}^{n} G_{1j} = G_{1.}$	$4500 + 4200 + 4800$	**13,500**
L2	$\sum_{j=1}^{n} G_{2j} = G_{2.}$	$4700 + 4300 + 4900$	**13,900**
L3	$\sum_{j=1}^{n} G_{3j} = G_{3.}$	$4400 + 4100 + 4700$	**13,200**

Demostración de la equivalencia de métodos de cálculo para el total de producción de grano

La suma total de producción de grano puede obtenerse mediante tres métodos equivalentes:

Método 1: Suma directa de todos los elementos

$\sum_{i=1}^{3}\sum_{j=1}^{3} G_{ij} = 4500+4700+4400+4200+4300+4100+4800+4900+4700 = 40,600$ kg

Método 2: Suma de totales por variedad

$\sum_{i=1}^{3}\sum_{j=1}^{3} G_{ij} = G_{..} = 13,600 + 12,600 + 14,400 = 40,600$ kg

Método 3: Suma de totales por localidad

$\sum_{i=1}^{3}\sum_{j=1}^{3} G_{ij} = G_{..} = 13,500 + 13,900 + 13,200 = 40,600$ kg

Esta equivalencia demuestra la propiedad de descomposición en sumas parciales y confirma la consistencia de los cálculos.

6.2.3 Aplicación de la propiedad de linealidad: Cálculo de biomasa total para la variedad 3

Para ilustrar la propiedad de linealidad, se calcula la biomasa total de la variedad 3 (V3), definida como la suma de producción de grano y rastrojo.

Datos para la variedad 3:

Localidad	Grano (G_{3j})	Rastrojo (R_{3j})	Biomasa (B_{3j})
L1	4,800	3,500	8,300
L2	4,900	3,600	8,500
L3	4,700	3,400	8,100

Planteamiento con notación sumatoria

La biomasa total de la variedad 3 se expresa como:

$\sum_{j=1}^{3} B_{3j} = \sum_{j=1}^{3}(G_{3j} + R_{3j})$

Aplicación de la propiedad de linealidad

Utilizando la propiedad de linealidad de las sumatorias:

$\sum_{j=1}^{3}(G_{3j} + R_{3j}) = \sum_{j=1}^{3} G_{3j} + \sum_{j=1}^{3} R_{3j}$

Cálculos detallados

1. Suma de producción de grano para V3:

   $\sum_{j=1}^{3} G_{3j} = 4,800 + 4,900 + 4,700 = 14,400$ kg

2. Suma de producción de rastrojo para V3:

   $\sum_{j=1}^{3} R_{3j} = 3,500 + 3,600 + 3,400 = 10,500$ kg

3. Biomasa total para V3:

   $\sum_{j=1}^{3} B_{3j} = \sum_{j=1}^{3} G_{3j} + \sum_{j=1}^{3} R_{3j} = 14,400 + 10,500 = 24,900$ kg

4. Verificación mediante suma directa

   $\sum_{j=1}^{3} B_{3j} = 8,300 + 8,500 + 8,100 = 24,900 \text{ kg}$

La coincidencia de resultados confirma la validez de la propiedad de linealidad.

Capítulo IV

Estadística descriptiva para datos sin agrupar

7 Estadística descriptiva para datos sin agrupar

En el análisis estadístico, la descripción y el resumen de conjuntos de datos constituyen pasos fundamentales para la comprensión de fenómenos en ciencias aplicadas, como la agronomía. Las medidas de tendencia central, dispersión y posición relativa permiten sintetizar la información, identificar patrones y tomar decisiones informadas en la gestión de recursos agrícolas (López & González, 2018; Montgomery, 2017). Estas herramientas facilitan la interpretación de datos experimentales y la comparación entre diferentes tratamientos o condiciones de cultivo.

Para ejemplificar el desarrollo de este tema, se utilizará la siguiente base de datos, correspondiente al rendimiento de maíz (en toneladas por hectárea) en ocho parcelas de diferente tamaño:

Parcela	Superficie (ha)	Rendimiento (t ha ^{1})
P	1.5	6.2
P	2.0	5.8
P	1.0	6.5
P	1.5	6.0
P	1.0	6.3
P	2.0	5.9
P	1.0	6.4
P	1.0	6.1

7.1 Medidas de Tendencia Central

7.1.1 Media aritmética

La media aritmética representa el valor promedio de un conjunto de datos y constituye la medida de tendencia central más utilizada en estadística descriptiva (López & González, 2018). Se define como la suma de todos los valores dividida entre el número total de observaciones.

Fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde x_i representa cada valor individual y n es el número total de observaciones.

Cálculo con los datos de ejemplo:

Datos ordenados: 5.8, 5.9, 6.0, 6.1, 6.2, 6.3, 6.4, 6.5

$$\bar{x} = \frac{6.2 + 5.8 + 6.5 + 6.0 + 6.3 + 5.9 + 6.4 + 6.1}{8} = \frac{49.2}{8} = 6.150 \text{ t ha}^{-1}$$

La media aritmética es útil para describir el valor central de un conjunto de datos homogéneo, aunque presenta sensibilidad a valores extremos (Montgomery, 2017).

7.1.2 Mediana

La mediana es el valor que divide un conjunto de datos ordenados en dos partes iguales, de manera que el 50% de las observaciones se encuentran por debajo y el 50% por encima de este valor (Steel & Torrie, 1980). Esta medida es especialmente útil cuando los datos presentan distribuciones asimétricas o contienen valores atípicos.

Procedimiento de cálculo:

Para un conjunto de datos con n observaciones ordenadas:

1. Si n es impar:

$$\text{Mediana} = x_{(n+1)/2}$$

2. Si n es par:

$$\text{Mediana} = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

Cálculo con los datos de ejemplo:

Como $n = 8$ (par), la mediana se calcula:

$$\text{Mediana} = \frac{x_4 + x_5}{2} = \frac{6.1 + 6.2}{2} = 6.150 \text{ t ha}^{-1}$$

La mediana es robusta ante valores atípicos y proporciona una medida de tendencia central más estable que la media aritmética en presencia de datos extremos (Anderson et al., 2018).

7.1.3 Moda

La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. En variables continuas, puede no existir moda o pueden existir múltiples modas (López & González, 2018). En el ejemplo presentado, todos los valores son únicos, por lo que no existe moda. La moda es particularmente útil para describir variables cualitativas o discretas, donde indica la categoría más frecuente.

7.2 Medidas de Dispersión

7.2.1 Rango

El rango es la medida de dispersión más simple y se define como la diferencia entre el valor máximo y el valor mínimo del conjunto de datos (López & González, 2018). Proporciona una idea general de la variabilidad, aunque es muy sensible a valores extremos.

Fórmula:

$$R = x_{\text{máx}} - x_{\text{mín}}$$

Cálculo con los datos de ejemplo:

$$R = 6.5 - 5.8 = 0.700 \,\text{t ha}^{-1}$$

7.2.2 Varianza

La varianza mide la dispersión promedio de los datos respecto a la media aritmética. Para datos muestrales, se utiliza el denominador $(n-1)$ para obtener un estimador insesgado de la varianza poblacional (Montgomery, 2017).

Fórmula de la varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Cálculo paso a paso:

$$\begin{aligned} s^2 &= \frac{0.0025 + 0.1225 + 0.1225 + 0.0225 + 0.0225 + 0.0625 + 0.0625 + 0.0025}{7} \\ &= \frac{0.4200}{7} = 0.0600 \ (\text{t ha}^{-1})^2 \end{aligned}$$

7.2.3 Desviación estándar

La desviación estándar es la raíz cuadrada positiva de la varianza y se expresa en las mismas unidades que los datos originales:

$$s = \sqrt{s^2} = \sqrt{0.0600} = 0.245 \,\text{t ha}^{-1}$$

7.2.4 Coeficiente de variación

El coeficiente de variación (CV) expresa la dispersión relativa respecto a la media, permitiendo comparar la variabilidad entre diferentes conjuntos de datos que pueden tener diferentes unidades o magnitudes (Steel & Torrie, 1980).

Fórmula:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Cálculo con los datos de ejemplo:

$$CV = \frac{0.245}{6.150} \times 100\% = 3.98\%$$

Este bajo coeficiente de variación indica una dispersión relativamente pequeña en los rendimientos, sugiriendo homogeneidad en el desempeño productivo de las parcelas.

7.3 Medidas de Posición Relativa

7.3.1 Cuartiles

Los cuartiles son valores que dividen un conjunto de datos ordenados en cuatro partes iguales. El primer cuartil (Q_1) es el valor por debajo del cual se encuentra el 25% de los datos, mientras que el tercer cuartil (Q_3) es el valor por debajo del cual se encuentra el 75% de los datos (López & González, 2018).

Método de cálculo:

Para un conjunto de datos ordenados de tamaño n, la posición de un cuartil se determina por:

$$\text{Posición} = p \times (n + 1)$$

donde $p = 0.25$ para Q_1 y $p = 0.75$ para Q_3.

Si la posición no es un número entero, se interpola linealmente entre los valores adyacentes.

Primer cuartil (Q_1):

$$\text{Posición de } Q_1 = 0.25 \times (8 + 1) = 2.25$$

$$Q_1 = x_2 + 0.25 \times (x_3 - x_2) = 5.9 + 0.25 \times (6.0 - 5.9) = 5.925 \text{ t ha}^{-1}$$

Tercer cuartil (Q_3):

$$\text{Posición de } Q_3 = 0.75 \times (8 + 1) = 6.75$$

$$Q_3 = x_6 + 0.75 \times (x_7 - x_6) = 6.3 + 0.75 \times (6.4 - 6.3) = 6.375 \text{ t ha}^{-1}$$

7.3.2 Rango intercuartílico (RIC)

El rango intercuartílico es una medida de dispersión que representa la diferencia entre el tercer cuartil y el primer cuartil. Esta medida indica la amplitud del 50% central de los datos y es menos sensible a valores extremos que el rango total (Steel & Torrie, 1980).

Fórmula:

$$RIC = Q_3 - Q_1$$

Cálculo con los datos de ejemplo:

$$RIC = 6.375 - 5.925 = 0.450 \text{ t ha}^{-1}$$

Interpretación del rango intercuartílico:

El rango intercuartílico de 0.450 t ha ^{1} indica que el 50% central de las parcelas presenta una variación de rendimiento de 0.450 toneladas por hectárea. Esta medida es particularmente útil para:

1. Identificar la dispersión de la porción central de los datos
2. Detectar valores atípicos (observaciones que se encuentran más allá de $Q_1 - 1.5 \times RIC$ o $Q_3 + 1.5 \times RIC$)
3. Comparar la variabilidad entre diferentes conjuntos de datos de manera robusta

En el contexto agronómico, un RIC relativamente pequeño sugiere que la mayoría de las parcelas tienen rendimientos similares, lo que puede indicar condiciones de cultivo homogéneas y prácticas de manejo consistentes (Montgomery, 2017).

7.3.3 Percentiles

Los percentiles dividen los datos en cien partes iguales, permitiendo identificar la posición relativa de cualquier observación dentro del conjunto de datos (Anderson et al., 2018). Son especialmente útiles para establecer rangos de referencia y realizar comparaciones.

Percentil 10 (P_{10}):

$$\text{Posición de } P_{10} = 0.10 \times (8 + 1) = 0.9$$

$$P_{10} = x_1 + 0.9 \times (x_2 - x_1) = 5.8 + 0.9 \times (5.9 - 5.8) = 5.890 \text{ t ha}^{-1}$$

Percentil 90 (P_{90}):

$$\text{Posición de } P_{90} = 0.90 \times (8 + 1) = 8.1$$

Como la posición (8.1) excede el número de datos (8), se utiliza el valor máximo:

$$P_{90} = x_8 = 6.5 \text{ t ha}^{-1}$$

7.3.4 Interpretación de los Resultados

Los resultados obtenidos proporcionan una descripción completa del comportamiento de los rendimientos de maíz:

1. **Tendencia central:** El rendimiento promedio es de 6.150 t ha [1], coincidiendo con la mediana, lo que sugiere una distribución simétrica.
2. **Dispersión:** La variabilidad es baja (CV = 3.98%), indicando homogeneidad en el desempeño productivo.
3. **Posición relativa:**
    a. El 25% de las parcelas tienen rendimientos por debajo de 5.925 t ha [1]
    b. El 75% de las parcelas tienen rendimientos por debajo de 6.375 t ha [1]
    c. Solo el 10% de las parcelas tienen rendimientos por debajo de 5.890 t ha [1]

Esta información es fundamental para la toma de decisiones en el manejo agronómico, permitiendo identificar parcelas de alto y bajo rendimiento, establecer metas productivas y evaluar la efectividad de diferentes prácticas de cultivo.

8 Otros tipos de medias

En el análisis estadístico, la media aritmética es comúnmente utilizada para representar el valor central de un conjunto de datos. Sin embargo, existen otros tipos de medias que resultan más apropiadas en ciertos contextos, especialmente cuando se trabaja con datos agronómicos que presentan características específicas (Steel & Torrie, 1980). Este apartado explora la media ponderada, la media geométrica y la media armónica, destacando sus aplicaciones y relevancia en el campo de la agronomía.

8.1 Media Ponderada

La media ponderada se define como un promedio en el cual cada valor del conjunto de datos recibe un peso que refleja su importancia relativa (Anderson et al., 2018). A diferencia de la media aritmética simple, que asigna igual peso a todos los valores, la media ponderada permite considerar la influencia desigual de cada observación en el resultado final.

Fórmula:

$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

donde x_i representa cada valor, w_i es el peso asignado a ese valor, y n es el número total de observaciones.

Aplicación en agronomía:

En el ámbito agronómico, la media ponderada es útil para calcular el rendimiento promedio de un cultivo en parcelas de diferentes tamaños. Por ejemplo, si se tienen cuatro parcelas con las siguientes superficies y rendimientos:

Parcela	Superficie (ha)	Rendimiento (t ha ^{1})
P	1.5	6.2
P	2.0	5.8
P	1.0	6.5
P	1.5	6.0

El rendimiento medio ponderado se calcula como:

$$
\begin{aligned}
\bar{x}_p &= \frac{(1.5)(6.2) + (2.0)(5.8) + (1.0)(6.5) + (1.5)(6.0)}{1.5 + 2.0 + 1.0 + 1.5} \\
&= \frac{9.3 + 11.6 + 6.5 + 9.0}{6.0} \\
&= \frac{36.4}{6.0} \\
&= 6.07 \text{ t ha}^{-1}
\end{aligned}
$$

Este resultado indica que el rendimiento medio ponderado por superficie es de 6.07 t ha ^{1}, lo cual proporciona una estimación más precisa del rendimiento global en comparación con una media aritmética simple.

8.2 Media Geométrica

La media geométrica es un tipo de promedio que se utiliza para calcular tasas de crecimiento o rendimientos relativos (Sokal & Rohlf, 1995). A diferencia de la media aritmética, que suma los valores y divide por el número de observaciones, la media geométrica multiplica los valores y extrae la raíz n-ésima del producto.

Fórmula:

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n}$$

donde x_i representa cada valor y n es el número total de observaciones.

Aplicación en agronomía:

La media geométrica es especialmente útil para calcular el crecimiento promedio de un cultivo a lo largo de varios años. Por ejemplo, si se tienen los siguientes rendimientos de maíz en cuatro años consecutivos:

Año	Rendimiento (t ha ^{1})
1	5.0
2	5.5
3	6.0
4	6.6

La tasa de crecimiento promedio se calcula como:

$$
\begin{aligned}
\bar{x}_g &= \sqrt[4]{5.0 \times 5.5 \times 6.0 \times 6.6} \\
&= \sqrt[4]{1089.0} \\
&\approx 5.74 \text{ t ha}^{-1}
\end{aligned}
$$

Este resultado indica que el rendimiento promedio anual, considerando el crecimiento a lo largo de los años, es de aproximadamente 5.74 t ha ^{1}.

8.3 Media Armónica

La media armónica es un tipo de promedio que se utiliza para calcular tasas o razones, especialmente cuando el denominador varía (Montgomery, 2017). A diferencia de la media aritmética, que promedia los valores directamente, la media armónica promedia los inversos de los valores y luego toma el inverso del resultado.

Fórmula:

$\bar{x}_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

donde x_i representa cada valor y n es el número total de observaciones.

Aplicación en agronomía:

La media armónica es útil para calcular el rendimiento promedio por unidad de tiempo o distancia. Por ejemplo, si se tienen los siguientes rendimientos de un cultivo en diferentes parcelas con distintas áreas:

Parcela	Área (ha)	Rendimiento (t)	Rendimiento (t/ha)
P	2.0	10.0	5.00
P	2.5	12.0	4.80
P	1.5	7.0	4.67
P	3.0	14.0	4.67

El rendimiento promedio por hectárea se calcula como:

$$\begin{aligned}\bar{x}_a &= \frac{4}{\frac{1}{5.00} + \frac{1}{4.80} + \frac{1}{4.67} + \frac{1}{4.67}} \\ &= \frac{4}{0.2 + 0.2083 + 0.2141 + 0.2141} \\ &= \frac{4}{0.8365} \\ &\approx 4.78 \text{ t ha}^{-1}\end{aligned}$$

Este resultado indica que el rendimiento promedio por hectárea, considerando las diferentes áreas de las parcelas, es de aproximadamente 4.78 t ha [1].

8.4 Análisis Comparativo de los Diferentes Tipos de Medias

A continuación se presenta un análisis comparativo de los cuatro tipos principales de medias utilizadas en estadística descriptiva, empleando como ejemplo los datos de rendimiento de maíz presentados anteriormente (López & González, 2018).

8.4.1 Datos de Referencia para los Cálculos

Para ilustrar las diferencias entre los tipos de medias, se utilizan los siguientes datos:

1. **Rendimientos (t ha ^{1}):** 6.2, 5.8, 6.5, 6.0
2. **Pesos/Superficies (ha):** 1.5, 2.0, 1.0, 1.5

8.4.2 Fórmulas y Cálculos Detallados

1. **Media Aritmética**

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{6.2 + 5.8 + 6.5 + 6.0}{4} \\ &= \frac{24.5}{4} \\ &= 6.125 \text{ t ha}^{-1}\end{aligned}$$

2. **Media Ponderada**

$$\begin{aligned}\bar{x}_p &= \frac{(1.5)(6.2) + (2.0)(5.8) + (1.0)(6.5) + (1.5)(6.0)}{1.5 + 2.0 + 1.0 + 1.5} \\ &= \frac{9.3 + 11.6 + 6.5 + 9.0}{6.0} \\ &= \frac{36.4}{6.0} \\ &= 6.07 \text{ t ha}^{-1}\end{aligned}$$

3. **Media Geométrica**

$$\begin{aligned}\bar{x}_g &= \sqrt[4]{6.2 \times 5.8 \times 6.5 \times 6.0} \\ &= \sqrt[4]{1406.04} \\ &\approx 6.12 \text{ t ha}^{-1}\end{aligned}$$

4. **Media Armónica**

$$\begin{aligned}\bar{x}_a &= \frac{4}{\frac{1}{6.2} + \frac{1}{5.8} + \frac{1}{6.5} + \frac{1}{6.0}} \\ &= \frac{4}{0.1613 + 0.1724 + 0.1538 + 0.1667} \\ &= \frac{4}{0.6542} \\ &\approx 6.11 \text{ t ha}^{-1}\end{aligned}$$

8.4.3 Análisis Comparativo de los Resultados

Los resultados obtenidos muestran diferencias sutiles pero importantes entre los tipos de medias:

1. **Media Aritmética (6.125 t ha ^{1}):** Proporciona el valor más alto, ya que no considera pesos ni ajustes por la naturaleza de los datos.
2. **Media Ponderada (6.067 t ha ^{1}):** Presenta el valor más bajo debido a que la parcela con mayor superficie (2.0 ha) tiene el rendimiento más bajo (5.8 t ha ^{1}), lo que reduce el promedio ponderado.
3. **Media Geométrica (6.120 t ha ^{1}):** Ofrece un valor intermedio, siendo menos sensible a valores extremos que la media aritmética.
4. **Media Armónica (6.114 t ha ^{1}):** Proporciona un valor ligeramente inferior a la media aritmética, dando mayor peso a los valores menores.

8.4.4 Cuadro Comparativo

Aspecto	**Media Aritmética**	**Media Ponderada**	**Media Geométrica**	**Media Armónica**
Fórmula	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$	$\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n}$	$\bar{x}_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
Resultado (t ha ^{1})	6.125	6.067	6.120	6.114
Características	Suma todos los valores y divide por n. Sensible a valores extremos	Considera la importancia relativa de cada valor mediante pesos	Apropiada para tasas de crecimiento y proporciones. Menos sensible a extremos	Útil para promediar razones y velocidades. Da mayor peso a valores pequeños
Aplicaciones Agronómicas	Rendimiento promedio simple, altura promedio de plantas	Rendimiento promedio considerando tamaño de parcelas	Tasa de crecimiento promedio, índices de productividad	Velocidad promedio de aplicación, eficiencia por unidad de tiempo
Ventajas	Fácil cálculo e interpretación. Ampliamente conocida	Refleja la importancia real de cada observación	Reduce el efecto de valores extremos. Apropiada para datos multiplicativos	Apropiada cuando se promedian razones. Conservativa con valores bajos

Desventajas	Muy sensible a valores atípicos	Requiere definir pesos apropiados	Solo aplicable a valores positivos. Interpretación menos intuitiva	Solo aplicable a valores positivos. Puede ser muy conservadora

8.4.5 Recomendaciones para la Selección del Tipo de Media

La elección del tipo de media apropiado depende del contexto específico del análisis (Steel & Torrie, 1980):

1. **Media Aritmética:** Para análisis descriptivos generales donde todos los valores tienen igual importancia.
2. **Media Ponderada:** Cuando las observaciones tienen diferente importancia o representan grupos de distinto tamaño.
3. **Media Geométrica:** Para datos que representan tasas de crecimiento, índices o proporciones.
4. **Media Armónica:** Para promediar velocidades, tasas con denominador variable o cuando se desea ser conservador con valores bajos.

9 Cálculos en R

9.1 Base de datos

El conjunto de datos IRIS es uno de los conjuntos de datos más utilizados en la literatura de estadística y aprendizaje automático. Fue introducido por Ronald Fisher en 1936 y contiene mediciones de cuatro características morfológicas de flores de tres especies distintas de iris: *Iris setosa*, *Iris versicolor* e *Iris virginica*. Este dataset es ampliamente empleado para ilustrar técnicas de análisis estadístico y clasificación supervisada (Fisher, 1936).

Referencia del dataset: Fisher, R. (1936). Iris [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76

Acceso a recursos: El script completo con el ejemplo desarrollado y la base de datos IRIS pueden descargarse en el siguiente repositorio: https://github.com/Ludwing-MJ/MTCDPR_sin_agrupar

9.2 Configuración del Entorno de Trabajo

Antes de comenzar cualquier análisis, es fundamental configurar adecuadamente el entorno de trabajo. Esto implica instalar y cargar los paquetes necesarios, así como explorar y comprender la estructura del conjunto de datos que se utilizará. En esta sección, se detallarán los pasos para configurar el entorno de trabajo y realizar una exploración inicial del conjunto de datos.

Se recomienda crear un proyecto nuevo en R para organizar adecuadamente el trabajo de estadística descriptiva. Se sugiere seguir los siguientes pasos para establecer un entorno de trabajo ordenado:

1. **Crear una nueva carpeta** en el directorio de trabajo denominada "Estadistica_Descriptiva_Iris"
2. **Crear un nuevo proyecto de R** dentro de esta carpeta utilizando RStudio
3. **Crear un script** donde se realizará y documentará el análisis estadístico

9.2.1 Instalación y Carga de Paquetes

Se procede a instalar y cargar los paquetes necesarios para el análisis estadístico descriptivo. Se utiliza la función condicional `if(!require())` para verificar si el paquete está instalado antes de proceder con la instalación:

```r
# Instalación y carga de paquetes necesarios
## Para manipulación y visualización de datos
if (!require(tidyverse)) install.packages("tidyverse")
## Para estadísticas descriptivas
if (!require(psych)) install.packages("psych")
```

9.2.2 Carga y exploración de los Datos

El dataset `iris` es un conjunto de datos clásico en estadística que contiene mediciones de características morfológicas de flores de tres especies del género *Iris*. Este dataset está incluido por defecto en R, lo que facilita su acceso para fines didácticos y de análisis estadístico.

```r
# Cargar el dataset iris
data(iris)

# Explorar la estructura del dataset
str(iris)
```

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

La función `str()` proporciona información sobre la estructura del dataset:

1. `object`: Nombre del objeto a examinar

```r
# Visualizar las primeras observaciones
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

La función `head()` muestra las primeras filas del dataset:

1. `x`: Objeto del cual mostrar las primeras filas
2. `n`: Número de filas a mostrar (por defecto 6)

```r
# Resumen básico del dataset
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

La función `summary()` proporciona un resumen estadístico básico:

1. `object`: Objeto del cual generar el resumen
2. `maxsum`: Número máximo de elementos a mostrar para factores (por defecto 7)
3. `digits`: Número de dígitos significativos para valores numéricos

9.3 Medidas de Tendencia Central

9.3.1 Media Aritmética

La función `mean()` calcula la media aritmética de un vector numérico.

1. `x`: Vector numérico del cual se calculará la media.
2. `trim`: Fracción de valores a recortar de cada extremo del vector (por defecto 0).
3. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).

```r
# Calcular la media para la longitud del sépalo
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

9.3.2 Mediana

La función `median()` calcula la mediana de un vector numérico.

1. `x`: Vector numérico del cual se calculará la mediana.
2. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).
3. `type`: Tipo de algoritmo para calcular la mediana (entero entre 1 y 9).

```
# Calcular la mediana para la longitud del sépalo
median(iris$Sepal.Length)
```

```
[1] 5.8
```

9.3.3 Moda

No existe una función base en R para calcular la moda directamente. Se puede crear una función personalizada para calcular la moda que maneja valores faltantes y múltiples modas:

```
# Función para calcular la moda
moda <- function(x) {
  # Eliminar valores NA
  x <- na.omit(x)

  # Verificar si el vector está vacío
  if (length(x) == 0) return(NA_character_)

  # Calcular la frecuencia de cada valor
  tabla <- table(x)

  # Identificar el/los valores con mayor frecuencia
  max_frecuencia <- max(tabla)
  modas <- names(tabla[tabla == max_frecuencia])

  # Verificar si todos los valores son únicos (sin moda)
  if (max_frecuencia == 1) return(NA_character_)

  # Retornar la moda como un string separado por comas
  return(paste(modas, collapse = ", "))
}
```

Una vez ya definida la función para calcular la moda (tarea que se realiza la cada vez que se abre el software y se desea cargar la función en el entorno de trabajo). Se procede a calcular la moda para la longitud del sépalo:

```
# Calcular la moda para la longitud del sépalo
moda(iris$Sepal.Length)
```

```
[1] "5"
```

9.4 Medidas de Dispersión

9.4.1 Rango

La función `range()` devuelve los valores mínimo y máximo de un vector numérico. La función `diff()` calcula la diferencia entre los valores máximo y mínimo.

1. `x`: Vector numérico del cual se calculará el rango.
2. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).

```
# Obtener el valor minimo y máximo de la logitud del sépalo
range(iris$Sepal.Length)
```

```
[1] 4.3 7.9
```

```
# Calcular el rango para la longitud del sépalo
diff(range(iris$Sepal.Length))
```

```
[1] 3.6
```

9.4.2 Varianza

La función `var()` calcula la varianza de un vector numérico.

1. `x`: Vector numérico del cual se calculará la varianza.
2. `y`: Vector numérico opcional para calcular la covarianza.
3. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).
4. `use`: Método para manejar valores faltantes (por defecto “everything”).

```
# Calcular la varianza para la longitud del sépalo
var(iris$Sepal.Length)
```

```
[1] 0.6856935
```

9.4.3 Desviación Estándar

La función `sd()` calcula la desviación estándar de un vector numérico.

1. `x`: Vector numérico del cual se calculará la desviación estándar.
2. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).

```
# Calcular la desviación estándar para la longitud del sépalo
sd(iris$Sepal.Length)
```

```
[1] 0.8280661
```

9.4.4 Coeficiente de Variación

No existe una función base en R para calcular el coeficiente de variación directamente. Se puede crear una función personalizada o calcularse mediante operaciones aritméticas:

```
# Función para calcular el coeficiente de variación
cv <- function(x) {
  (sd(x) / mean(x)) * 100
}
# Calcular el coeficiente de variación para la longitud del sépalo
cv(iris$Sepal.Length)
```

```
[1] 14.17113
```

```
# Calcular el coeficiente de variación para la longitud del sépalo
(sd(iris$Sepal.Length) / mean(iris$Sepal.Length)) * 100
```

```
[1] 14.17113
```

9.5 Medidas de Posición Relativa

9.5.1 Cuartiles

La función `quantile()` calcula los cuartiles y otros percentiles de un vector numérico.

1. `x`: Vector numérico del cual se calcularán los cuantiles.
2. `probs`: Vector de probabilidades (0.25 para Q , 0.75 para Q).
3. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).
4. `type`: Método de cálculo (entero entre 1 y 9, por defecto 7).

```
# Calcular cuartiles para la longitud del sépalo
quantile(iris$Sepal.Length, probs = c(0.25, 0.5, 0.75))
```

```
25% 50% 75%
5.1 5.8 6.4
```

```
# Calcular Q1 y Q3 por separado
Q1 <- quantile(iris$Sepal.Length, 0.25); Q1
```

```
25%
5.1
```

```r
Q3 <- quantile(iris$Sepal.Length, 0.75); Q3
```

```
75%
6.4
```

9.5.2 Rango intercuartílico

La función `IQR()` calcula en automático el rango intercuartílico (Q1-Q3) de un vector numérico.

1. `x`: Vector numérico del cual se calcularán los cuantiles.
2. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).
3. `type`: Método de cálculo (entero entre 1 y 9, por defecto 7).

```r
# Rango intercuartílico
IQR(iris$Sepal.Length)
```

```
[1] 1.3
```

9.5.3 Percentiles

La función `quantile()` también se utiliza para calcular percentiles.

1. `x`: Vector numérico del cual se calcularán los cuantiles.
2. `probs`: Vector de probabilidades (ej. 0.10 para el percentil 10, 0.90 para el percentil 90).
3. `na.rm`: Valor lógico que indica si se deben remover los valores `NA` (por defecto `FALSE`).
4. `type`: Método de cálculo (entero entre 1 y 9, por defecto 7).

```r
# Calcular percentiles específicos
quantile(iris$Sepal.Length, c(0.10, 0.90))
```

```
10% 90%
4.8 6.9
```

```r
# Calcular el percentil 95
quantile(iris$Sepal.Length, 0.95)
```

```
  95%
7.255
```

9.6 Análisis Completo con el Paquete psych

La función `describe()` del paquete `psych` calcula múltiples estadísticas descriptivas en una sola línea de código.

1. `x`: Data frame o vector numérico.
2. `na.rm`: Remover valores faltantes (por defecto `TRUE`).
3. `trim`: Fracción para la media recortada (por defecto 0.1).
4. `skew`: Calcular asimetría (por defecto `TRUE`).
5. `kurtosis`: Calcular curtosis (por defecto `TRUE`).
6. `ranges`: Calcular rangos (por defecto `TRUE`).

```r
# Análisis descriptivo completo del dataset iris
describe(iris[,1:4])
```

```
             vars   n mean   sd median trimmed  mad min max range  skew
Sepal.Length    1 150 5.84 0.83   5.80    5.81 1.04 4.3 7.9   3.6  0.31
Sepal.Width     2 150 3.06 0.44   3.00    3.04 0.44 2.0 4.4   2.4  0.31
Petal.Length    3 150 3.76 1.77   4.35    3.76 1.85 1.0 6.9   5.9 -0.27
Petal.Width     4 150 1.20 0.76   1.30    1.18 1.04 0.1 2.5   2.4 -0.10
             kurtosis   se
Sepal.Length    -0.61 0.07
Sepal.Width      0.14 0.04
Petal.Length    -1.42 0.14
Petal.Width     -1.36 0.06
```

Interpretación de los Estimadores de la Función describe()

La función `describe()` proporciona los siguientes estimadores estadísticos:

1. **n**: Número de observaciones válidas (sin valores faltantes)
2. **mean**: Media aritmética de los datos
3. **sd**: Desviación estándar muestral
4. **median**: Mediana o percentil 50
5. **trimmed**: Media recortada al 10% (elimina el 10% de valores extremos de cada cola)
6. **mad**: Desviación absoluta mediana, medida robusta de dispersión
7. **min**: Valor mínimo observado
8. **max**: Valor máximo observado
9. **range**: Diferencia entre el valor máximo y mínimo

10. **skew**: Coeficiente de asimetría. Valores cercanos a 0 indican distribución simétrica, valores positivos indican asimetría hacia la derecha, valores negativos hacia la izquierda
11. **kurtosis**: Coeficiente de curtosis. Valores cercanos a 0 indican distribución normal, valores positivos indican distribución leptocúrtica (más puntiaguda), valores negativos indican distribución platicúrtica (más aplanada)
12. **se**: Error estándar de la media, calculado como sd/√n

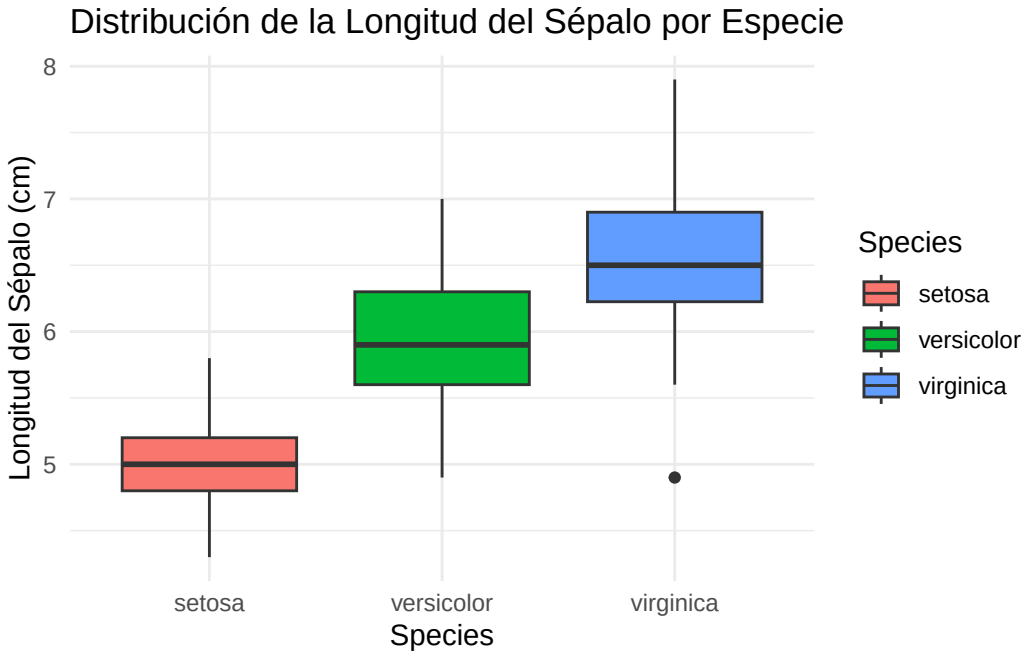
9.7 Visualización de Datos

La visualización de datos es una parte esencial del análisis estadístico descriptivo, ya que permite identificar patrones, tendencias y anomalías en los datos de manera gráfica. A continuación, se presentan ejemplos de diferentes tipos de gráficos que se pueden utilizar para visualizar el dataset `iris`.

9.7.1 Diagrama de Caja (Boxplot)

El diagrama de caja es una herramienta útil para visualizar la distribución de una variable numérica y comparar distribuciones entre diferentes grupos. Este gráfico muestra la mediana, los cuartiles (Q1 y Q3), los valores atípicos y los bigotes (valores mínimo y máximo dentro de un rango razonable).

```
# Diagrama de caja para visualizar la longitud del sépalo por especie
ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Distribución de la Longitud del Sépalo por Especie",
       y = "Longitud del Sépalo (cm)")
```



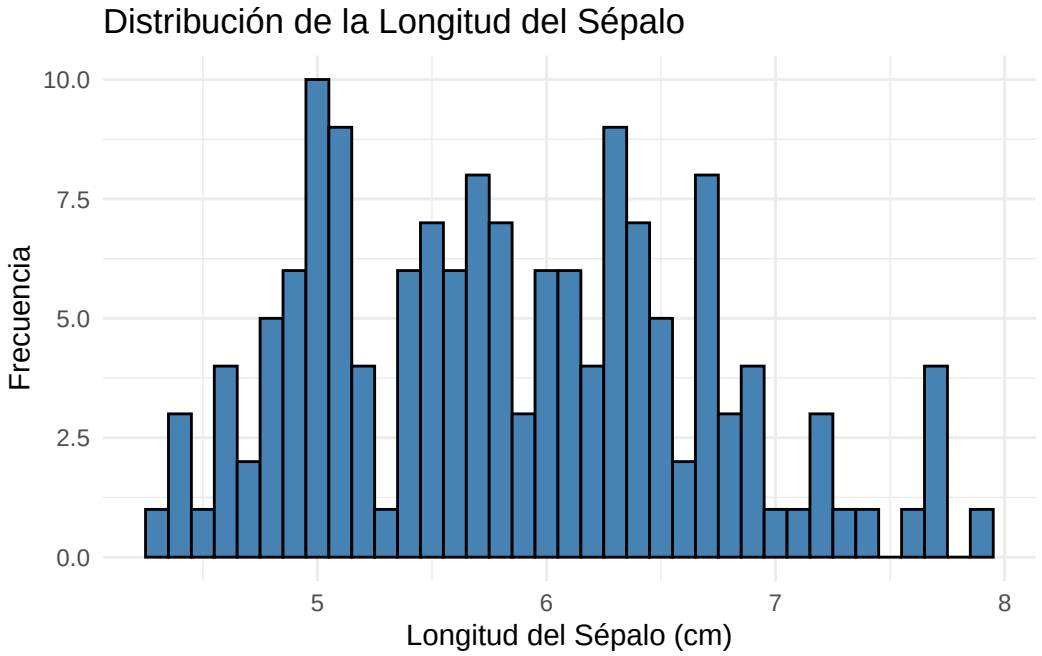
Interpretación:

1. La línea dentro de la caja representa la mediana.
2. Los bordes de la caja representan los cuartiles Q1 (25%) y Q3 (75%).
3. Los bigotes se extienden hasta los valores mínimo y máximo dentro de 1.5 veces el rango intercuartílico (IQR).
4. Los puntos fuera de los bigotes son considerados valores atípicos.

9.7.2 Histograma

El histograma es un gráfico que muestra la distribución de frecuencia de una variable numérica. Divide los datos en intervalos (bins) y muestra la frecuencia de observaciones en cada intervalo.

```
# Histograma de la longitud del sépalo
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(binwidth = 0.1, fill = "steelblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribución de la Longitud del Sépalo",
       x = "Longitud del Sépalo (cm)",
       y = "Frecuencia")
```



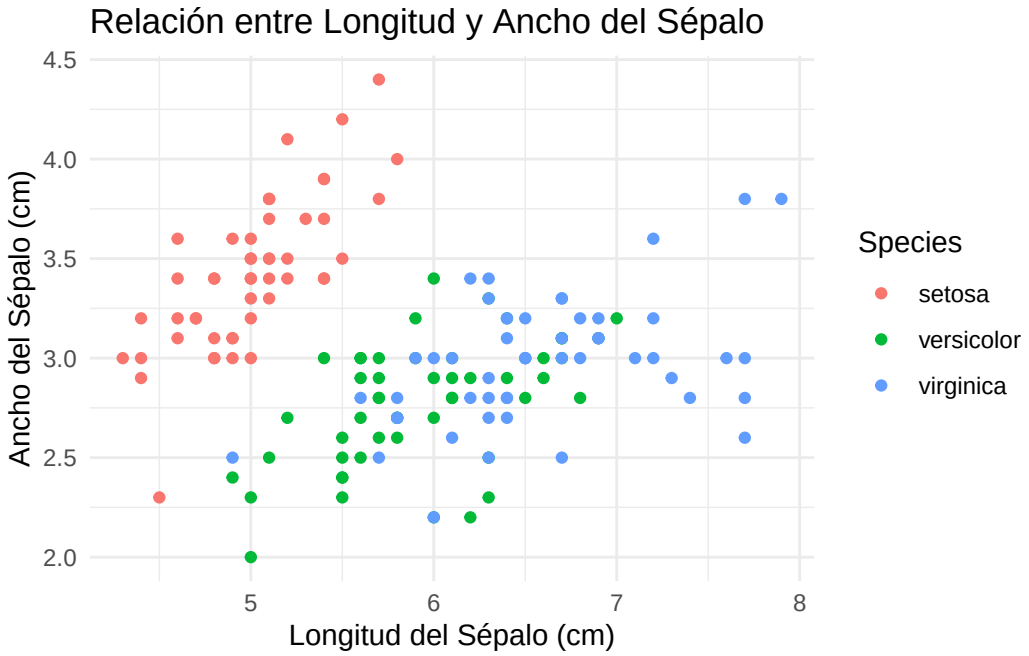
Interpretación:

1. El eje x representa los valores de la variable.
2. El eje y representa la frecuencia de observaciones en cada intervalo.
3. La forma del histograma puede indicar la simetría, asimetría y curtosis de la distribución.

9.7.3 Gráfico de Dispersión (Scatter Plot)

El gráfico de dispersión se utiliza para visualizar la relación entre dos variables numéricas. Cada punto en el gráfico representa una observación, con la posición del punto determinada por los valores de las dos variables.

```
# Gráfico de dispersión entre la longitud y el ancho del sépalo
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Relación entre Longitud y Ancho del Sépalo",
       x = "Longitud del Sépalo (cm)",
       y = "Ancho del Sépalo (cm)")
```



Interpretación:

1. El gráfico muestra la relación entre dos variables.
2. Los patrones en el gráfico pueden indicar correlación positiva, negativa o ninguna correlación.
3. Se pueden utilizar diferentes colores o formas para representar diferentes grupos.

Capítulo V

Estadística descriptiva para datos agrupados

10 Introducción y formulario

El análisis de datos agrupados es una técnica que permite resumir y describir conjuntos extensos de observaciones mediante la organización de los valores en intervalos o clases (Lind, Marchal, & Wathen, 2017). Esta agrupación facilita la identificación de patrones y tendencias generales, así como la comparación entre diferentes conjuntos de datos. Para caracterizar la distribución de los datos agrupados, se emplean varios tipos de medidas: las medidas de tendencia central, las medidas de dispersión y las medidas de posición relativa (Triola, 2018).

Aplicaciones de la estadística descriptiva para datos agrupados

A pesar de la disponibilidad de software estadístico avanzado, el análisis de datos agrupados sigue siendo relevante en diversas situaciones prácticas (Anderson et al., 2018). A continuación, se presentan algunos campos de aplicación:

1. **Estudios epidemiológicos:** Los datos sobre la incidencia de enfermedades a menudo se presentan en forma de rangos de edad o niveles de exposición, lo que requiere el uso de técnicas de datos agrupados para calcular tasas y comparar poblaciones (Triola, 2018).
2. **Investigaciones de mercado:** Los datos sobre ingresos o gastos de los consumidores pueden estar disponibles solo en forma de intervalos, lo que exige el uso de métodos de datos agrupados para estimar promedios y evaluar la distribución del gasto (Lind et al., 2017).
3. **Agronomía:** Los datos sobre rendimientos de cultivos o características del suelo pueden estar agrupados debido a limitaciones en la precisión de la medición o a la necesidad de proteger la confidencialidad de los datos (López & González, 2018).
4. **Control de calidad:** En la industria, los datos sobre las dimensiones de productos o el tiempo de vida útil pueden estar agrupados en rangos para facilitar el análisis y la toma de decisiones (Anderson et al., 2018).
5. **Análisis demográfico:** Los datos sobre la distribución de la población por grupos de edad y nivel socioeconómico se analizan mediante técnicas de datos agrupados para comprender las características de una población (Lind et al., 2017).
6. **Análisis de riesgos:** En finanzas y seguros, los datos sobre pérdidas o siniestros se agrupan para evaluar la probabilidad de eventos extremos y establecer primas o reservas adecuadas (Triola, 2018).

10.1 Construcción de la Tabla de Frecuencia

Antes de calcular estas medidas, es necesario construir una tabla de frecuencia para los datos agrupados. Este proceso implica los siguientes pasos:

10.1.1 Determinación del número de clases

El número de clases, denotado como k, se estima frecuentemente mediante la regla de Sturges, que se expresa como:

$$k = 1 + 3.322 \log_{10}(N)$$

donde N representa el número total de observaciones. Esta fórmula proporciona una guía para seleccionar un número de clases que permita un análisis adecuado, evitando tanto la excesiva fragmentación como la pérdida de información relevante (Triola, 2014).

Recomendación para la aproximación del resultado de la regla de Sturges: Dado que el número de clases debe ser un entero, es común redondear el resultado de la fórmula de Sturges al entero más cercano. Sin embargo, es importante considerar el contexto del análisis y la naturaleza de los datos. En algunos casos, puede ser preferible redondear hacia arriba o hacia abajo para obtener un número de clases que facilite la interpretación y la comparación. Por ejemplo, si el resultado de la fórmula es 6.2, se podría optar por 6 o 7 clases, dependiendo de si se prefiere una representación más resumida o más detallada de los datos. En general, se recomienda experimentar con diferentes números de clases y evaluar el impacto en la claridad y la utilidad del análisis (Anderson et al., 2018).

10.1.2 Cálculo del intervalo de clase

El intervalo de clase, simbolizado como c, corresponde a la amplitud de cada clase y se calcula dividiendo el rango de los datos entre el número de clases:

$$c = \frac{Rango}{k}$$

El rango se obtiene restando el valor mínimo del valor máximo del conjunto de datos. Es recomendable ajustar el valor de c a un número conveniente para facilitar la interpretación y la construcción de la tabla (Lind et al., 2017).

10.1.3 Definición de los límites de clase

Cada clase se define por un límite inferior y un límite superior. Para evitar ambigüedades en la asignación de los datos a las clases, se utiliza una notación estándar:

1. **Corchete [:** Indica que el límite está incluido en el intervalo.
2. **Paréntesis) :** Indica que el límite no está incluido en el intervalo.

Por ejemplo, el intervalo [10, 20) incluye todos los valores desde 10 hasta 19.999..., pero no incluye el valor 20. Esta distinción es crucial para evitar ambigüedades y asegurar que cada dato se clasifique en un único intervalo. La correcta definición de los límites de clase garantiza que cada observación se asigne a una única clase, evitando la superposición y facilitando el análisis (López & González, 2018).

10.1.4 Frecuencia Absoluta

La frecuencia absoluta, denotada como f_i, representa el número de observaciones que pertenecen a la clase i. Este valor proporciona una medida directa de la concentración de datos en cada intervalo y es fundamental para el cálculo de las demás medidas estadísticas (Triola, 2014).

10.1.5 Frecuencia Relativa

La frecuencia relativa, denotada como fr_i, se calcula dividiendo la frecuencia absoluta de la clase i entre el número total de observaciones N:

$$fr_i = \frac{f_i}{N}$$

La frecuencia relativa expresa la proporción de observaciones que pertenecen a cada clase y permite comparar la distribución de diferentes conjuntos de datos, independientemente de su tamaño. La suma de las frecuencias relativas de todas las clases debe ser igual a 1 (Lind et al., 2017).

10.1.6 Frecuencia Acumulada

La frecuencia acumulada, denotada como Fa_i, representa el número total de observaciones que son menores o iguales al límite superior de la clase i. Se calcula sumando las frecuencias absolutas de todas las clases anteriores a la clase i y la frecuencia absoluta de la clase i:

$$Fa_i = \sum_{j=1}^{i} f_j$$

La frecuencia acumulada proporciona información sobre la distribución de los datos a lo largo de todo el rango de valores y es útil para identificar percentiles y cuartiles (Anderson et al., 2018).

10.1.7 Frecuencia Relativa Acumulada

La frecuencia relativa acumulada, denotada como Fra_i, se calcula dividiendo la frecuencia acumulada de la clase i entre el número total de observaciones N:

$$Fra_i = \frac{Fa_i}{N}$$

La frecuencia relativa acumulada expresa la proporción de observaciones que son menores o iguales al límite superior de la clase i y permite comparar la distribución de diferentes conjuntos de datos en términos de proporciones acumuladas. La frecuencia relativa acumulada del último intervalo debe ser igual a 1 (Triola, 2014).

10.2 Medidas de Tendencia Central para Datos Agrupados

Una vez construida la tabla de frecuencia, se procede a calcular las medidas de tendencia central, que resumen la posición central de la distribución de los datos. Las principales medidas de tendencia central para datos agrupados son la media aritmética, la mediana y la moda.

10.2.1 Media Aritmética

La media aritmética para datos agrupados, denotada como $\bar{x}$, se calcula como la suma ponderada de los puntos medios de cada clase, donde los pesos son las frecuencias absolutas de cada clase:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{N}$$

donde f_i es la frecuencia absoluta de la clase i, x_i es el punto medio de la clase i, k es el número de clases y N es el número total de observaciones. El punto medio de cada clase se calcula como el promedio de los límites inferior y superior de la clase (Lind et al., 2017).

10.2.2 Mediana

La mediana es el valor que divide la distribución de los datos en dos partes iguales. Para calcular la mediana en datos agrupados, primero se identifica la clase mediana, que es la primera clase cuya frecuencia acumulada es mayor o igual a $N/2$. Luego, se aplica la siguiente fórmula:

$$Me = L_{inf} + \frac{\frac{N}{2} - Fa_{ant}}{f_m} \cdot c$$

donde L_{inf} es el límite inferior de la clase mediana, N es el número total de observaciones, Fa_{ant} es la frecuencia acumulada de la clase anterior a la clase mediana, f_m es la frecuencia absoluta de la clase mediana y c es el intervalo de clase (Triola, 2014).

10.2.3 Moda

La moda es el valor que ocurre con mayor frecuencia en la distribución de los datos. Para calcular la moda en datos agrupados, primero se identifica la clase modal, que es la clase con la mayor frecuencia absoluta. Luego, se aplica la siguiente fórmula:

$$Mo = L_{inf} + \frac{d_1}{d_1 + d_2} \cdot c$$

donde L_{inf} es el límite inferior de la clase modal, d_1 es la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase anterior, d_2 es la diferencia entre la frecuencia de la clase modal y la frecuencia de la clase posterior, y c es el intervalo de clase (Anderson et al., 2018).

10.3 Medidas de Dispersión para Datos Agrupados

Las medidas de dispersión cuantifican el grado de variabilidad o dispersión de los datos respecto a las medidas de tendencia central. Las principales medidas de dispersión para datos agrupados son el rango, la varianza, la desviación estándar y el coeficiente de variación.

10.3.1 Rango

El rango es la medida de dispersión más simple y se calcula como la diferencia entre el valor máximo y el valor mínimo de los datos. Para datos agrupados, el rango se aproxima restando el límite inferior de la primera clase al límite superior de la última clase:

$$Rango = L_{sup,k} - L_{inf,1}$$

donde $L_{sup,k}$ es el límite superior de la última clase y $L_{inf,1}$ es el límite inferior de la primera clase. Aunque es fácil de calcular, el rango es sensible a los valores extremos y no proporciona información sobre la distribución de los datos entre los extremos (Triola, 2014).

10.3.2 Varianza

La varianza es una medida que cuantifica la dispersión de los datos respecto a la media aritmética. En el caso de datos agrupados, la varianza muestral se calcula considerando la frecuencia de cada clase y el punto medio correspondiente. La fórmula clásica para la varianza es la siguiente:

$$s^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{N-1}$$

donde f_i es la frecuencia absoluta de la clase i, x_i es el punto medio de la clase i, $\bar{x}$ es la media aritmética y N es el número total de observaciones. La varianza proporciona una medida de la dispersión de los datos alrededor de la media, pero se expresa en unidades al cuadrado, lo que dificulta su interpretación directa (Lind et al., 2017).

Como alternativa, la varianza también puede calcularse utilizando una fórmula operativa, que resulta especialmente útil cuando se dispone de la suma de los productos de las frecuencias por los puntos medios y sus cuadrados. Esta fórmula es algebraicamente equivalente a la anterior y se expresa así:

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{N}}{N-1}$$

donde f_i es la frecuencia absoluta de la clase i, x_i es el punto medio de la clase i, N es el número total de observaciones y kkk es el número de clases. En esta fórmula, $\sum_{i=1}^{k} f_i x_i^2$ representa la suma de los productos de la frecuencia por el cuadrado del punto medio de

cada clase, mientras que $\sum_{i=1}^{k} f_i x_i$ corresponde a la suma de los productos de la frecuencia por el punto medio de cada clase.

Ambas fórmulas, la clásica y la operativa, son equivalentes y proporcionan el mismo resultado si se aplican correctamente (Lind et al., 2017; López & González, 2018; Triola, 2018).

10.3.3 Desviación Estándar

La desviación estándar es la raíz cuadrada de la varianza y se expresa en las mismas unidades que los datos originales. Para datos agrupados, la desviación estándar se calcula como:

$$s = \sqrt{s^2}$$

La desviación estándar proporciona una medida de la dispersión de los datos alrededor de la media y es más fácil de interpretar que la varianza. Un valor alto de la desviación estándar indica una mayor dispersión de los datos, mientras que un valor bajo indica una menor dispersión (Anderson et al., 2018).

10.3.4 Coeficiente de Variación

El coeficiente de variación es una medida relativa de dispersión que se calcula dividiendo la desviación estándar entre la media aritmética:

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

El coeficiente de variación expresa la dispersión de los datos como un porcentaje de la media y permite comparar la variabilidad de diferentes conjuntos de datos, independientemente de sus unidades de medida. Un valor alto del coeficiente de variación indica una mayor variabilidad relativa, mientras que un valor bajo indica una menor variabilidad relativa (Triola, 2014).

10.4 Medidas de Posición Relativa para Datos Agrupados

Las medidas de posición relativa describen la ubicación de un valor específico en relación con el resto de los datos. Las principales medidas de posición relativa son los cuartiles y los percentiles.

10.4.1 Cuartiles

Los cuartiles dividen la distribución de los datos en cuatro partes iguales, cada una conteniendo el 25% de las observaciones. Los tres cuartiles se denotan como Q_1, Q_2 y Q_3.

1. Q_1 (Primer Cuartil): Es el valor que separa el 25% inferior de los datos del 75% superior.
2. Q_2 (Segundo Cuartil): Es el valor que coincide con la mediana y separa el 50% inferior de los datos del 50% superior.
3. Q_3 (Tercer Cuartil): Es el valor que separa el 75% inferior de los datos del 25% superior.

Para calcular los cuartiles en datos agrupados, primero se identifica la clase cuartil, que es la primera clase cuya frecuencia acumulada es mayor o igual a $i \cdot N/4$, donde i es el número del cuartil (1, 2 o 3). Luego, se aplica la siguiente fórmula:

$$Q_i = L_{inf} + \frac{\frac{i \cdot N}{4} - Fa_{ant}}{f_q} \cdot c$$

donde L_{inf} es el límite inferior de la clase cuartil, N es el número total de observaciones, Fa_{ant} es la frecuencia acumulada de la clase anterior a la clase cuartil, f_q es la frecuencia absoluta de la clase cuartil y c es el intervalo de clase (Lind et al., 2017).

10.4.2 Percentiles

Los percentiles dividen la distribución de los datos en cien partes iguales, cada una conteniendo el 1% de las observaciones. El percentil P_p es el valor que separa el p% inferior de los datos del $(100 - p)$% superior.

Para calcular los percentiles en datos agrupados, primero se identifica la clase percentil, que es la primera clase cuya frecuencia acumulada es mayor o igual a $p \cdot N/100$. Luego, se aplica la siguiente fórmula:

$$P_p = L_{inf} + \frac{\frac{p \cdot N}{100} - Fa_{ant}}{f_p} \cdot c$$

donde L_{inf} es el límite inferior de la clase percentil, N es el número total de observaciones, Fa_{ant} es la frecuencia acumulada de la clase anterior a la clase percentil, f_p es la frecuencia absoluta de la clase percentil y c es el intervalo de clase (Triola, 2014).

11 Ejemplo empleando el formulario

11.1 Base de datos

Referencia del dataset: Fisher, R. (1936). Iris [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76

Acceso a recursos: El script completo con el ejemplo desarrollado y la base de datos IRIS pueden descargarse en el siguiente repositorio:

A continuación, se presenta un conjunto de datos correspondientes a la longitud del pétalo (en cm) de 150 flores de la especie *Iris*, organizados en formato matricial para facilitar su visualización y análisis. Estos datos serán utilizados para ilustrar el cálculo de estadísticos descriptivos para datos agrupados, siguiendo las metodologías propuestas en la sección aterior.

1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5
1.5	1.6	1.4	1.1	1.2	1.5	1.3	1.4	1.7	1.5
1.7	1.5	1.0	1.7	1.9	1.6	1.6	1.5	1.4	1.6
1.6	1.5	1.5	1.4	1.5	1.2	1.3	1.4	1.3	1.5
1.3	1.3	1.3	1.6	1.9	1.4	1.6	1.4	1.5	1.4
4.7	4.5	4.9	4.0	4.6	4.5	4.7	3.3	4.6	3.9
3.5	4.2	4.0	4.7	3.6	4.4	4.5	4.1	4.5	3.9
4.8	4.0	4.9	4.7	4.3	4.4	4.8	5.0	4.5	3.5
3.8	3.7	3.9	5.1	4.5	4.5	4.7	4.4	4.1	4.0
4.4	4.6	4.0	3.3	4.2	4.2	4.2	4.3	3.0	4.1
6.0	5.1	5.9	5.6	5.8	6.6	4.5	6.3	5.8	6.1
5.1	5.3	5.5	5.0	5.1	5.3	5.5	6.7	6.9	5.0
5.7	4.9	6.7	4.9	5.7	6.0	4.8	4.9	5.6	5.8
6.1	6.4	5.6	5.1	5.6	6.1	5.6	5.5	4.8	5.4
5.6	5.1	5.1	5.9	5.7	5.2	5.0	5.2	5.4	5.1

11.2 Construcción de la Tabla de Frecuencias

11.2.1 Determinación del rango (R)

El rango es la diferencia entre el valor máximo y el valor mínimo de la variable:

$$R = X_{\max} - X_{\min}$$

Para la longitud de pétalo:

$$R = 6.9 - 1.0 = 5.9$$

11.2.2 Cálculo del número de clases (K)

El número de clases se determina con la Regla de Sturges:

$$k = 1 + 3.322 \log_{10} N$$

Donde n es el número total de datos:

$$k = 1 + 3.322 \log_{10} 150 \approx 1 + 3.322 \times 2.1761 \approx 8.22$$

Se redondea al entero más cercano:

$$k = 8$$

11.2.3 Cálculo de la amplitud de clase (C)

La amplitud de clase se calcula así:

$$C = \frac{R}{k}$$

Sustituyendo valores:

$$C = \frac{5.9}{8} = 0.7375 \approx 0.75$$

11.2.4 Determinación de los límites de clase

El primer límite inferior es el valor mínimo (1.0). Los siguientes se obtienen sumando la amplitud de clase (C).
Para evitar que un valor pertenezca a dos clases al mismo tiempo, se utiliza la notación de intervalos semiabiertos:

1. El corchete [indica que el límite inferior está incluido en la clase.
2. El paréntesis) indica que el límite superior no está incluido en la clase.

Por ejemplo, el primer intervalo se escribe:

$$[1.00, 1.75)$$

Esto significa que la clase incluye todos los valores x tales que $1.00 \leq x < 1.75$.

Los intervalos de clase quedan así:

$$\begin{aligned}
&\text{Clase 1:} \quad [1.00, 1.75) \\
&\text{Clase 2:} \quad [1.75, 2.50) \\
&\text{Clase 3:} \quad [2.50, 3.25) \\
&\text{Clase 4:} \quad [3.25, 4.00) \\
&\text{Clase 5:} \quad [4.00, 4.75) \\
&\text{Clase 6:} \quad [4.75, 5.50) \\
&\text{Clase 7:} \quad [5.50, 6.25) \\
&\text{Clase 8:} \quad [6.25, 7.00]
\end{aligned}$$

Nótese que en la última clase se utiliza el corchete de cierre] para incluir el valor máximo.

11.2.5 Cálculo de la marca de clase

La marca de clase es el punto medio de cada intervalo:

$$x_i = \frac{L_i + L_s}{2}$$

Por ejemplo, para la primera clase:

$$x_1 = \frac{1.00 + 1.75}{2} = 1.375$$

11.2.6 Cálculo de la frecuencia absoluta

La frecuencia absoluta es el número de datos en cada clase, obtenida por conteo directo.

11.2.7 Cálculo de la frecuencia relativa

La frecuencia relativa se calcula así:

$$fr_i = \frac{f_i}{N}$$

Por ejemplo, para la primera clase:

$$fr_i = \frac{48}{150} = 0.32$$

11.2.8 Cálculo de la frecuencia acumulada

La frecuencia acumulada es la suma de las frecuencias absolutas hasta la clase i:

$$fa_i = \sum_{j=1}^{i} f_j$$

Por ejemplo, para la cuarta clase:

$$fa_4 = f_1 + f_2 + f_3 + f_4 = 48 + 0 + 0 + 15 = 63$$

11.2.9 Cálculo de $f_i x_i$ y $f_i x_i^2$

Estos productos se utilizan para cálculos posteriores:

$$f_i x_i = f_i \times x_i$$

$$f_i x_i^2 = f_i \times (x_i)^2$$

Por ejemplo, para la primera clase:

$$f_1 x_1 = 48 \times 1.375 = 66.00$$

$$f_1 x_1^2 = 48 \times (1.375)^2 = 48 \times 1.890625 = 90.75$$

11.3 Tabla de frecuencia

Clase	Límite Inferior (LI)	Límite Superior (LS)	Marca de clase (x_i)	Frecuencia (f_i)	Frecuencia relativa (fr_i)	Frecuencia acumulada (fa_i)	$f_i x_i$	$f_i x_i^2$
1	1.000	1.750	1.375	48	0.320	48	66.000	90.750
2	1.750	2.500	2.125	2	0.013	50	4.250	9.031
3	2.500	3.250	2.875	1	0.007	51	2.875	8.266
4	3.250	4.000	3.625	10	0.067	61	36.250	131.406
5	4.000	4.750	4.375	34	0.227	95	148.750	650.781
6	4.750	5.500	5.125	27	0.180	122	138.375	709.172
7	5.500	6.250	5.875	22	0.147	144	129.250	759.344
8	6.250	7.000	6.625	6	0.040	150	39.750	263.344
Total				150	1.000		565.500	2622.094

11.4 Medidas de tendencia central

Una vez construida la tabla de frecuencia, se procede a calcular las medidas de tendencia central, que resumen la posición central de la distribución de los datos.

11.4.1 Media Aritmética

La formula para calcular la media aritmpetica es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i \cdot x_i}{N}$$

Sustituyendo valores

$$\bar{x} = \frac{565.50}{150} = 3.77$$

11.4.2 Mediana

Para el calculo de la mediana hay que identificar la primera clase donde la frecuencia acumulada fa_i supera $N/2$. Para este ejemplo $N/2$ al sustituir valores equivale a $150/2 = 75$ siendo la clase numero 5 aquella donde la frecuencia acumulada supera $N/2$ siendo la formula para el cálculo de la mediana la siguiente:

$$Me = L_{inf} + \frac{\frac{N}{2} - Fa_{ant}}{f_m} \cdot c$$

Sustituyendo valores

$$Me = 4.00 + \frac{\frac{150}{2} - 61}{34} \cdot 0.75 = 4.31$$

11.4.3 Moda

Para el calculo de la moda hay que identificar clase con mayor frecuencia absoluta siendo la clase numero 1 para este ejemplo. Siendo la formula para el cálculo de la moda la siguiente:

$$Mo = L_{inf} + \frac{d_1}{d_1 + d_2} \cdot c$$

Sustituyendo valores:

$$Mo = 1.00 + \frac{(48 - 0)}{(48 - 0) + (48 - 2)} \cdot 0.75 = 1.383$$

11.5 Medidas de dispersión

11.5.1 Rango

El rango se aproxima restando el límite inferior de la primera clase al límite superior de la última clase siendo su formula la siguiente:

$$Rango = L_{sup,k} - L_{inf,1}$$

Sustituyendo valores:

$$Rango = 7.00 - 1.00 = 6.00$$

11.5.2 Varianza

Para el calculo de la varianza se utilizará la siguiente formula operativa, que resulta especialmente útil porque se dispone de la suma de los productos de las frecuencias por los puntos medios y sus cuadrados.

$$s^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{N}}{N-1}$$

Sustituyendo valores:

$$s^2 = \frac{2622.094 - \frac{(565.500)^2}{150}}{150-1} = 3.29$$

11.5.3 Desviación Estándar

La desviación estándar es la raíz cuadrada de la varianza y se expresa en las mismas unidades que los datos originales. Para datos agrupados, la desviación estándar se calcula como:

$$s = \sqrt{s^2}$$

Sustituyendo valores

$$s = \sqrt{3.29} = 1.645$$

11.5.4 Coeficiente de Variación

El coeficiente de variación es una medida relativa de dispersión que se calcula dividiendo la desviación estándar entre la media aritmética:

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Sustituyendo valores:

$$CV = \frac{1.645}{3.77} \cdot 100\% = 43.63\%$$

11.6 Medidas de posición relativa

11.6.1 Cuartiles

Para calcular los cuartiles en datos agrupados, primero se identifica la clase cuartil, que es la primera clase cuya frecuencia acumulada es mayor o igual a $i \cdot N/4$, donde i es el número del cuartil (1, 2 o 3). Luego, se aplica la siguiente fórmula:

$$Q_i = L_{inf} + \frac{\frac{i \cdot N}{4} - Fa_{ant}}{f_q} \cdot c$$

Para el ejemplo se calculará el cuartil 1 (Q1) por lo que primero se identifica la clase dentro de la que se encuentra, para ello se usa la formula $i \cdot N/4$, sustituyendo valores esto sería $1 \cdot 150/4 = 38.5$ siendo la clase donde la frecuencia acumulada supera este valor por primera vez la clase 1, ya con esta información se procede a sustituir valores en la fórmula.

$$Q_1 = 1.0 + \frac{\frac{1 \cdot 150}{4} - 0}{48} \cdot 0.75 = 1.59$$

11.6.2 Percentiles

Para calcular los percentiles en datos agrupados, primero se identifica la clase percentil, que es la primera clase cuya frecuencia acumulada es mayor o igual a $p \cdot N/100$. Luego, se aplica la siguiente fórmula:

$$P_p = L_{inf} + \frac{\frac{p \cdot N}{100} - Fa_{ant}}{f_p} \cdot c$$

Para el ejemplo se calculará el percentil 80 ($p = 80$) para ello primero se identifica la clase a la que pertenece este percentil usando la formula $p \cdot N/100$ la cual al sustituir los valores equivale a: $80 \cdot 150/100 = 120$ con este dato se ubica la clase 6 como la clase donde se

encuentra el percentil 80 al ser la primera donde la frecuencia acumulada supera 120. Una vez obtenida esta información se procede a sustituir valores en la formula:

$$P_{80} = 4.75 + \frac{120 - 95}{27} \cdot 0.75 = 5.44$$

11.7 Interpretación de Resultados

11.7.1 Media aritmética

La media aritmética obtenida fue de 3.77 cm. Esto indica que, en promedio, la longitud del pétalo de las flores analizadas es de 3.77 centímetros. Esta medida representa el valor central alrededor del cual tienden a agruparse los datos y es útil para describir el comportamiento general de la variable en estudio (López & González, 2018).

11.7.2 Mediana

La mediana calculada fue de 4.31 cm. Esto significa que el 50% de las flores tiene una longitud de pétalo menor o igual a 4.31 cm, mientras que el otro 50% tiene una longitud mayor o igual a este valor. La mediana es especialmente útil cuando la distribución de los datos es asimétrica o presenta valores extremos, ya que no se ve afectada por estos (López & González, 2018).

11.7.3 Moda

La moda resultó ser 1.375 cm, correspondiente a la clase con mayor frecuencia absoluta. Esto indica que la longitud de pétalo más común entre las flores analizadas se encuentra en el intervalo de 1.00 a 1.75 cm. La moda es relevante para identificar el valor o rango de valores que se presentan con mayor frecuencia en el conjunto de datos (López & González, 2018).

11.7.4 Rango

El rango calculado fue de 6.00 cm, lo que representa la diferencia entre la longitud máxima y mínima de los pétalos observados. Este valor proporciona una idea general de la dispersión de los datos, mostrando el intervalo total en el que se distribuyen las observaciones (López & González, 2018).

11.7.5 Varianza y desviación estándar

La varianza obtenida fue de 3.67 cm^2 y la desviación estándar fue de 1.83 cm. Estos valores indican que, en promedio, las longitudes de los pétalos se desvían 1.83 cm respecto a la media. Una desviación estándar relativamente alta, como en este caso, sugiere que existe una considerable variabilidad en la longitud de los pétalos dentro del grupo analizado (López & González, 2018).

11.7.6 Coeficiente de variación

El coeficiente de variación fue de 46.88%. Este valor, al ser mayor al 30%, indica que la dispersión relativa de los datos respecto a la media es alta. En términos prácticos, esto significa que la longitud de los pétalos presenta una considerable heterogeneidad dentro del conjunto de flores estudiadas (López & González, 2018).

11.7.7 Cuartil 1 (Q1)

El primer cuartil (Q1) se ubicó en 1.59 cm. Esto implica que el 25% de las flores tiene una longitud de pétalo menor o igual a 1.59 cm. El análisis de los cuartiles permite identificar la distribución de los datos en segmentos y facilita la detección de posibles asimetrías o concentraciones de valores (López & González, 2018).

11.7.8 Percentil 80 (P80)

El percentil 80 se calculó en 5.78 cm, lo que significa que el 80% de las flores tiene una longitud de pétalo menor o igual a 5.78 cm. El percentil 80 es útil para identificar valores altos dentro de la distribución y para realizar comparaciones entre diferentes grupos o tratamientos (López & González, 2018).

12 Ejemplo en R

12.1 Base de datos

Referencia del dataset: Fisher, R. (1936). Iris [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76

Acceso a recursos: El script completo con el ejemplo desarrollado y la base de datos IRIS pueden descargarse en el siguiente repositorio: https://github.com/Ludwing-MJ/MTCDPR_datos_agrupados

A continuación, se presenta un conjunto de datos correspondientes a la longitud del pétalo (en cm) de 150 flores de la especie *Iris*, organizados en formato matricial para facilitar su visualización y análisis. Estos datos serán utilizados para ilustrar el cálculo de estadísticos descriptivos para datos agrupados, siguiendo las metodologías propuestas en la sección anterior.

1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5
1.5	1.6	1.4	1.1	1.2	1.5	1.3	1.4	1.7	1.5
1.7	1.5	1.0	1.7	1.9	1.6	1.6	1.5	1.4	1.6
1.6	1.5	1.5	1.4	1.5	1.2	1.3	1.4	1.3	1.5
1.3	1.3	1.3	1.6	1.9	1.4	1.6	1.4	1.5	1.4
4.7	4.5	4.9	4.0	4.6	4.5	4.7	3.3	4.6	3.9
3.5	4.2	4.0	4.7	3.6	4.4	4.5	4.1	4.5	3.9
4.8	4.0	4.9	4.7	4.3	4.4	4.8	5.0	4.5	3.5
3.8	3.7	3.9	5.1	4.5	4.5	4.7	4.4	4.1	4.0
4.4	4.6	4.0	3.3	4.2	4.2	4.2	4.3	3.0	4.1
6.0	5.1	5.9	5.6	5.8	6.6	4.5	6.3	5.8	6.1
5.1	5.3	5.5	5.0	5.1	5.3	5.5	6.7	6.9	5.0
5.7	4.9	6.7	4.9	5.7	6.0	4.8	4.9	5.6	5.8
6.1	6.4	5.6	5.1	5.6	6.1	5.6	5.5	4.8	5.4
5.6	5.1	5.1	5.9	5.7	5.2	5.0	5.2	5.4	5.1

12.2 Preparación del entorno de trabajo

```
# Instalación y carga de paquetes necesarios
## Para manipulación y visualización de datos
if (!require(tidyverse)) install.packages("tidyverse")
## Para exportar archivos en excel
if (!require(writexl)) install.packages("writexl")
## Para importar archivos en excel
if (!require(readxl)) install.packages("readxl")
```

12.3 Carga y Preparación de Datos

Primero, se carga el conjunto de datos iris y se extrae la variable de interés, en este caso, la longitud del pétalo.

```
# Cargar el dataset iris
data(iris)

# Extraer la variable longitud de pétalo
longitud_petalo <- iris$Petal.Length
```

12.4 Determinación de parámetros básicos para la agrupación

Se define una función personalizada para calcular los parámetros necesarios para agrupar los datos: número de observaciones, valores mínimo y máximo, rango, número de clases (usando la regla de Sturges) y amplitud de clase.

```
# Función para calcular parámetros de agrupamiento
calcular_parametros_agrupamiento <- function(datos) {
  n <- length(datos)
  x_min <- min(datos)
  x_max <- max(datos)
  rango <- x_max - x_min

  # Regla de Sturges para número de clases
  k <- round(1 + 3.322 * log10(n))

  # Amplitud de clase
  amplitud <- rango / k

  return(list(
    n = n,
    x_min = x_min,
    x_max = x_max,
    rango = rango,
    k = k,
    amplitud = amplitud
  ))
}
```

Una vez ya definida la función para calcular los parámetros necesarios para la agrupación de los datos (tarea que se realiza la cada vez que se abre el software y se desea cargar la función en el entorno de trabajo). Se procede a calcularlos:

```r
# Aplicar función
parametros <- calcular_parametros_agrupamiento(longitud_petalo)
# Visualizar el resultado
parametros
```

```
$n
[1] 150

$x_min
[1] 1

$x_max
[1] 6.9

$rango
[1] 5.9

$k
[1] 8

$amplitud
[1] 0.7375
```

12.5 Construcción de la tabla de frecuencias

Se utiliza una función personalizada para construir la tabla de frecuencias, calculando los límites de clase, marcas de clase, frecuencias absolutas, relativas y acumuladas, así como sumas necesarias para los cálculos posteriores.

```r
# Función corregida para construir tabla de frecuencias
construir_tabla_frecuencias <- function(datos, parametros) {

  # Crear breaks (puntos de corte) para las clases
  # Esto garantiza exactamente k clases
  breaks <- seq(parametros$x_min,
                parametros$x_max,
                length.out = parametros$k + 1)

  # Crear límites de clase a partir de los breaks
  limite_inferior <- breaks[-length(breaks)]  # Todos excepto el último
  limite_superior <- breaks[-1]               # Todos excepto el primero

  # Calcular marcas de clase
  marca_clase <- (limite_inferior + limite_superior) / 2

  # Calcular frecuencias absolutas usando cut()
  intervalos <- cut(datos,
```

```
                    breaks = breaks,
                    include.lowest = TRUE,
                    right = FALSE,
                    labels = FALSE)  # Usar números en lugar de etiquetas

  # Contar frecuencias por clase
  frecuencia_absoluta <- as.numeric(table(factor(intervalos,
                                                 levels = 1:parametros$k)))

  # Reemplazar NA por 0 si alguna clase queda vacía
  frecuencia_absoluta[is.na(frecuencia_absoluta)] <- 0

  # Calcular frecuencias derivadas
  frecuencia_relativa <- frecuencia_absoluta / parametros$n
  frecuencia_acumulada <- cumsum(frecuencia_absoluta)
  fi_xi <- frecuencia_absoluta * marca_clase
  fi_xi2 <- frecuencia_absoluta * (marca_clase^2)

  # Crear tabla
  tabla <- data.frame(
    Clase = 1:parametros$k,
    Limite_Inferior = round(limite_inferior, 3),
    Limite_Superior = round(limite_superior, 3),
    Marca_Clase = round(marca_clase, 3),
    Frecuencia_Absoluta = frecuencia_absoluta,
    Frecuencia_Relativa = round(frecuencia_relativa, 4),
    Frecuencia_Acumulada = frecuencia_acumulada,
    fi_xi = round(fi_xi, 3),
    fi_xi2 = round(fi_xi2, 3)
  )

  return(tabla)
}
```

Una vez ya definida la función para construir la tabla de frecuencias (tarea que se realiza la cada vez que se abre el software y se desea cargar la función en el entorno de trabajo). Se procede a emplear la función para construir la tabla:

```
# Construir tabla de frecuencias
tabla_freq <- construir_tabla_frecuencias(longitud_petalo, parametros)

# Mostrar tabla
tabla_freq
```

```
  Clase Limite_Inferior Limite_Superior Marca_Clase Frecuencia_Absoluta
1     1           1.000           1.738       1.369                  48
2     2           1.738           2.475       2.106                   2
3     3           2.475           3.213       2.844                   1
```

```
4     4           3.213           3.950       3.581                  10
5     5           3.950           4.688       4.319                  29
6     6           4.688           5.425       5.056                  32
7     7           5.425           6.163       5.794                  22
8     8           6.163           6.900       6.531                   6
  Frecuencia_Relativa Frecuencia_Acumulada   fi_xi  fi_xi2
1              0.3200                   48  65.700  89.927
2              0.0133                   50   4.213   8.873
3              0.0067                   51   2.844   8.087
4              0.0667                   61  35.812 128.254
5              0.1933                   90 125.244 540.896
6              0.2133                  122 161.800 818.101
7              0.1467                  144 127.463 738.486
8              0.0400                  150  39.188 255.943
```

La tabla de frecuencias es la base para calcular las medidas de tendencia central y dispersión en datos agrupados. Cada fila representa un intervalo de clase y sus frecuencias asociadas. Si se desea exportar la tabla de frecuencias en un formato tabular para su presentación se utiliza la función `write_xlsx` como se muestra a continuación.

```r
# Exportar la tabla de frecuencias
write_xlsx(tabla_freq, "tabla_frecuencias.xlsx")
```

Al ejecutar esta linea de código R automáticamente guardará un archivo .xlsx en la carpeta del proyecto.

12.6 Medidas de Tendencia Central

Se define una función personalizada para calcular la media, mediana y moda a partir de la tabla de frecuencias.

```r
# Función para calcular medidas de tendencia central
calcular_tendencia_central <- function(tabla, parametros) {
  # Media aritmética
  media <- sum(tabla$fi_xi) / parametros$n

  # Mediana
  n <- parametros$n
  posicion_mediana <- n / 2
  clase_mediana <- which(tabla$Frecuencia_Acumulada >= posicion_mediana)[1]
  L <- tabla$Limite_Inferior[clase_mediana]
  F_anterior <- ifelse(clase_mediana == 1,
                       0, tabla$Frecuencia_Acumulada[
                         clase_mediana - 1])
  f_m <- tabla$Frecuencia_Absoluta[clase_mediana]
  A <- tabla$Limite_Superior[clase_mediana] -
    tabla$Limite_Inferior[clase_mediana]
```

```
  mediana <- L + ((posicion_mediana - F_anterior) / f_m) * A

  # Moda
  clase_modal <- which.max(tabla$Frecuencia_Absoluta)
  fa_ant <- ifelse(clase_modal == 1,
                   0, tabla$Frecuencia_Absoluta[
                     clase_modal - 1])
  fa_sig <- ifelse(clase_modal == parametros$k,
                   0, tabla$Frecuencia_Absoluta[
                     clase_modal + 1])
  d1 <- tabla$Frecuencia_Absoluta[clase_modal] - fa_ant
  d2 <- tabla$Frecuencia_Absoluta[clase_modal] - fa_sig
  if ((d1 + d2) == 0) {
    moda <- NA
  } else {
    moda <- tabla$Limite_Inferior[clase_modal] + (d1 / (d1 + d2)) * A
  }

  return(list(media = media, mediana = mediana, moda = moda))
}
```

En esta función la media se calcula como el promedio ponderado de las marcas de clase. La mediana y la moda se estiman usando fórmulas específicas para datos agrupados, considerando la posición dentro de la clase correspondiente, una vez ya definida la función se procede a utilizarla para calcular las medidas de tendencia central.

```
# Calcular medidas
tendencia <- calcular_tendencia_central(tabla_freq, parametros)

# Mostrar resultados
tendencia
```

```
$media
[1] 3.748427

$mediana
[1] 4.306276

$moda
[1] 1.376851
```

12.7 Medidas de Dispersión

Se utiliza una función personalizada para calcular el rango, la varianza, la desviación estándar y el coeficiente de variación.

```r
# Función para calcular medidas de dispersión
calcular_dispersion <- function(tabla, parametros, media) {
  # Rango aproximado
  rango_aprox <- tabla$Limite_Superior[parametros$k] -
    tabla$Limite_Inferior[1]

  # Varianza
  varianza <- (sum(tabla$fi_xi2) - (sum(tabla$fi_xi)^2 / parametros$n)) /
    (parametros$n - 1)

  # Desviación estándar
  desviacion_std <- sqrt(varianza)

  # Coeficiente de variación
  cv <- (desviacion_std / media) * 100

  return(list(
    rango = rango_aprox,
    varianza = varianza,
    desviacion_std = desviacion_std,
    cv = cv
  ))
}
```

El rango es la diferencia entre el límite superior del último intervalo y el límite inferior del primero. La varianza y la desviación estándar se calculan usando las sumas ponderadas de las marcas de clase al cuadrado. El coeficiente de variación expresa la dispersión relativa respecto a la media. Para estos cálculos la función emplea las formulas presentadas en la sección anterior y una vez definida se procede al cálculo de las medidas de dispersión:

```r
# Calcular medidas de dispersión
dispersion <- calcular_dispersion(tabla_freq, parametros, tendencia$media)

# Mostrar los resultados
dispersion
```

```
$rango
[1] 5.9

$varianza
[1] 3.22793

$desviacion_std
[1] 1.796644

$cv
[1] 47.93062
```

12.8 Medidas de Posición Relativa

Finalmente, se puede calcular cualquier cuartil o percentil usando una función personalizada.

```
# Función para calcular cuartiles y percentiles
calcular_posicion_relativa <- function(tabla,
                                       parametros, posicion,
                                       tipo = "cuartil") {
  if (tipo == "cuartil") {
    pos_valor <- posicion * parametros$n / 4
  } else if (tipo == "percentil") {
    pos_valor <- posicion * parametros$n / 100
  }

  clase_objetivo <- which(tabla$Frecuencia_Acumulada >= pos_valor)[1]
  fa_anterior <- ifelse(clase_objetivo == 1, 0,
                        tabla$Frecuencia_Acumulada[clase_objetivo - 1])

  valor <- tabla$Limite_Inferior[clase_objetivo] +
    ((pos_valor - fa_anterior) /
       tabla$Frecuencia_Absoluta[clase_objetivo]) * parametros$amplitud

  return(valor)
}
```

Esta función permite calcular cualquier medida de posición relativa, como cuartiles o percentiles, utilizando la tabla de frecuencias y la fórmula correspondiente para datos agrupados. Una vez definida en el entorno de trabajo se procede a utilizar para calcular Q1 y P80 como en el ejemplo anterior:

```
# Calcular Q1
Q1 <- calcular_posicion_relativa(tabla_freq, parametros, 1, "cuartil")
Q1
```

```
[1] 1.576172
```

```
# Calcular P80
P80 <- calcular_posicion_relativa(tabla_freq, parametros, 80, "percentil")
P80
```

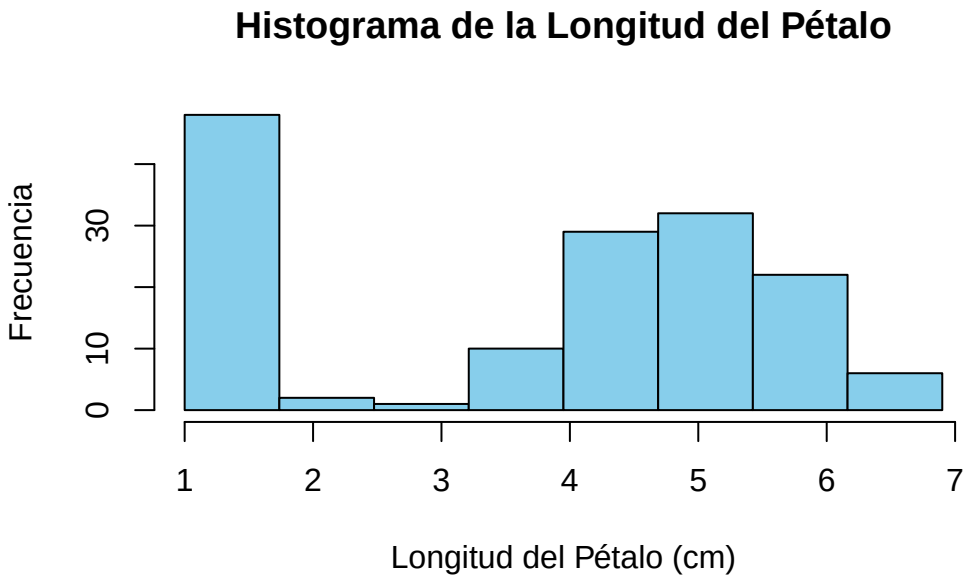
```
[1] 5.379406
```

12.9 Histograma

El histograma es un gráfico de barras que representa la distribución de frecuencias de los datos agrupados. Cada barra corresponde a un intervalo de clase, y su altura es proporcional a la frecuencia absoluta o relativa de ese intervalo.

Construcción en R:

```
hist(longitud_petalo,
     breaks = seq(min(longitud_petalo),
                  max(longitud_petalo),
                  length.out = parametros$k + 1),
     main = "Histograma de la Longitud del Pétalo",
     xlab = "Longitud del Pétalo (cm)",
     ylab = "Frecuencia",
     col = "skyblue",
     border = "black")
```



Explicación:

1. `hist()`: Función para crear histogramas en R.
2. `longitud_pedalo`: Variable a graficar.
3. `breaks`: Define los límites de los intervalos de clase. Se utiliza `seq()` para generar una secuencia de valores desde el mínimo hasta el máximo de la variable, dividida en `k + 1` puntos (donde `k` es el número de clases).
4. `main`, `xlab`, `ylab`: Títulos y etiquetas de los ejes.
5. `col`, `border`: Colores de las barras y del borde.

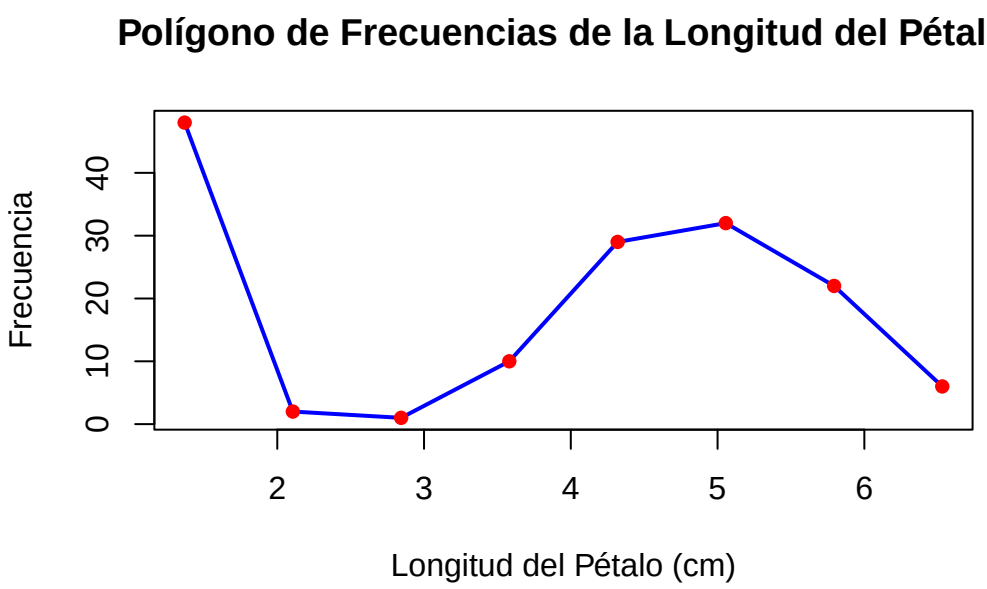
12.10 Polígono de Frecuencias

El polígono de frecuencias es un gráfico de líneas que conecta los puntos medios de las barras del histograma. Se construye uniendo los puntos correspondientes a las marcas de clase y sus respectivas frecuencias.

Construcción en R:

```
# Crear el polígono de frecuencias
plot(tabla_freq$Marca_Clase,
     tabla_freq$Frecuencia_Absoluta,
     type = "l",  # "l" para líneas
     main = "Polígono de Frecuencias de la Longitud del Pétalo",
     xlab = "Longitud del Pétalo (cm)",
     ylab = "Frecuencia",
     col = "blue",
     lwd = 2)  # Grosor de la línea

# Agregar puntos en las marcas de clase
points(tabla_freq$Marca_Clase,
       tabla_freq$Frecuencia_Absoluta,
       col = "red", pch = 16)
```




```
# pch = 16 para círculos rellenos
```

Explicación:

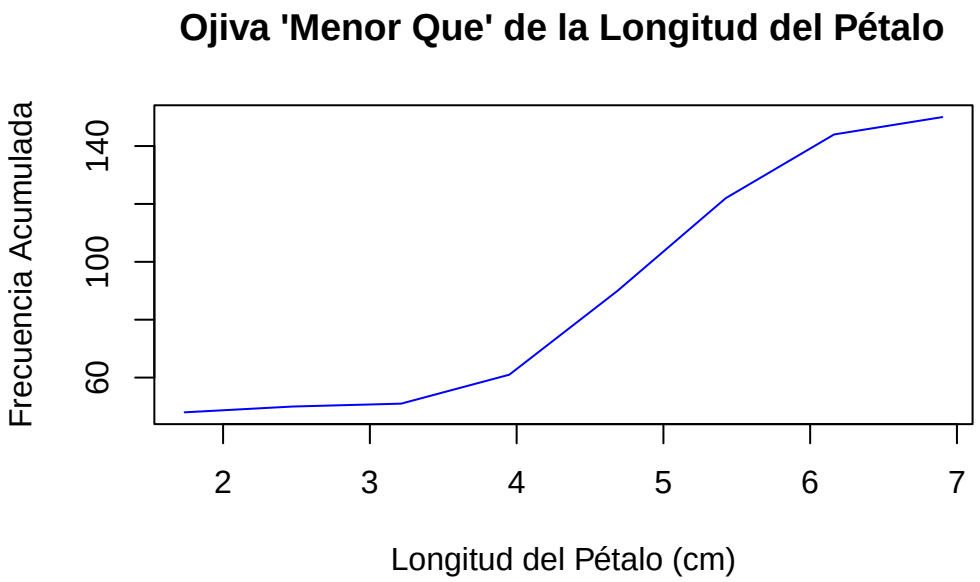
1. `plot(type = "l")`: Crea un gráfico de líneas.
2. `tabla_freq$Marca_Clase`y `tabla_freq$Frecuencia_Absoluta`: Vectores con las marcas de clase y las frecuencias absolutas.
3. `points()`: Agrega puntos en las marcas de clase para resaltar los valores.

12.11 Ojiva (Polígono de Frecuencias Acumuladas)

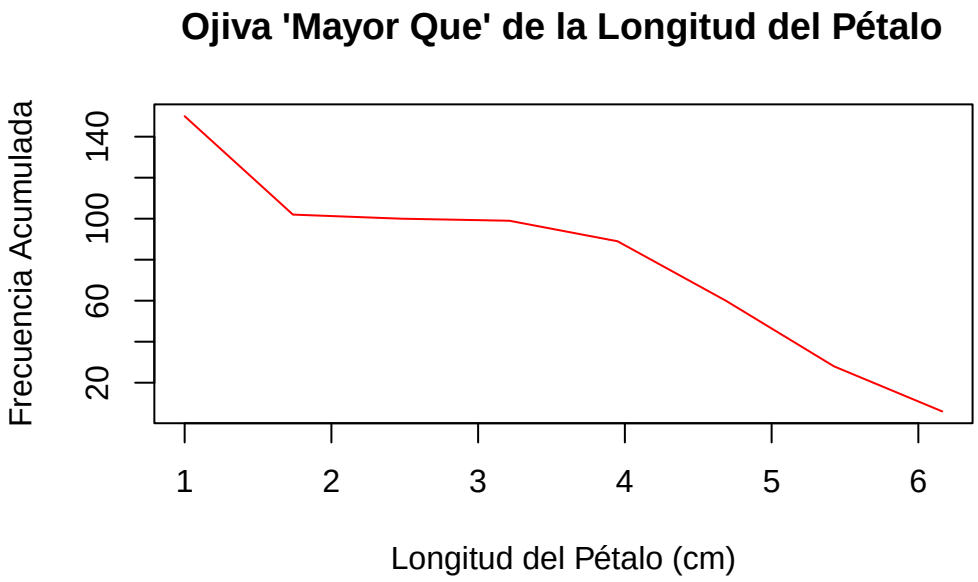
La ojiva es un gráfico de líneas que representa las frecuencias acumuladas. Se construye uniendo los puntos correspondientes a los límites superiores de los intervalos de clase y sus respectivas frecuencias acumuladas.

Construcción en R:

```
# Ojiva "Menor Que"
plot(tabla_freq$Limite_Superior, tabla_freq$Frecuencia_Acumulada,
     type = "l",
     main = "Ojiva 'Menor Que' de la Longitud del Pétalo",
     xlab = "Longitud del Pétalo (cm)",
     ylab = "Frecuencia Acumulada",
     col = "blue")
```




```
# Ojiva "Mayor Que"
frecuencia_acumulada_mayor_que <- rev(cumsum(
  rev(tabla_freq$Frecuencia_Absoluta)))
plot(tabla_freq$Limite_Inferior, frecuencia_acumulada_mayor_que,
     type = "l",
     main = "Ojiva 'Mayor Que' de la Longitud del Pétalo",
     xlab = "Longitud del Pétalo (cm)",
     ylab = "Frecuencia Acumulada",
     col = "red")
```

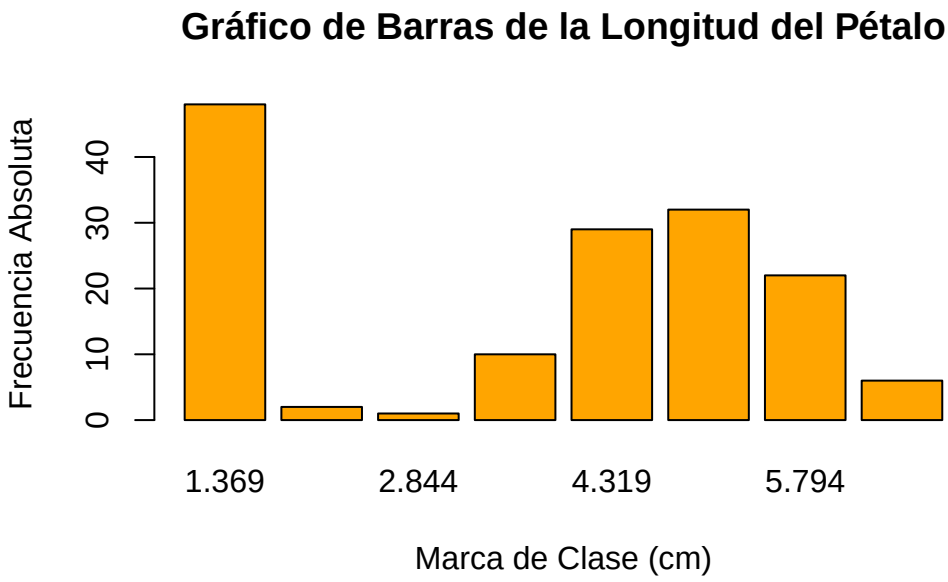


12.12 Gráfico de Barras

Aunque el histograma es el gráfico más común para datos agrupados, también se puede utilizar un gráfico de barras para representar las frecuencias de cada clase.

Construcción en R:

```
barplot(tabla_freq$Frecuencia_Absoluta,
        names.arg = tabla_freq$Marca_Clase,
        main = "Gráfico de Barras de la Longitud del Pétalo",
        xlab = "Marca de Clase (cm)",
        ylab = "Frecuencia Absoluta",
        col = "orange",
        border = "black")
```



Explicación:

1. `barplot()`: Función para crear gráficos de barras en R.
2. `tabla_freq$Frecuencia_Absoluta`: Vector con las frecuencias absolutas.
3. `names.arg`: Etiquetas para cada barra (en este caso, las marcas de clase).

12.13 Cálculos a partir de una tabla de frecuencias

No siempre es posible encontrar la base de datos completa para poder construir la tabla de frecuencias y realizar las estimaciones, muchas veces se parte de una tabla de frecuencias debido a la sensibilidad de los datos, privacidad o porque los datos son muy antiguos y se han perdido los registos, para este ejemplo se va a explicar como usar las funciones partiendo de la tabla de frecuencias exportada a un archivo excel previamente en esta sección:

12.13.1 Importar la tabla de frecuencias

Cabe resaltar que para que esto funcione la tabla de frecuencias que se vaya a importar debe tener el mismo formato (numero y nombre de columnas) que la tabla que se muestra a continuación:

Clase	Limite_Inferior	Limite_Superior	Marca_Clase	Frecuencia_Absoluta	Frecuencia_Relativa	Frecuencia_Acumulada	fi_xi	fi_xi2
1	1	1.7375	1.369	48	0.32	48	65.7	89.927
2	1.7375	2.475	2.106	2	0.0133	50	4.213	8.873
3	2.475	3.2125	2.844	1	0.0067	51	2.844	8.087
4	3.2125	3.95	3.581	10	0.0667	61	35.812	128.254
5	3.95	4.6875	4.319	29	0.1933	90	125.244	540.896
6	4.6875	5.425	5.056	32	0.2133	122	161.8	818.101
7	5.425	6.1625	5.794	22	0.1467	144	127.463	738.486
8	6.1625	6.9	6.531	6	0.04	150	39.188	255.943

```r
#Importar tabla de frecuencias
tabla<-read_excel("tabla_frecuencias.xlsx")
```

12.13.2 Estimación de los parámetros de agrupación

Una vez importada la tabla de frecuencias adecuadamente se procede a estimar los parámetros de agrupación a partir de ella, ya que estos son indispensables para las funciones elaboradas para estimar las medidas de tendencia central, dispersión y posición relativa.

```r
# Funcion personalizada para calcular los parametros
calcular_parametros_desde_tabla <- function(tabla) {
  n <- sum(tabla$Frecuencia_Absoluta)
  x_min <- min(tabla$Limite_Inferior)
  x_max <- max(tabla$Limite_Superior)
  rango <- x_max - x_min
  k <- nrow(tabla)
  amplitud <- (tabla$Limite_Superior[1] - tabla$Limite_Inferior[1])

  return(list(
    n = n,
    x_min = x_min,
    x_max = x_max,
    rango = rango,
    k = k,
    amplitud = amplitud
  ))
}
```

Una vez cargada la función al entorno de trabajo esta se utiliza con la tabla de frecuencias previamente importada para estimar los parámetros de agrupación

```r
# Estimar los parametros de agrupacion a partir de la tabla de frecuencias
parametros_tabla <- calcular_parametros_desde_tabla(tabla)
```

12.13.3 Estimación de los parámetros con las mismas funciones

Una vez ya se ha importado la tabla de frecuencias y estimado los parámetros de agrupación a partir de la tabla de frecuencias es posible usar las funciones previamente establecidas para calculas los parámetros como se muestra a continuación:

1. **Medidas de tendencia central**

   ```r
   # Calcular medidas
   tendencia_tabla <- calcular_tendencia_central(tabla, parametros_tabla)

   # Mostrar resultados
   tendencia_tabla
   ```

```
$media
[1] 3.748427

$mediana
[1] 4.306034

$moda
[1] 1.376596
```

2. **Medidas de dispersión**

```
# Calcular medidas de dispersión
dispersion_tabla <- calcular_dispersion(tabla,
                                        parametros_tabla,
                                        tendencia_tabla$media)

# Mostrar los resultados
dispersion_tabla
```

```
$rango
[1] 5.9

$varianza
[1] 3.22793

$desviacion_std
[1] 1.796644

$cv
[1] 47.93062
```

3. **Medidas de posición relativa**

```
# Calcular Q1 y P80
Q1_tabla <- calcular_posicion_relativa(tabla, parametros_tabla,
                                       1, "cuartil");Q1_tabla
```

```
[1] 1.576172
```

```
P80_tabla <- calcular_posicion_relativa(tabla, parametros_tabla,
                                        80, "percentil");P80_tabla
```

```
[1] 5.378906
```

Como se puede observar siempre y cuando la tabla de frecuencias siga el formato propuesto las funciones seguirán operando con normalidad partiendo desde una base de datos completa o únicamente desde una tabla de frecuencias, cabe resaltar que el ajustar el formato de la tabla de frecuencias cuando se trabaja con una tabla de frecuencias y no con una base de datos completa es una tarea adicional que se debe llevar a cabo previo al análisis.

Capítulo VI

Introducción a probabilidades

13 Introducción a probabilidades

La mayor parte de los problemas en estadística involucran elementos de incertidumbre, ya que usualmente no es posible determinar anticipadamente las características de una población desconocida o prever las consecuencias exactas de la toma de una decisión. Por lo tanto, es conveniente disponer de una medida que exprese esa incertidumbre en términos de una escala numérica. Esta medida es la **probabilidad** (López & González, 2018).

En el contexto agronómico, la probabilidad permite modelar y cuantificar la variabilidad inherente en los procesos biológicos, climáticos y productivos. Por ejemplo, se puede utilizar para evaluar la probabilidad de ocurrencia de plagas, el éxito de tratamientos fitosanitarios, o la variabilidad en rendimientos de cultivos.

13.1 Conceptos Fundamentales

13.1.1 Experimento y Experimento Aleatorio

Un **experimento** es el proceso mediante el cual se obtiene una observación o medida de un fenómeno. Cuando el resultado del experimento no puede preverse con certeza debido a la variabilidad inherente del fenómeno, se denomina **experimento aleatorio** (López & González, 2018).

Ejemplos de experimentos aleatorios:

1. Lanzamiento de un dado y observación del número mostrado en la cara superior
2. Lanzamiento de una moneda cuatro veces y observación del número de caras obtenido
3. Prueba de duración de una lámpara, anotando el tiempo transcurrido desde que se enciende hasta que se quema
4. Cruzamiento de animales y observación del sexo del primero que nace
5. Conteo del número de larvas de gusano cogollero en plantas de maíz
6. Conteo del número de piezas defectuosas producidas en una línea de producción durante 24 horas

13.1.2 Espacio Muestral

El **espacio muestral** es el conjunto de todos los posibles resultados de un experimento aleatorio. Se denota con el símbolo Ω (López & González, 2018).

Ejemplos de espacios muestrales:

1. Para el lanzamiento de un dado: $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. Para el lanzamiento de una moneda cuatro veces: $\Omega = \{0, 1, 2, 3, 4\}$
3. Para la duración de una lámpara: $\Omega = \{t | t \geq 0\}$
4. Para el sexo de una cría: $\Omega = \{\text{Macho, Hembra}\}$
5. Para el conteo de larvas: $\Omega = \{0, 1, 2, ...\}$

13.1.3 Evento

Un **evento** A es un subconjunto del espacio muestral Ω. En terminología de conjuntos, un evento es simplemente un conjunto de resultados posibles del experimento aleatorio. Los eventos se denotan con letras mayúsculas como A, B, C, etc. (López & González, 2018).

Ejemplos de eventos:

1. A_1: Sale un número par en el lanzamiento de un dado, $A_1 = \{2, 4, 6\}$
2. A_2: Ocurren dos caras en cuatro lanzamientos, $A_2 = \{2\}$
3. A_3: La lámpara se quema en menos de 3 horas, $A_3 = \{t | 0 \leq t < 3\}$

13.2 Métodos para Asignar Probabilidades

Independientemente del método utilizado, se deben satisfacer dos requisitos básicos:

1. Los valores de probabilidad asignados a cada resultado experimental deben estar entre 0 y 1:

$$0 \leq P(E_i) \leq 1, \text{ para toda } i$$

2. La suma de todas las probabilidades de resultados experimentales debe ser 1:

$$\sum_{i=1}^{k} P(E_i) = 1$$

13.2.1 Método Clásico

Si un evento A puede ocurrir en h maneras diferentes de un número total de n maneras posibles, todas igualmente probables, entonces la probabilidad del evento es:

$$P(A) = \frac{h}{n} = \frac{\text{número de resultados favorables}}{\text{número de resultados posibles}}$$

Ejemplo: En el lanzamiento de dos dados honestos, calcule las probabilidades de los siguientes eventos:

A: La suma de los valores es igual a 7

B: Los resultados en los dados son iguales

C: La suma de los valores es 9 o más

El espacio muestral contiene 36 resultados posibles. Contando los casos favorables:

- Para A: $(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$, entonces $P(A) = \frac{6}{36} = 0.167$
- Para B: $(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)$, entonces $P(B) = \frac{6}{36} = 0.167$
- Para C: 10 casos favorables, entonces $P(C) = \frac{10}{36} = 0.278P$

13.2.2 Método de la Frecuencia Relativa

Si después de nnn repeticiones de un experimento donde nnn es "muy grande", un evento ocurre hhh veces, entonces la probabilidad del evento es:

$$P(A) = \frac{h}{n}$$

Ejemplo: Si se lanza una moneda 1000 veces y resultan 532 caras, se puede estimar que:

$$P(\text{cara}) = \frac{532}{1000} = 0.532$$

13.2.3 Método Subjetivo

Este método está basado en el juicio personal. Se puede usar cualquier dato disponible junto con la experiencia e intuición del investigador.

13.3 Relaciones Básicas de Probabilidad

13.3.1 Complemento de un Evento

Dado un evento A, el **complemento** de A se define como el evento formado por todos los puntos muestrales que no están en A, y se representa por A^c.

$$P(A) + P(A^c) = 1$$

Por lo tanto: $P(A) = 1 - P(A^c)$

13.3.2 Ley Aditiva

La ley aditiva es útil cuando se tienen dos eventos y se desea conocer la probabilidad de que ocurra por lo menos uno de ellos. Para eventos A y B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Ejemplo: El gerente de personal de una empresa agroforestal encontró que el 30% de los empleados que salieron lo hicieron por insatisfacción salarial, el 20% por insatisfacción laboral, y el 12% por ambas razones.

Sean:

1. S: evento de salida por salario
2. W: evento de salida por trabajo

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38$$

13.3.3 Eventos Mutuamente Excluyentes

Dos eventos son **mutuamente excluyentes** si no tienen puntos muestrales en común, es decir, $P(A \cap B) = 0$.

Para eventos mutuamente excluyentes:

$$P(A \cup B) = P(A) + P(B)$$

13.4 Probabilidad Condicional

La **probabilidad condicional** de un evento A dado que ha ocurrido B se denota como $P(A|B)$ y se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ siempre que } P(B) > 0$$

Esta notación se lee como "la probabilidad de A dado B" y representa la probabilidad de que ocurra A sabiendo que B ya ha ocurrido.

Ejemplo: En una facultad de agronomía se tiene la siguiente distribución de estudiantes:

Carrera	Masculino	Femenino	Total
Agronomía	160	40	200
Forestal	30	10	40
Agroindustrial	15	10	25
Total	**205**	**60**	**265**

Dado que un alumno cursa Agronomía (A), ¿cuál es la probabilidad de que sea masculino (H)?

$$P(H|A) = \frac{P(H \cap A)}{P(A)} = \frac{160/265}{200/265} = \frac{160}{200} = 0.80$$

13.5 Eventos Independientes

Dos eventos A y B son **independientes** si:

$$P(A|B) = P(A) \text{ o } P(B|A) = P(B)$$

De lo contrario, los eventos son dependientes.

13.6 Ley Multiplicativa

Mientras que la ley aditiva se utiliza para determinar la probabilidad de una unión entre dos eventos, la **ley multiplicativa** se usa para determinar la probabilidad de una intersección de dos eventos:

$$P(A \cap B) = P(B) \cdot P(A|B)$$

o también:

$$P(A \cap B) = P(A) \cdot P(B|A)$$

13.6.1 Ley Multiplicativa para Eventos Independientes

Para eventos independientes:

$$P(A \cap B) = P(A) \cdot P(B)$$

Ejemplo: El gerente de una gasolinera sabe que el 80% de los clientes usan tarjeta de crédito. ¿Cuál es la probabilidad de que dos clientes consecutivos usen tarjeta de crédito?

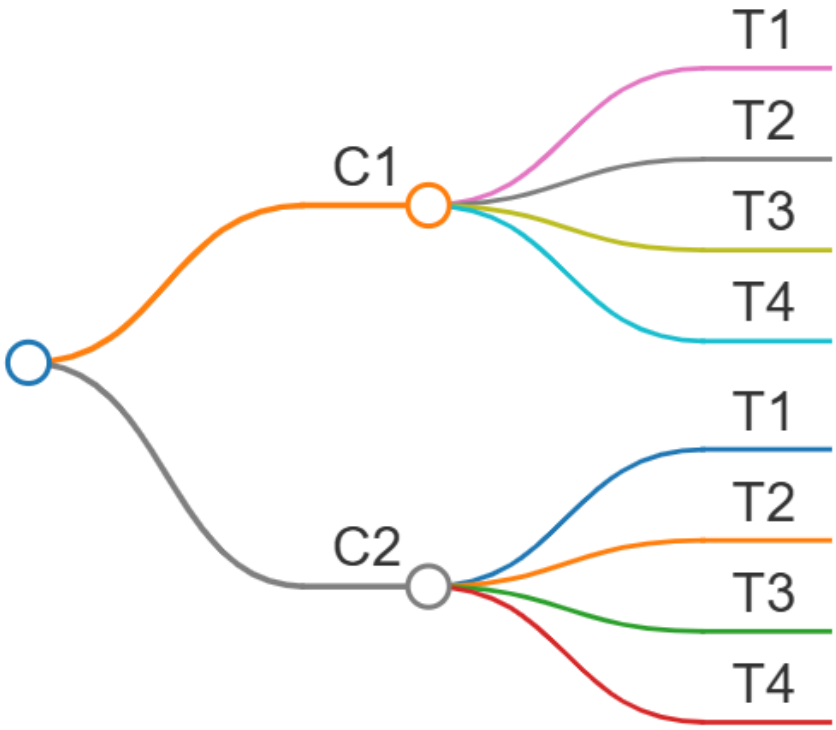
$$P(A \cap B) = P(A) \cdot P(B) = 0.8 \times 0.8 = 0.64$$

13.7 Diagramas de Árbol

Un **diagrama de árbol** es una herramienta gráfica que se emplea frecuentemente en conexión con el principio multiplicativo. Debido a su apariencia, permite visualizar todos los posibles resultados de un experimento compuesto y facilita el cálculo de probabilidades.

Ejemplo: Si un hombre tiene 2 camisas (S_1, S_2) y 4 corbatas (T_1, T_2, T_3, T_4), entonces tiene $2 \times 4 = 8$ maneras de escoger una camisa y luego una corbata.

El diagrama de árbol correspondiente sería:

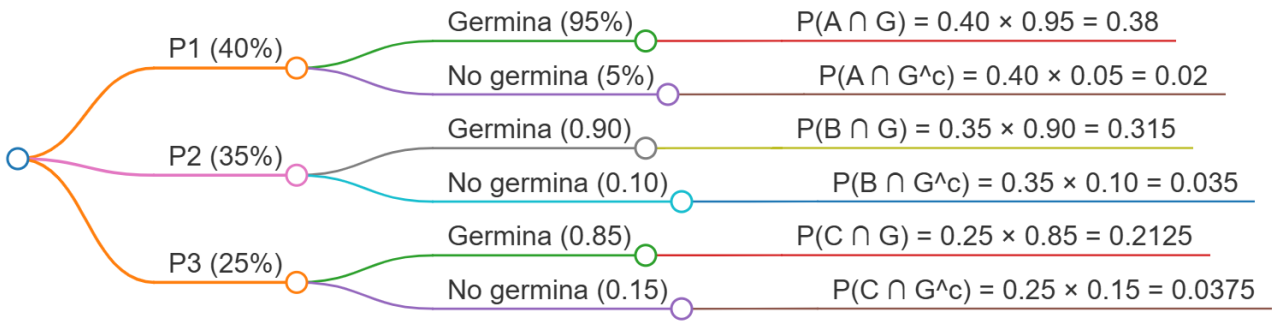


13.7.1 Uso de Diagramas de Árbol en Probabilidad Condicional

Los diagramas de árbol son especialmente útiles para problemas de probabilidad condicional, donde las probabilidades en las ramas posteriores dependen de los resultados de las ramas anteriores.

Ejemplo aplicado: Una empresa agrícola tiene tres proveedores de semillas. El proveedor A suministra el 40% de las semillas, el proveedor B el 35%, y el proveedor C el 25%. Las tasas de germinación son del 95%, 90%, y 85% respectivamente.

El diagrama de árbol sería:



13.8 Teorema de Bayes

El **teorema de Bayes** también se conoce como "probabilidad de las causas", es decir, la probabilidad de un hecho anterior sabiendo la probabilidad de un hecho posterior. Se basa en que los eventos definidos sobre un espacio muestral son particiones del mismo.

Si $A_1, A_2, A_3, ..., A_n$ son eventos mutuamente excluyentes y exhaustivos, y B es un evento observado, entonces:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{j=1}^{n} P(A_j) \cdot P(B|A_j)}$$

donde:

1. $P(A_i)$ son las probabilidades a priori
2. $P(B|A_i)$ son las probabilidades condicionales
3. $P(A_i|B)$ son las probabilidades a posteriori

Ejemplo resuelto: Una fábrica con 3 sucursales produce 40%, 35% y 25% del total de la producción. Tienen porcentajes de artículos defectuosos de 4%, 6% y 8%, respectivamente. Si se elige aleatoriamente un artículo:

a) ¿Cuál es la probabilidad de que no sea defectuoso?

Usando la ley de probabilidad total:

$$P(C) = P(A_1) \cdot P(C|A_1) + P(A_2) \cdot P(C|A_2) + P(A_3) \cdot P(C|A_3)$$

$$P(C) = 0.40 \times 0.96 + 0.35 \times 0.94 + 0.25 \times 0.92 = 0.943$$

b) Si resultó defectuoso, ¿cuál es la probabilidad de que proceda de la primera sucursal?

Primero calculamos

$$P(B) = 1 - P(C) = 1 - 0.943 = 0.057$$

Luego aplicamos Bayes:

$$P(A_1|B) = \frac{P(A_1) \cdot P(B|A_1)}{P(B)} = \frac{0.40 \times 0.04}{0.057} = 0.2807$$

Eventos	Probabilidades previas	Probabilidades condicionales	Probabilidades conjuntas	Probabilidades posteriores

13.8.1 Tabla de Análisis para el Teorema de Bayes

Eventos	Probabilidades previas	Probabilidades condicionales	Probabilidades conjuntas	Probabilidades posteriores
A_i	$P(A_i)$	$P(C)$	$P(A_i) \cdot P(C\|A_i)$	$P(A_i\|B)$
A_1	0.40	0.04	0.016	0.2807
A_2	0.35	0.06	0.021	0.3684
A_3	0.25	0.08	0.020	0.3509
Total	**1.00**		$\mathbf{P(B) = 0.057}$	**1.0000**

13.9 Notación Correcta para Probabilidades

Es fundamental utilizar la notación correcta para evitar confusiones:

1. $P(A)$: Probabilidad marginal del evento A
2. $P(A \cap B)$: Probabilidad conjunta de A y B (intersección)
3. $P(A \cup B)$: Probabilidad de la unión de A y B
4. $P(A|B)$: Probabilidad condicional de A dado B
5. $P(A^c)$: Probabilidad del complemento de A
6. $A \perp B$: A y B son independientes

Capítulo VII

Distribuciones de probabilidad discretas

14 Distribuciones Binomial y Poisson en R

14.1 Introducción

El estudio de las distribuciones de probabilidad constituye uno de los pilares fundamentales de la estadística aplicada. En el contexto de las ciencias agronómicas, estas herramientas matemáticas permiten modelar y analizar fenómenos aleatorios que ocurren frecuentemente en la investigación y práctica agrícola. Las distribuciones binomial y de Poisson, como distribuciones discretas, son especialmente útiles para describir eventos de conteo, tales como el número de semillas germinadas, la cantidad de plagas observadas en una parcela, o la ocurrencia de eventos climáticos adversos.

Según López y González (2018), las variables aleatorias discretas son aquellas que pueden tomar un conjunto finito o numerable de valores, generalmente asociados a conteos de eventos. El software estadístico R proporciona funciones específicas para el cálculo de probabilidades en estas distribuciones, facilitando el análisis estadístico y la toma de decisiones basada en evidencia.

14.2 Distribución Binomial

14.2.1 Características y definición

La distribución binomial describe el número de éxitos obtenidos en una secuencia de ensayos independientes, donde cada ensayo presenta únicamente dos posibles resultados: éxito o fracaso. López y González (2018) establecen que un experimento binomial posee las siguientes características fundamentales:

1. Consta de n ensayos o pruebas idénticas (ensayos de Bernoulli)
2. Cada prueba puede tener uno de dos resultados posibles (éxito o fracaso)
3. La probabilidad de un éxito en una sola prueba es igual a p, y permanece constante de una prueba a otra. La probabilidad de fracaso es igual a $(1-p)$ y se denota con la letra q
4. El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente

La distribución binomial se representa como $B(n, p)$, siendo n y p los parámetros de dicha distribución.

14.2.2 Función de probabilidad

La función de probabilidad de la distribución binomial se expresa matemáticamente como:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

donde:

1. X es la variable aleatoria que representa el número de éxitos
2. x es el número de éxitos observados
3. n es el número total de ensayos
4. p es la probabilidad de éxito en cada ensayo
5. $q = 1 - p$ es la probabilidad de fracaso
6. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ es el coeficiente binomial

14.2.3 Parámetros de la distribución binomial

Los parámetros de tendencia central y dispersión de la distribución binomial son:

1. Media: $E(X) = np$
2. Varianza: $V(X) = npq$

14.3 Cálculo de probabilidades binomiales en R

El software R proporciona funciones específicas para el cálculo de probabilidades binomiales. A continuación se describen las funciones principales y sus argumentos.

14.3.1 Función para calcular P(X = x)

Para calcular la probabilidad de obtener exactamente xxx éxitos, se utiliza la función:

```
dbinom(x, size, prob)
```

Argumentos en orden:

1. `x`: número de éxitos deseados (valor específico de la variable aleatoria)
2. `size`: número total de ensayos (n)
3. `prob`: probabilidad de éxito en cada ensayo (p)

14.3.2 Función para calcular P(X x) y P(X > x)

Para calcular probabilidades acumuladas, se utiliza la función:

```
pbinom(q, size, prob, lower.tail)
```

Argumentos en orden:

1. `q`: valor hasta el cual se desea calcular la probabilidad acumulada
2. `size`: número total de ensayos (nnn)
3. `prob`: probabilidad de éxito en cada ensayo (ppp)
4. `lower.tail`: argumento lógico que indica si se calcula $P(X \leq x)$(`TRUE`, $por defecto$) o $P(X > x)$(`FALSE`)

14.3.3 Ejemplo práctico: Germinación de semillas

Supóngase que se siembran 20 semillas de maíz y se sabe que la probabilidad de germinación de cada semilla es de 0.8. Se desea calcular las siguientes probabilidades:

14.3.3.1 Caso 1: P(X = 16) - Probabilidad de que germinen exactamente 16 semillas

```
dbinom(16, 20, 0.8)
```

```
[1] 0.2181994
```

14.3.3.2 Caso 2: P(X 15) - Probabilidad de que germinen 15 o menos semillas

```
pbinom(15, 20, 0.8)
```

```
[1] 0.3703517
```

14.3.3.3 Caso 3: P(X > 18) - Probabilidad de que germinen más de 18 semillas

```
pbinom(18, 20, 0.8, lower.tail = FALSE)
```

```
[1] 0.06917529
```

14.4 Distribución de Poisson

14.4.1 Características y definición

La distribución de Poisson, desarrollada por Simeón Dennis Poisson (1781-1840), es un modelo apropiado para describir el número de eventos raros que ocurren en un intervalo de tiempo o espacio específico. López y González (2018) indican que esta distribución es útil para modelar eventos con las siguientes características:

1. Los eventos ocurren de manera independiente
2. La tasa promedio de ocurrencia permanece constante
3. Los eventos son raros o poco frecuentes

14.4.2 Función de probabilidad

Una variable aleatoria X tiene distribución de Poisson con parámetro $\lambda > 0$, si su función de probabilidad está dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ...$$

donde:

1. λ representa el número medio de ocurrencias por intervalo de tiempo
2. $e = 2.71828$ es la base de los logaritmos naturales

La notación utilizada es: $X \sim Po(\lambda)$

14.4.3 Parámetros de la distribución de Poisson

Los parámetros de la distribución de Poisson son:

1. Media:$E(X) = \lambda$
2. Varianza: $V(X) = \lambda$

14.5 Cálculo de probabilidades de Poisson en R

14.5.1 Función para calcular P(X = x)

Para calcular la probabilidad de observar exactamente x eventos, se utiliza:

```
dpois(x, lambda)
```

Argumentos en orden:

1. `x`: número de eventos observados
2. `lambda`: tasa promedio de ocurrencia (λ)

14.5.2 Función para calcular P(X x) y P(X > x)

Para probabilidades acumuladas, se utiliza:

```
ppois(q, lambda, lower.tail)
```

Argumentos en orden:

1. `q`: valor hasta el cual se desea calcular la probabilidad acumulada
2. `lambda`: tasa promedio de ocurrencia ( )
3. `lower.tail`: argumento lógico para $P(X \leq x)$ (`TRUE`) o $P(X > x)$ (`FALSE`)

14.5.3 Ejemplo práctico: Incidencia de plagas

Supóngase que en un cultivo de tomate se observa un promedio de 3 plagas por metro cuadrado. Se desea calcular las siguientes probabilidades:

14.5.3.1 Caso 1: P(X = 5) - Probabilidad de observar exactamente 5 plagas

```
dpois(5, 3)
```

```
[1] 0.1008188
```

14.5.3.2 Caso 2: P(X 2) - Probabilidad de observar 2 o menos plagas

```
ppois(2, 3)
```

```
[1] 0.4231901
```

14.5.3.3 Caso 3: P(X > 4) - Probabilidad de observar más de 4 plagas

```
ppois(4, 3, lower.tail = FALSE)
```

```
[1] 0.1847368
```

14.6 Interpretación y aplicaciones en agronomía

Las distribuciones binomial y de Poisson encuentran numerosas aplicaciones en el campo agronómico. La distribución binomial es particularmente útil para modelar situaciones donde se evalúa el éxito o fracaso de un proceso, como la germinación de semillas, la supervivencia de plantas trasplantadas, o la efectividad de tratamientos fitosanitarios.

Por su parte, la distribución de Poisson es apropiada para modelar la ocurrencia de eventos raros, tales como la aparición de plagas específicas, la incidencia de enfermedades en cultivos, o la ocurrencia de eventos climáticos extremos.

Capítulo VIII

Distribución normal

15 Distribución normal

15.1 Introducción

La distribución normal, también conocida como distribución gaussiana o campana de Gauss, es una de las distribuciones de probabilidad continua más importantes en estadística. Su relevancia radica en que muchos fenómenos naturales y sociales tienden a seguir esta distribución, y además, sirve como base para numerosas pruebas y modelos estadísticos.

Según López y González (2018), la distribución normal es fundamental en bioestadística debido a que muchas variables biométricas tienden a distribuirse normalmente, la distribución de las medias muestrales de una variable cualquiera tiende a ser normal (Teorema del Límite Central), y muchas pruebas estadísticas asumen la normalidad de los datos.

15.2 Características y definición

La distribución normal se caracteriza por ser simétrica y tener forma de campana. Está completamente definida por dos parámetros: la media (μ) y la desviación estándar (σ). La función de densidad de probabilidad de la distribución normal se expresa como:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

donde:

1. x es la variable aleatoria continua
2. μ es la media de la distribución
3. σ es la desviación estándar de la distribución
4. e es la base del logaritmo natural (aproximadamente 2.71828)
5. π es la constante pi (aproximadamente 3.14159)

La notación utilizada es: $X \sim N(\mu, \sigma^2)$, donde μ es la media y σ^2 es la varianza.

15.2.1 Propiedades de la distribución normal

López y González (2018) destacan las siguientes propiedades de la distribución normal:

1. Existe una familia de distribuciones normales, cada una definida por su media (μ) y desviación estándar (σ).
2. El punto más alto de la curva normal es la media, que coincide con la mediana y la moda.
3. La distribución es simétrica alrededor de la media.
4. Los extremos de la distribución se extienden indefinidamente sin tocar el eje horizontal.
5. La desviación estándar (σ) determina el ancho de la curva; valores mayores indican mayor dispersión.
6. El área total bajo la curva es igual a 1.
7. Las probabilidades se determinan mediante áreas bajo la curva.
8. La regla empírica establece que aproximadamente el 68% de las observaciones se encuentran dentro de una desviación estándar de la media ($\mu \pm \sigma$), el 95% dentro de dos desviaciones estándar ($\mu \pm 2\sigma$), y el 99.7% dentro de tres desviaciones estándar ($\mu \pm 3\sigma$).

15.3 Cálculo de probabilidades normales en R

El software R proporciona funciones para calcular probabilidades asociadas a la distribución normal.

15.3.1 Función para calcular la función de densidad de probabilidad

Para calcular la función de densidad de probabilidad en un punto x, se utiliza la función:

```
dnorm(x, mean, sd)
```

Argumentos en orden:

1. `x`: valor de la variable aleatoria en el que se evalúa la función de densidad
2. `mean`: media de la distribución (μ)
3. `sd`: desviación estándar de la distribución (σ)

15.3.2 Función para calcular probabilidades acumuladas

Para calcular la probabilidad acumulada $P(X \leq x)$, se utiliza la función:

$$\texttt{pnorm(q, mean, sd, lower.tail)}$$

Argumentos en orden:

1. `q`: valor hasta el cual se desea calcular la probabilidad acumulada
2. `mean`: media de la distribución (μ)
3. `sd`: desviación estándar de la distribución (σ)
4. `lower.tail`: argumento lógico que indica si se calcula $P(X \leq x)$ (`TRUE`, $por defecto$) o $P(X > x)$ (`FALSE`)

15.3.3 Ejemplo práctico: Estatura de estudiantes

Supóngase que la estatura de los estudiantes de una universidad se distribuye normalmente con una media de 170 cm y una desviación estándar de 10 cm. Se desea calcular las siguientes probabilidades:

15.3.3.1 Caso 1: P(X 180) - Probabilidad de que un estudiante mida 180 cm o menos

```
pnorm(180, 170, 10)
```

```
[1] 0.8413447
```

15.3.3.2 Caso 2: P(X $>$ 160) - Probabilidad de que un estudiante mida más de 160 cm

```
pnorm(160, 170, 10, lower.tail = FALSE)
```

```
[1] 0.8413447
```

15.3.3.3 Caso 3: P(165 X 175) - Probabilidad de que un estudiante mida entre 165 cm y 175 cm

Para calcular esta probabilidad, se resta la probabilidad acumulada hasta 165 cm de la probabilidad acumulada hasta 175 cm:

```r
pnorm(175, 170, 10) - pnorm(165, 170, 10)
```

```
[1] 0.3829249
```

15.4 Estandarización de la variable normal

15.4.1 Ejemplo práctico: Duración de la temporada de heladas en Guatemala

El Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología (INSIVUMEH) de Guatemala ha determinado que la duración de la temporada de heladas sigue una distribución normal. Se conoce la siguiente información:

1. La duración promedio de la temporada de heladas es de 120 días ($\mu = 120$)
2. La probabilidad de que la temporada dure más de 133 días es del 25.78% ($P(X > 133) = 0.2578$)

Objetivo: Determinar la desviación estándar (σ) de la distribución normal.

15.4.1.1 Paso 1: Estandarización de la variable

Para resolver este problema, se debe estandarizar la variable X (duración de la temporada de heladas) utilizando la transformación a Z:

$$Z = \frac{X - \mu}{\sigma}$$

donde:

1. $X = 133$ días
2. $\mu = 120$ días
3. $\sigma =$ desviación estándar (valor a determinar)

Sustituyendo los valores conocidos:

$$Z = \frac{133 - 120}{\sigma} = \frac{13}{\sigma}$$

15.4.1.2 Paso 2: Cálculo de la probabilidad acumulada

Dado que P(X > 133) = 0.2578, se puede determinar la probabilidad acumulada hasta 133:

$$P(X \leq 133) = 1 - P(X > 133) = 1 - 0.2578 = 0.7422$$

Por lo tanto:

$$P\left(Z \leq \frac{13}{\sigma}\right) = 0.7422) = 0.7422$$

15.4.1.3 Paso 3: Encontrar el valor Z correspondiente

Se debe encontrar el valor z tal que $P(Z \leq z) = 0.7422$ en la distribución normal estándar.

En R, se utiliza la función:

`qnorm(p, mean, sd)`

Argumentos:

1. `p`: probabilidad acumulada deseada
2. `mean`: media de la distribución (0 para la normal estándar)
3. `sd`: desviación estándar de la distribución (1 para la normal estándar)

```
qnorm(0.7422, mean = 0, sd = 1)
```

```
[1] 0.6501428
```

15.4.1.4 Paso 4: Despejar la desviación estándar

Igualando la expresión estandarizada con el valor z encontrado:

$$\frac{13}{\sigma} = 0.65$$

Despejando:

$$\sigma = \frac{13}{0.65} = 20$$

15.4.1.5 Verificación en R

Para verificar el resultado, se puede calcular la probabilidad $P(X > 133)$ con los parámetros encontrados:

```r
pnorm(133, mean = 120, sd = 20, lower.tail = FALSE)
```

```
[1] 0.2578461
```

Este resultado confirma que la desviación estándar calculada es correcta.

15.4.1.6 Interpretación

La desviación estándar de la duración de la temporada de heladas en Guatemala es de 20 días ($\sigma = 20$). Esto significa que la duración de la temporada de heladas varía alrededor de la media (120 días) con una dispersión de 20 días.

Con esta información, se puede establecer que la duración de la temporada de heladas en Guatemala sigue una distribución $N(120, 20^2)$, lo que permite realizar predicciones y análisis probabilísticos para la planificación agrícola y la gestión de riesgos climáticos.

15.5 Interpretación y aplicaciones en agronomía

La distribución normal es ampliamente utilizada en agronomía para modelar variables continuas como la altura de las plantas, el rendimiento de los cultivos, el peso de los frutos, y las temperaturas. Permite realizar inferencias estadísticas, como la estimación de intervalos de confianza y la realización de pruebas de hipótesis, que son fundamentales para la investigación y la toma de decisiones en el sector agropecuario.

Capítulo IX

Intervalos de confianza

16 Estimación puntual e intervalos de confianza en R

16.1 Introducción

La estimación de parámetros poblacionales a partir de muestras es una de las tareas fundamentales en la estadística aplicada, especialmente en la investigación agronómica. En la toma de decisiones sobre producción, selección de variedades o evaluación de innovaciones tecnológicas, es común que el investigador disponga únicamente de datos muestrales. Por ello, resulta esencial contar con herramientas que permitan inferir, con un nivel de confianza conocido, los valores verdaderos de la población a partir de la información obtenida en el laboratorio o en campo (López & González, 2018).

El uso de intervalos de confianza permite cuantificar la incertidumbre inherente a la estimación de parámetros, como la media o la varianza, y facilita la comunicación de resultados de manera rigurosa y transparente. El software R proporciona funciones específicas para calcular estimaciones puntuales e intervalos de confianza, lo que agiliza el análisis estadístico y la interpretación de los datos en contextos agronómicos.

16.2 Fundamentos teóricos

16.2.1 Estimación puntual

La estimación puntual consiste en asignar un único valor numérico, calculado a partir de los datos muestrales, como mejor aproximación del parámetro poblacional de interés. Por ejemplo, la media muestral ($\bar{x}$) se utiliza como estimador puntual de la media poblacional (μ).

16.2.2 Intervalo de confianza

Un intervalo de confianza es un rango de valores, calculado a partir de los datos muestrales, que con una determinada probabilidad (nivel de confianza) contiene al verdadero valor del parámetro poblacional. Matemáticamente, para la media poblacional, el intervalo de confianza se expresa como:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

1. $\bar{x}$ es la media muestral,

2. $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar para el nivel de confianza deseado,
3. σ es la desviación estándar poblacional (o muestral, si σ es desconocida),
4. n es el tamaño de la muestra.

Cuando la desviación estándar poblacional es desconocida y el tamaño de la muestra es pequeño (n < 30), se utiliza la distribución t de Student en lugar de la normal estándar.

16.2.3 Nivel de confianza y significancia

El nivel de confianza $(1 - \alpha)$ representa la probabilidad de que el intervalo calculado contenga al verdadero parámetro poblacional. Comúnmente, se utilizan niveles de confianza del 90%, 95% o 99%. El valor α representa la significancia, es decir, la probabilidad de que el intervalo no contenga al parámetro poblacional.

Factores que afectan la amplitud del intervalo

La amplitud del intervalo de confianza está influenciada por varios factores:

1. **Tamaño de la muestra (n):** A mayor tamaño de la muestra, menor es la amplitud del intervalo.
2. **Desviación estándar (σ) :** A mayor variabilidad en los datos, mayor es la amplitud del intervalo.
3. **Nivel de confianza ($1-\alpha$):** A mayor nivel de confianza, mayor es la amplitud del intervalo.

16.3 Formulas para el calculo de intervalos de confianza

16.3.1 Intervalos de confianza para la media con desviación estándar conocida

Cuando la desviación estándar de la población (σ) es conocida, el intervalo de confianza para la media poblacional (μ) se calcula utilizando la distribución normal estándar (z):

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Para automatizar este proceso en R se puede emplear la siguiente formula personalizada:

```
# Función personalizada para intervalo de confianza (sigma conocida)
ic_media_sigma <- function(x_barra,
                           sigma,
                           n,
                           confianza = 0.95) {
  # Cálculos
  error_estandar <- sigma / sqrt(n)
  alpha <- 1 - confianza
```

```r
  z_critico <- qnorm(1 - alpha/2)
  margen_error <- z_critico * error_estandar

  # Límites del intervalo
  limite_inf <- x_barra - margen_error
  limite_sup <- x_barra + margen_error

  # Resultados organizados
  resultados <- list(
    media_muestra = x_barra,
    error_estandar = error_estandar,
    z_critico = z_critico,
    margen_error = margen_error,
    limite_inferior = limite_inf,
    limite_superior = limite_sup,
    intervalo = c(limite_inf, limite_sup),
    confianza = confianza * 100
  )

  # Mostrar resultados
  cat("=== INTERVALO DE CONFIANZA PARA LA MEDIA ===\n")
  cat("Desviación estándar poblacional conocida\n\n")
  cat("Datos:\n")
  cat("- Media muestral:", x_barra, "\n")
  cat("- Desviación estándar poblacional:", sigma, "\n")
  cat("- Tamaño de muestra:", n, "\n")
  cat("- Nivel de confianza:", confianza*100, "%\n\n")
  cat("Cálculos:\n")
  cat("- Error estándar:", round(error_estandar, 4), "\n")
  cat("- Valor z crítico:", round(z_critico, 4), "\n")
  cat("- Margen de error:", round(margen_error, 4), "\n\n")
  cat("RESULTADO:\n")
  cat("IC al", confianza*100, "%: [", round(limite_inf, 4),
      ",", round(limite_sup, 4), "]\n")

  return(invisible(resultados))
}
```

Esta función cuenta con la siguiente sintaxis para su uso:

ic_media_sigma(x_barra, sigma, n, confianza)

Argumentos en orden:

1. x_barra: Media muestral
2. sigma: Desviación estandar poblacional conocida
3. n: Tamaño de la muestra

4. confianza: Nivel de confianza

16.3.2 Intervalos de confianza para la media con desviación estándar desconocida

Cuando la desviación estándar de la población (σ) es desconocida, se utiliza la desviación estándar muestral (s) como estimación. La elección de la distribución apropiada para calcular el intervalo de confianza depende del tamaño de la muestra:

Criterio de selección de distribución:

Para muestras pequeñas (n $<$ 30): Se utiliza la distribución t de Student:

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

donde $t_{\alpha/2,n-1}$ es el valor crítico de la distribución t de Student con $n-1$ grados de libertad.

Para muestras grandes (n 30): Se puede utilizar la distribución normal estándar (Z)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

donde $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar.

Para automatizar este proceso en R se puede emplear la siguiente formula personalizada:

```
# Función robusta que decide automáticamente entre Z y t
ic_media_s <- function(datos = NULL,
                       x_barra = NULL,
                       s = NULL,
                       n = NULL,
                       confianza = 0.95) {

  # Si se proporcionan los datos directamente
  if (!is.null(datos)) {
    n <- length(datos)
    x_barra <- mean(datos)
    s <- sd(datos)
  }

  # Verificar que tenemos todos los parámetros necesarios
  if (is.null(x_barra) || is.null(s) || is.null(n)) {
    stop("Debe proporcionar los datos o los valores de x_barra, s y n")
  }

  # Decidir qué distribución usar
  usar_z <- n >= 30
```

```r
# Cálculos comunes
alpha <- 1 - confianza
error_estandar <- s / sqrt(n)

if (usar_z) {
  # Usar distribución Z
  valor_critico <- qnorm(1 - alpha/2)
  distribucion <- "Z (Normal estándar)"
  gl <- NA
} else {
  # Usar distribución t
  gl <- n - 1
  valor_critico <- qt(1 - alpha/2, gl)
  distribucion <- "t de Student"
}

margen_error <- valor_critico * error_estandar

# Límites del intervalo
limite_inf <- x_barra - margen_error
limite_sup <- x_barra + margen_error

# Resultados organizados
resultados <- list(
  datos = if(!is.null(datos)) datos else "No proporcionados",
  n = n,
  media_muestra = x_barra,
  desv_estandar_muestra = s,
  distribucion_usada = distribucion,
  grados_libertad = if(usar_z) NA else gl,
  error_estandar = error_estandar,
  valor_critico = valor_critico,
  margen_error = margen_error,
  limite_inferior = limite_inf,
  limite_superior = limite_sup,
  intervalo = c(limite_inf, limite_sup),
  confianza = confianza * 100
)

# Mostrar resultados
cat("=== INTERVALO DE CONFIANZA PARA LA MEDIA ===\n")
cat("Desviación estándar poblacional desconocida\n")
cat("Distribución utilizada:", distribucion, "\n")
cat("Criterio: n", if(usar_z) " " else "<", "30\n\n")

if (!is.null(datos)) {
  cat("Datos originales:\n")
  if (length(datos) <= 20) {
    cat(paste(datos, collapse = ", "), "\n\n")
```

```
    } else {
      cat("Muestra de", length(datos), "observaciones\n\n")
    }
  }

  cat("Estadísticos calculados:\n")
  cat("- Tamaño de muestra (n):", n, "\n")
  cat("- Media muestral (x̄):", round(x_barra, 4), "\n")
  cat("- Desviación estándar muestral (s):", round(s, 4), "\n")
  if (!usar_z) cat("- Grados de libertad:", gl, "\n")
  cat("- Nivel de confianza:", confianza*100, "%\n\n")

  cat("Cálculos del intervalo:\n")
  cat("- Error estándar:", round(error_estandar, 4), "\n")
  cat("- Valor", if(usar_z) "z" else "t", "crítico:",
      round(valor_critico, 4), "\n")
  cat("- Margen de error:", round(margen_error, 4), "\n\n")

  cat("RESULTADO:\n")
  cat("IC al", confianza*100, "%: [", round(limite_inf, 4),
      ",", round(limite_sup, 4), "]\n\n")

  return(invisible(resultados))
}
```

Esta función cuenta con la siguiente sintaxis para su uso:

ic_media_s(datos, confianza)

Argumentos en orden:

1. datos: Vector con los datos muestrales
2. confianza: Nivel de confianza

También tiene la siguiente sintaxis cuando no se cuenta con los datos de la muestra directamente:

ic_media_s(x_barra, s, n, confianza)

Argumentos en orden:

1. x_barra: Media muestral
2. s: Desviación estandar muestral
3. n: Tamaño de la muestra
4. confianza: Nivel de confianza

16.3.3 Intervalos de confianza para la varianza

El intervalo de confianza para la varianza poblacional (σ^2) se calcula utilizando la distribución chi-cuadrado (χ^2):

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

donde:

1. s^2 es la varianza muestral,
2. $\chi^2_{\alpha/2,n-1}$ y $\chi^2_{1-\alpha/2,n-1}$ son los valores críticos de la distribución chi-cuadrado con $n-1$ grados de libertad.

Para automatizar este proceso en R se puede emplear la siguiente formula personalizada:

```
# Función personalizada para intervalo de confianza de varianza
ic_varianza <- function(datos = NULL,
                        s = NULL,
                        n = NULL,
                        confianza = 0.95) {

  # Si se proporcionan los datos directamente
  if (!is.null(datos)) {
    n <- length(datos)
    s <- sd(datos)
  }

  # Verificar que tenemos todos los parámetros necesarios
  if (is.null(s) || is.null(n)) {
    stop("Debe proporcionar los datos o los valores de s y n")
  }

  # Cálculos básicos
  s2 <- s^2  # varianza muestral
  gl <- n - 1  # grados de libertad
  alpha <- 1 - confianza

  # Valores críticos de chi-cuadrado
  chi2_inf <- qchisq(alpha/2, gl)        # límite inferior
  chi2_sup <- qchisq(1 - alpha/2, gl)    # límite superior

  # Intervalos de confianza para la varianza
  ic_var_inf <- (gl * s2) / chi2_sup
  ic_var_sup <- (gl * s2) / chi2_inf

  # Intervalos de confianza para la desviación estándar
  ic_sd_inf <- sqrt(ic_var_inf)
  ic_sd_sup <- sqrt(ic_var_sup)
```

```r
# Resultados organizados
resultados <- list(
  datos = if(!is.null(datos)) datos else "No proporcionados",
  n = n,
  desv_estandar_muestra = s,
  varianza_muestra = s2,
  grados_libertad = gl,
  chi2_inferior = chi2_inf,
  chi2_superior = chi2_sup,
  ic_varianza = c(ic_var_inf, ic_var_sup),
  ic_desv_estandar = c(ic_sd_inf, ic_sd_sup),
  confianza = confianza * 100
)

# Mostrar resultados
cat("=== INTERVALO DE CONFIANZA PARA LA VARIANZA ===\n")
cat("Distribución Chi-cuadrado\n\n")

if (!is.null(datos)) {
  cat("Datos originales:\n")
  if (length(datos) <= 20) {
    cat(paste(datos, collapse = ", "), "\n\n")
  } else {
    cat("Muestra de", length(datos), "observaciones\n\n")
  }
}

cat("Estadísticos calculados:\n")
cat("- Tamaño de muestra (n):", n, "\n")
cat("- Desviación estándar muestral (s):", round(s, 4), "\n")
cat("- Varianza muestral (s²):", round(s2, 4), "\n")
cat("- Grados de libertad:", gl, "\n")
cat("- Nivel de confianza:", confianza*100, "%\n")
cat("- Nivel de significancia ( ):", alpha, "\n\n")

cat("Valores críticos de Chi-cuadrado:\n")
cat("-  ²", alpha/2, ",", gl, "=", round(chi2_inf, 4), "\n")
cat("-  ²", 1-alpha/2, ",", gl, "=", round(chi2_sup, 4), "\n\n")

cat("Cálculos del intervalo:\n")
cat("- Límite inferior varianza: (", gl, "×", round(s2,1), ") /",
    round(chi2_sup,3), "=", round(ic_var_inf, 1), "\n")
cat("- Límite superior varianza: (", gl, "×", round(s2,1), ") /",
    round(chi2_inf,3), "=", round(ic_var_sup, 1), "\n\n")

cat("RESULTADOS:\n")
cat("IC al", confianza*100, "% para  ²: [", round(ic_var_inf, 1),
    ",", round(ic_var_sup, 1), "] ")
```

```
  cat("IC al", confianza*100, "% para  :  [", round(ic_sd_inf, 2),
      ",", round(ic_sd_sup, 2), "] ")

  return(invisible(resultados))
}
```

Esta función cuenta con la siguiente sintaxis para su uso:

ic_varianza(datos, confianza)

Argumentos en orden:

1. datos: Vector con los datos de la muestra
2. confianza: Nivel de confianza

También tiene la siguiente sintaxis cuando no se cuenta con los datos de la muestra directamente:

ic_varianza(s, n, confianza)

Argumentos en orden:

1. s: Desviación estándar muestral
2. n: Tamaño de la muestra
3. confianza: Nivel de confianza

16.3.4 Intervalos de confianza para la proporción

El intervalo de confianza para la proporción poblacional (p) se calcula como:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde:

1. $\hat{p}$ es la proporción muestral,
2. $z_{\alpha/2}$ es el valor crítico de la distribución normal estándar.

Para automatizar este proceso en R se puede emplear la siguiente formula personalizada:

```r
# Función personalizada para intervalo de confianza de una proporción
ic_proporcion <- function(x,
                          n,
                          confianza = 0.95) {
  p_hat <- x / n
  alpha <- 1 - confianza
  z_critico <- qnorm(1 - alpha/2)
  error_estandar <- sqrt(p_hat * (1 - p_hat) / n)
  margen_error <- z_critico * error_estandar
  limite_inf <- p_hat - margen_error
  limite_sup <- p_hat + margen_error

  # Resultados organizados
  resultados <- list(
    proporcion_muestral = p_hat,
    error_estandar = error_estandar,
    z_critico = z_critico,
    margen_error = margen_error,
    limite_inferior = limite_inf,
    limite_superior = limite_sup,
    intervalo = c(limite_inf, limite_sup),
    confianza = confianza * 100
  )

  # Mostrar resultados
  cat("=== INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN ===\n")
  cat("Datos:\n")
  cat("- Éxitos (x):", x, "\n")
  cat("- Tamaño de muestra (n):", n, "\n")
  cat("- Proporción muestral (p̂):", round(p_hat, 4), "\n")
  cat("- Nivel de confianza:", confianza*100, "%\n\n")

  cat("Cálculos:\n")
  cat("- Error estándar:", round(error_estandar, 4), "\n")
  cat("- Valor z crítico:", round(z_critico, 4), "\n")
  cat("- Margen de error:", round(margen_error, 4), "\n\n")

  cat("RESULTADO:\n")
  cat("IC al", confianza*100, "%: [", round(limite_inf, 4),
      ",", round(limite_sup, 4), "]\n")

  return(invisible(resultados))
}
```

Esta función cuenta con la siguiente sintaxis para su uso:

ic_proporcion(x, n, confianza)

Argumentos en orden:

1. x: numero de observaciones “exitosas” en la muestra
2. n: Tamaño de la muestra
3. confianza: Nivel de confianza

16.4 Ejemplos de cálculo de intervalos de confianza en R

16.4.1 Ejemplo 1: Intervalo de confianza para la media con desviación estándar conocida

Contexto agronómico: Una empresa productora de semillas de maíz conoce que la desviación estándar del peso de las semillas es de 0.15 gramos. Se toma una muestra aleatoria de 25 semillas y se obtiene un peso promedio de 0.85 gramos. Se desea construir un intervalo de confianza del 95% para el peso promedio poblacional.

Cálculo manual:

$$
\begin{aligned}
n &= 25 \\
\bar{x} &= 0.85 \text{ g} \\
\sigma &= 0.15 \text{ g} \\
\alpha &= 0.05 \\
z_{\alpha/2} &= z_{0.025} = 1.96
\end{aligned}
$$

Cálculo del error estándar:

$$\text{Error estándar} = \frac{\sigma}{\sqrt{n}} = \frac{0.15}{\sqrt{25}} = \frac{0.15}{5} = 0.03$$

Margen de error:

$$\text{Margen de error} = z_{\alpha/2} \times \text{Error estándar} = 1.96 \times 0.03 = 0.0588$$

Intervalo de confianza al 95%:

$$
\begin{aligned}
IC_{95\%} &= \bar{x} \pm \text{Margen de error} \\
IC_{95\%} &= 0.85 \pm 0.0588 \\
IC_{95\%} &= [0.7912,\ 0.9088]
\end{aligned}
$$

Implementación en R:

```
# Uso de la función con los datos del ejemplo
resultado <- ic_media_sigma(x_barra = 0.85,
                            sigma = 0.15,
                            n = 25,
                            confianza = 0.95)
```

```
=== INTERVALO DE CONFIANZA PARA LA MEDIA ===
Desviación estándar poblacional conocida

Datos:
- Media muestral: 0.85
- Desviación estándar poblacional: 0.15
- Tamaño de muestra: 25
- Nivel de confianza: 95 %

Cálculos:
- Error estándar: 0.03
- Valor z crítico: 1.96
- Margen de error: 0.0588

RESULTADO:
IC al 95 %: [ 0.7912 , 0.9088 ]
```

16.4.2 Ejemplo 2: Intervalo de confianza para la media con desviación estándar desconocida

Contexto agronómico: Un investigador desea estimar la altura promedio de plantas de frijol a los 30 días después de la siembra. Se selecciona una muestra aleatoria de 15 plantas y se registran las siguientes alturas en centímetros:

12.5	14.2	13.8
15.1	12.9	14.7
13.3	14.9	13.6
14.4	12.8	15.3
13.9	14.1	13.7

Datos:

$$
\begin{aligned}
n &= 15 \\
\bar{x} &= 13.89 \text{ cm} \\
s &= 0.85 \text{ cm} \\
\alpha &= 0.05
\end{aligned}
$$

Grados de libertad:

$$gl = n - 1 = 15 - 1 = 14$$

Valor crítico de t:

$$t_{\alpha/2,14} = t_{0.025,14} = 2.145$$

Error estándar:

$$\text{Error estándar} = \frac{s}{\sqrt{n}} = \frac{0.85}{\sqrt{15}} = 0.2195$$

Margen de error:

$$\text{Margen de error} = t_{\alpha/2,14} \times \text{Error estándar} = 2.145 \times 0.2195 = 0.4708$$

Intervalo de confianza al 95%:

$$IC_{95\%} = \bar{x} \pm \text{Margen de error} = 13.89 \pm 0.4708 = [13.42,\ 14.36]$$

Implementación en R:

Usando la función por defecto de R:

```r
# Datos del problema
alturas <- c(12.5, 14.2, 13.8, 15.1, 12.9, 14.7, 13.3, 14.9,
             13.6, 14.4, 12.8, 15.3, 13.9, 14.1, 13.7)

# Cálculo directo con función incorporada
resultado <- t.test(alturas, conf.level = 0.95)
print(resultado)
```

```
    One Sample t-test

data:  alturas
t = 63.729, df = 14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 13.47730 14.41604
sample estimates:
mean of x
 13.94667
```

Usando la función personalizada:

```r
# Opción 1: Pasando los datos directamente
resultado1 <- ic_media_s(datos = alturas,
                         confianza = 0.95)
```

```
=== INTERVALO DE CONFIANZA PARA LA MEDIA ===
Desviación estándar poblacional desconocida
Distribución utilizada: t de Student
Criterio: n < 30

Datos originales:
```

```
12.5, 14.2, 13.8, 15.1, 12.9, 14.7, 13.3, 14.9, 13.6, 14.4, 12.8, 15.3, 13.9, 14.1, 13.7

Estadísticos calculados:
- Tamaño de muestra (n): 15
- Media muestral (x̄): 13.9467
- Desviación estándar muestral (s): 0.8476
- Grados de libertad: 14
- Nivel de confianza: 95 %

Cálculos del intervalo:
- Error estándar: 0.2188
- Valor t crítico: 2.1448
- Margen de error: 0.4694

RESULTADO:
IC al 95 %: [ 13.4773 , 14.416 ]
```

```r
# Opción 2: Pasando los estadísticos calculados
resultado2 <- ic_media_s(x_barra = 13.89,
                         s = 0.85,
                         n = 15,
                         confianza = 0.95)
```

```
=== INTERVALO DE CONFIANZA PARA LA MEDIA ===
Desviación estándar poblacional desconocida
Distribución utilizada: t de Student
Criterio: n < 30

Estadísticos calculados:
- Tamaño de muestra (n): 15
- Media muestral (x̄): 13.89
- Desviación estándar muestral (s): 0.85
- Grados de libertad: 14
- Nivel de confianza: 95 %

Cálculos del intervalo:
- Error estándar: 0.2195
- Valor t crítico: 2.1448
- Margen de error: 0.4707

RESULTADO:
IC al 95 %: [ 13.4193 , 14.3607 ]
```

16.4.3 Ejemplo 3: Intervalo de confianza para la varianza

Contexto agronómico: Se evalúa la variabilidad en el peso de 20 tomates con una desviación estándar muestral de 45 gramos.

Datos:

$$
\begin{aligned}
n &= 20 \\
s &= 45 \text{ g} \\
s^2 &= 2025 \text{ g}^2 \\
\alpha &= 0.10
\end{aligned}
$$

Grados de libertad:

$$gl = n - 1 = 20 - 1 = 19$$

Valores críticos de la distribución chi-cuadrado:

$$\chi^2_{0.05,19} = 30.144, \quad \chi^2_{0.95,19} = 10.117$$

Límite inferior para la varianza:

$$\frac{(n-1) \cdot s^2}{\chi^2_{\alpha/2,\, n-1}} = \frac{19 \cdot 2025}{30.144} = 1276.9$$

Límite superior para la varianza:

$$\frac{(n-1) \cdot s^2}{\chi^2_{1-\alpha/2,\, n-1}} = \frac{19 \cdot 2025}{10.117} = 3803.8$$

Intervalo de confianza del 90 % para la varianza σ^2:

$$IC_{90\%} \text{ para } \sigma^2 = [1276.9,\ 3803.8] \text{ g}^2$$

Intervalo de confianza del 90 % para la desviación estándar σ:

$$IC_{90\%} \text{ para } \sigma = [\sqrt{1276.9},\ \sqrt{3803.8}] = [35.73,\ 61.68] \text{ g}$$

Implementación en R:

```r
# Ejemplo con los datos del problema
resultado <- ic_varianza(s = 45,
                         n = 20,
                         confianza = 0.90)
```

```
=== INTERVALO DE CONFIANZA PARA LA VARIANZA ===
Distribución Chi-cuadrado

Estadísticos calculados:
- Tamaño de muestra (n): 20
- Desviación estándar muestral (s): 45
- Varianza muestral (s²): 2025
```

```
- Grados de libertad: 19
- Nivel de confianza: 90 %
- Nivel de significancia ( ): 0.1

Valores críticos de Chi-cuadrado:
-  ² 0.05 , 19 = 10.117
-  ² 0.95 , 19 = 30.1435

Cálculos del intervalo:
- Límite inferior varianza: ( 19 × 2025 ) / 30.144 = 1276.4
- Límite superior varianza: ( 19 × 2025 ) / 10.117 = 3803

RESULTADOS:
IC al 90 % para  ²: [ 1276.4 , 3803 ] IC al 90 % para  :  [ 35.73 , 61.67 ]
```

16.4.4 Ejemplo 4: Intervalo de confianza para la proporción

Contexto agronómico: De 200 plantas de trigo evaluadas, 156 mostraron resistencia a una enfermedad.

Cálculo manual:

Datos:

$$\begin{aligned} n &= 200 \\ x &= 156 \\ \hat{p} &= \frac{156}{200} = 0.78 \\ \alpha &= 0.05 \\ z_{\alpha/2} &= z_{0.025} = 1.96 \end{aligned}$$

Error estándar:

$$\text{Error estándar} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.78 \cdot 0.22}{200}} = \sqrt{0.000858} = 0.0293$$

Margen de error:

$$\text{Margen de error} = z_{\alpha/2} \cdot \text{Error estándar} = 1.96 \cdot 0.0293 = 0.0574$$

Intervalo de confianza al 95 %:

$$IC_{95\%} = \hat{p} \pm \text{Margen de error} = 0.78 \pm 0.0574 = [0.7226,\ 0.8374]$$

Implementación en R:

Usando la función base de R:

```r
# Cálculo directo con función incorporada en R base
prop.test(x = 156,
          n = 200,
          conf.level = 0.95,
          correct = FALSE)
```

```
    1-sample proportions test without continuity correction

data:  156 out of 200, null probability 0.5
X-squared = 62.72, df = 1, p-value = 2.383e-15
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.7176120 0.8318346
sample estimates:
   p
0.78
```

Nota: `correct = FALSE` evita la corrección de continuidad para que el resultado sea idéntico al cálculo manual.

Usando la función personalizada:

```r
# Uso con los datos del ejemplo
ic_proporcion(x = 156,
              n = 200,
              confianza = 0.95)
```

```
=== INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN ===
Datos:
- Éxitos (x): 156
- Tamaño de muestra (n): 200
- Proporción muestral (p̂): 0.78
- Nivel de confianza: 95 %

Cálculos:
- Error estándar: 0.0293
- Valor z crítico: 1.96
- Margen de error: 0.0574

RESULTADO:
IC al 95 %: [ 0.7226 , 0.8374 ]
```

Capítulo X

Pruebas de hipótesis

17 Pruebas de Hipótesis Paramétricas en R

En la investigación agronómica, es fundamental tomar decisiones basadas en datos. Las pruebas de hipótesis permiten evaluar si los resultados observados en una muestra pueden generalizarse a la población de interés o si son producto del azar. Este capítulo guía al estudiante en la aplicación de pruebas de hipótesis paramétricas utilizando R, desde la formulación de hipótesis hasta la interpretación de resultados, empleando ejemplos prácticos y reales del ámbito agronómico (López & González, 2018).

17.1 Fundamentos de las pruebas de hipótesis

Una prueba de hipótesis es un procedimiento estadístico que permite decidir, con un nivel de confianza predefinido, si una afirmación sobre un parámetro poblacional es compatible con los datos muestrales. El proceso general incluye:

1. Plantear la hipótesis nula (H_0) y la alternativa (H_a).
2. Seleccionar el estadístico de prueba adecuado según el tipo de dato y los supuestos.
3. Calcular el valor del estadístico y el valor-p.
4. Comparar el valor-p con el nivel de significancia (α), generalmente 0.05.
5. Tomar una decisión: rechazar o no rechazar H_0.

Criterios de selección de la prueba:

1. Tipo de variable (cuantitativa o cualitativa).
2. Tamaño de la muestra.
3. Conocimiento de la varianza poblacional.
4. Independencia o dependencia entre muestras.
5. Homogeneidad de varianzas.

17.2 Prueba de hipótesis sobre una media

Esta prueba se utiliza para determinar si la media de una población difiere de un valor específico. Es útil, por ejemplo, para verificar si el peso promedio de semillas, el rendimiento de un cultivo o el contenido de un nutriente cumple con un estándar.

17.2.1 Criterios de selección

1. Variable cuantitativa continua.
2. La muestra debe ser aleatoria.
3. Si la varianza poblacional es conocida y la muestra es grande ($n \geq 30$), se usa la prueba z.
4. Si la varianza es desconocida y la muestra es pequeña ($n < 30$), se usa la prueba t de Student.

17.2.2 Fórmulas

a) Prueba z (varianza conocida):

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

b) Prueba t (varianza desconocida):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

donde $gl = n - 1$.

17.2.3 Ejemplo hipotético

Supóngase que se afirma que el peso promedio de semillas de maíz es de 250 mg. Se toma una muestra de 20 semillas, obteniéndose una media de 242 mg y una desviación estándar de 15 mg. Se desea saber, con un nivel de significancia del 5%, si el peso medio difiere del valor afirmado.

1. Planteamiento de hipótesis:

1. $H_0 : \mu = 250$ mg
2. $H_a : \mu \neq 250$ mg

2. Cálculo del estadístico:

$$t = \frac{242 - 250}{15/\sqrt{20}} = \frac{-8}{3.354} = -2.39$$

3. Región crítica:

Para $gl = 19$ y $\alpha = 0.05$ (bilateral), el valor crítico es ± 2.093.

4. Decisión:

Como $|-2.39| > 2.093$, se rechaza H_0.

5. Conclusión:

Con un 5% de significancia, existe evidencia de que el peso medio difiere de 250 mg.

17.2.4 Código en R explicado

```r
# Instalar paquete si no está instalado
## Para realizar pruebas de hipotesis
if (!require(BSDA)) install.packages("BSDA")
if (!require(EnvStats)) install.packages("EnvStats")


# Prueba t con estadísticos resumidos usando tsum.test()
tsum.test(mean.x = 242,
          s.x = 15,
          n.x = 20,
          mu = 250,
          alternative = "two.sided",
          conf.level = 0.95)
```

```
    One-sample t-Test

data:  Summarized x
t = -2.3851, df = 19, p-value = 0.02765
alternative hypothesis: true mean is not equal to 250
95 percent confidence interval:
 234.9798 249.0202
sample estimates:
mean of x
      242
```

Parámetros de `tsum.test()`:

1. **mean.x**: media muestral (242 mg)
2. **s.x**: desviación estándar muestral (15 mg)
3. **n.x**: tamaño de muestra (20)
4. **mu**: valor hipotético bajo H (250 mg)
5. **alternative**: tipo de prueba ("two.sided" para bilateral)
6. **conf.level**: nivel de confianza (0.95 para 95%)

17.3 Prueba de hipótesis sobre dos medias

Permite comparar si las medias de dos poblaciones son iguales o diferentes. Es útil, por ejemplo, para comparar el rendimiento de dos variedades de cultivo, el efecto de dos tratamientos o la altura de plantas de dos especies.

17.3.1 Criterios de selección

1. Las muestras pueden ser independientes (grupos distintos) o dependientes (mediciones pareadas).
2. Se debe verificar si las varianzas son iguales o diferentes.
3. Si las muestras son grandes ($n_1, n_2 \geq 30$), se puede usar la prueba z; si son pequeñas y la varianza es desconocida, se usa la prueba t.

17.3.2 Fórmulas

a) Muestras independientes, varianzas iguales (t "pooled"):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

b) Muestras independientes, varianzas diferentes (Welch):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

c) Muestras dependientes (pareadas):

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

donde $\bar{d}$ es la media de las diferencias y s_d su desviación estándar.

17.3.3 Ejemplo hipotético (independientes, varianzas iguales)

Se comparan las alturas de dos especies forestales.

1. Especie 1: $\bar{x}_1 = 25.97$ m, $s_1 = 1.36$, $n_1 = 13$
2. Especie 2: $\bar{x}_2 = 25.39$ m, $s_2 = 1.77$, $n_2 = 14$

1. Hipótesis:

1. $H_0 : \mu_1 = \mu_2$
2. $H_a : \mu_1 \neq \mu_2$

2. Cálculo:

$$s_p^2 = \frac{12 \times 1.36^2 + 13 \times 1.77^2}{25} = 2.30$$

$$s_p = \sqrt{2.30} = 1.52$$

$$t = \frac{25.97 - 25.39}{1.52\sqrt{\frac{1}{13} + \frac{1}{14}}} = 0.94$$

3. Decisión: Para $gl = 25$, $t_{0.025} = 2.060$. Como $0.94 < 2.060$, no se rechaza H_0.

17.3.4 Código en R explicado

```
# Datos del ejercicio
mean1 <- 25.97; s1 <- 1.36; n1 <- 13  # Especie 1
mean2 <- 25.39; s2 <- 1.77; n2 <- 14  # Especie 2

# Prueba t para dos muestras independientes con varianzas iguales
tsum.test(mean.x = mean1, s.x = s1, n.x = n1,
          mean.y = mean2, s.y = s2, n.y = n2,
          alternative = "two.sided",
          mu = 0,        # diferencia hipotética (H :   -   = 0)
          var.equal = TRUE,          # asume varianzas iguales (pooled)
          conf.level = 0.95)         # nivel de confianza
```

```
    Standard Two-Sample t-Test

data:  Summarized x and y
t = 0.94918, df = 25, p-value = 0.3516
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -0.678492  1.838492
sample estimates:
mean of x mean of y
    25.97     25.39
```

Parámetros de `tsum.test()` para dos muestras:

1. **mean.x, s.x, n.x**: estadísticos de la muestra 1
2. **mean.y, s.y, n.y**: estadísticos de la muestra 2
3. **mu**: diferencia hipotética bajo H (0 para igualdad de medias)
4. **var.equal = TRUE**: usa varianza pooled (varianzas iguales)
5. **alternative**: "two.sided" para prueba bilateral

17.3.5 Ejemplo hipotético (pareadas)

Se evalúa el efecto de una capacitación en 10 empleados, midiendo el puntaje antes y después.

1. Hipótesis:

1. $H_0 : \mu_D = 0$ (no hay diferencia)
2. $H_a : \mu_D \neq 0$ (hay diferencia)

2. Cálculo:
Supóngase que la media de las diferencias es -0.4 y la desviación estándar 0.8 .

$$t = \frac{-0.4}{0.8/\sqrt{10}} = -1.58$$

3. Decisión:
Para $gl = 9$, $t_{0.05} = 2.262$. Como $|-1.58| < 2.262$, no se rechaza H_0.

17.3.6 Código en R explicado

```r
# Datos del ejercicio (estadísticos de las diferencias)
n <- 10
mean_diff <- -0.4    # media de diferencias (antes - después)
sd_diff <- 0.8       # desviación estándar de diferencias

# Prueba t pareada usando estadísticos resumidos
# Para muestras pareadas, usamos tsum.test() con una sola muestra
# (las diferencias)
tsum.test(mean.x = mean_diff,
          s.x = sd_diff,
```

```
          n.x = n,
          mu = 0,                      # H :  _D = 0
          alternative = "two.sided",
          conf.level = 0.95)
```

```
Warning in tsum.test(mean.x = mean_diff, s.x = sd_diff, n.x = n, mu = 0, :
argument 'var.equal' ignored for one-sample test.
```

```
    One-sample t-Test

data:  Summarized x
t = -1.5811, df = 9, p-value = 0.1483
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.9722855  0.1722855
sample estimates:
mean of x
     -0.4
```

Una prueba t pareada es equivalente a una prueba t de una muestra sobre las diferencias.

Por eso usamos `tsum.test()` con:

1. **mean.x**: media de las diferencias (-0.4)
2. **s.x**: desviación estándar de las diferencias (0.8)
3. **n.x**: número de pares (10)
4. **mu = 0**: hipótesis nula (no hay diferencia promedio)

Alternativa con datos individuales:

Si tuvieras los datos originales:

```
# Datos
antes <- c(9.0,7.3,6.7,5.3,8.7,6.3,7.9,7.3,8.0,8.5)
despues <- c(9.2,8.2,8.5,4.9,8.9,5.8,8.2,7.8,9.5,8.0)
# Test para datos pareados
t.test(antes, despues,
       paired = TRUE,
       alternative = "two.sided",
       conf.level = 0.95)
```

```
    Paired t-test

data:  antes and despues
t = -1.5784, df = 9, p-value = 0.1489
```

```
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.9732782  0.1732782
sample estimates:
mean difference
          -0.4
```

Por eso usamos `t.test()` con:

1. **antes**: Vector numerico con los datos iniciales.
2. **despues**: Vector numerico con los datos finales o pareados.
3. **paired**: si es una prueba de t pareada (TRUE)
4. **alternative**: "two.sided" hace referencia a que compara una igualdad.

17.4 Prueba de hipótesis sobre una proporción

Permite determinar si la proporción de una característica en la población es igual a un valor específico. Por ejemplo, si la proporción de agricultores que adopta una tecnología supera el 60%.

17.4.1 Criterios de selección

1. Variable cualitativa dicotómica.
2. Tamaño muestral suficiente para aproximación normal ($np_0 > 5np$ y $n(1-p_0) > 5$).

Fórmula

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

17.4.2 Ejemplo hipotético

De 180 agricultores, 120 adoptaron un fertilizante. Se desea saber si la proporción es diferente de 0.60.

$$\hat{p} = \frac{120}{180} = 0.667$$

$$z = \frac{0.667 - 0.60}{\sqrt{0.60 \times 0.40/180}} = 1.56$$

Para $\alpha = 0.05$, $z_{0.025} = 1.96$. Como $1.56 < 1.96$, no se rechaza H_0.

17.4.3 Código en R explicado

```r
prop.test(x = 120, n = 180,
          p = 0.60,                # valor bajo H0
          alternative = "two.sided",
          correct = FALSE)         # sin corrección de continuidad
```

```
    1-sample proportions test without continuity correction

data:  120 out of 180, null probability 0.6
X-squared = 3.3333, df = 1, p-value = 0.06789
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.5949523 0.7314158
sample estimates:
        p
0.6666667
```

1. **x**: número de éxitos.
2. **n**: tamaño de la muestra.
3. **p**: proporción bajo H_0.

17.5 Prueba de hipótesis sobre dos proporciones

Permite comparar si la proporción de una característica es igual en dos poblaciones. Por ejemplo, comparar la proporción de adopción de una tecnología entre hombres y mujeres.

17.5.1 Criterios de selección

1. Variable cualitativa dicotómica.
2. Muestras independientes.
3. Tamaño muestral suficiente.

17.5.2 Fórmulas

$$\hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

17.5.3 Ejemplo hipotético

En una encuesta, 110 de 200 hombres y 210 de 300 mujeres respondieron. ¿Existe diferencia en las proporciones?

$$\hat{p}_1 = 0.55, ; \hat{p}_2 = 0.70$$

$$\hat{p}_c = \frac{110 + 210}{200 + 300} = 0.64$$

$$z = \frac{0.55 - 0.70}{\sqrt{0.64 \times 0.36 \left(\frac{1}{200} + \frac{1}{300}\right)}} = -3.42$$

Como $|-3.42| > 1.96$, se rechaza H_0.

17.5.4 Código en R explicado

```r
xp <- c(110, 210)   # éxitos en cada grupo
np <- c(200, 300)   # tamaño de cada grupo

prop.test(xp, np,
          alternative = "two.sided",
          correct = FALSE)
```

```
    2-sample test for equality of proportions without continuity correction

data:  xp out of np
X-squared = 11.719, df = 1, p-value = 0.0006187
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.23627182 -0.06372818
sample estimates:
prop 1 prop 2
  0.55   0.70
```

1. **xp**: vector de éxitos.
2. **np**: vector de tamaños.

17.6 Prueba de hipótesis sobre varianzas

La prueba de hipótesis sobre varianzas permite evaluar si la variabilidad observada en una muestra es compatible con un valor de referencia o si existen diferencias en la variabilidad entre dos poblaciones. Este tipo de prueba es fundamental en agronomía para analizar la uniformidad de procesos, como la comparación de la variabilidad en el rendimiento de cultivos bajo diferentes métodos de riego o el control de calidad de productos agrícolas.

17.6.1 Criterios de selección

1. La variable de interés debe ser cuantitativa y continua.
2. Los datos deben provenir de poblaciones con distribución normal.
3. Para comparar dos varianzas, las muestras deben ser independientes.

17.6.2 Fórmulas

a) Una varianza (χ^2): Esta prueba se utiliza para determinar si la varianza de una población es igual a un valor específico, generalmente un estándar de calidad o una especificación técnica.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

donde:

1. n es el tamaño de la muestra,
2. s^2 es la varianza muestral,
3. σ_0^2 es la varianza poblacional bajo la hipótesis nula.

Para este calculo no existe una función predefinida en R que lo realice con fines prácticos se desarrollo la siguiente función personalizada para esta tarea:

```
# Función personalizada para prueba de hipótesis de una varianza
var_test_chi <- function(x = NULL,
                         n = NULL,
                         s2 = NULL,
                         sigma0_2,
                         alternative = "two.sided",
                         alpha = 0.05) {

  # Validación de argumentos
  if (is.null(x) && (is.null(n) || is.null(s2))) {
    stop("Debe proporcionar 'x' (vector de datos) o
         'n' y 's2' (estadísticos muestrales)")
  }
```

```r
if (!is.null(x) && (!is.null(n) || !is.null(s2))) {
  warning("Se proporcionaron datos y estadísticos.
          Se usarán los datos 'x'")
}

# Calcular estadísticos si se proporcionan los datos
if (!is.null(x)) {
  n <- length(x)
  s2 <- var(x)
}

# Validar alternative
alternative <- match.arg(alternative, c("two.sided", "less", "greater"))

# Calcular estadístico chi-cuadrado
chi_sq <- (n - 1) * s2 / sigma0_2
df <- n - 1

# Calcular valor-p según el tipo de prueba
if (alternative == "two.sided") {
  # Para prueba bilateral
  p_value <- 2 * min(pchisq(chi_sq, df),
                     pchisq(chi_sq, df, lower.tail = FALSE))
} else if (alternative == "greater") {
  p_value <- pchisq(chi_sq, df, lower.tail = FALSE)
} else { # alternative == "less"
  p_value <- pchisq(chi_sq, df, lower.tail = TRUE)
}

# Decisión
decision <- ifelse(p_value < alpha, "Rechazar H0", "No rechazar H0")

# Valor crítico
if (alternative == "two.sided") {
  crit_lower <- qchisq(alpha/2, df)
  crit_upper <- qchisq(1 - alpha/2, df)
  critical_value <- c(crit_lower, crit_upper)
} else if (alternative == "greater") {
  critical_value <- qchisq(1 - alpha, df)
} else { # alternative == "less"
  critical_value <- qchisq(alpha, df)
}

# Crear objeto de resultado
result <- list(
  statistic = chi_sq,
  parameter = df,
  p.value = p_value,
```

```
    critical.value = critical_value,
    alternative = alternative,
    method = "Prueba de hipótesis para una varianza (Chi-cuadrado)",
    data.name = deparse(substitute(x)),
    sample.size = n,
    sample.variance = s2,
    null.variance = sigma0_2,
    alpha = alpha,
    decision = decision
  )

  class(result) <- "var_test_custom"
  return(result)
}

# Método print personalizado para mostrar resultados de forma clara
print.var_test_custom <- function(x, ...) {
  cat("\n")
  cat(x$method, "\n")
  cat("Datos:", x$data.name, "\n")
  cat("\n")
  cat("Hipótesis:\n")
  if (x$alternative == "two.sided") {
    cat("H0: sigma^2 =", x$null.variance, "\n")
    cat("Ha: sigma^2  ", x$null.variance, "\n")
  } else if (x$alternative == "greater") {
    cat("H0: sigma^2  ", x$null.variance, "\n")
    cat("Ha: sigma^2 >", x$null.variance, "\n")
  } else {
    cat("H0: sigma^2  ", x$null.variance, "\n")
    cat("Ha: sigma^2 <", x$null.variance, "\n")
  }
  cat("\n")
  cat("Estadísticos de la muestra:\n")
  cat("n =", x$sample.size, "\n")
  cat("s^2 =", round(x$sample.variance, 4), "\n")
  cat("\n")
  cat("Estadístico de prueba:\n")
  cat("Chi-cuadrado =", round(x$statistic, 4), "\n")
  cat("Grados de libertad =", x$parameter, "\n")
  cat("\n")
  cat("Valor crítico(s):\n")
  if (length(x$critical.value) == 2) {
    cat("Chi^2(", x$alpha/2, ",", x$parameter, ") =",
        round(x$critical.value[1], 4), "\n")
    cat("Chi^2(", 1-x$alpha/2, ",", x$parameter, ") =",
        round(x$critical.value[2], 4), "\n")
  } else {
    cat("Chi^2 crítico =", round(x$critical.value, 4), "\n")
```

```
  }
  cat("\n")
  cat("Valor-p =", round(x$p.value, 6), "\n")
  cat("Nivel de significancia =", x$alpha, "\n")
  cat("\n")
  cat("Decisión:", x$decision, "\n")
  cat("\n")
}
```

Esta función cuenta con la siguiente sintaxis para su uso:

$$\text{var\_test\_chi}(x, sigma0_2, alternative, alpha)$$

Argumentos en orden:

1. **x**: Vector con los datos de la muestra
2. **sigma0_2**: Varianza poblacional con la que se está comparando
3. **alternative**: Opción en carácter que indica el tipo de prueba utilizando los mismos argumentos que las otras funciones "greater", "less" y "two.sided".
4. **alpha**: nivel de significancia

También tiene la siguiente sintaxis cuando no se cuenta con los datos de la muestra directamente:

$$\text{var\_test\_chi}(n, s2, sigma0_2, alternative, alpha)$$

Argumentos en orden:

1. **n**: Tamaño de la muestra
2. **s2**: Varianza muestral
3. **sigma0_2**: Varianza poblacional con la que se está comparando
4. **alternative**: Opción en carácter que indica el tipo de prueba utilizando los mismos argumentos que las otras funciones "greater", "less" y "two.sided".
5. **alpha**: nivel de significancia

b) Dos varianzas (F de Fisher):

$$F = \frac{s_1^2}{s_2^2}$$

donde:

1. s_1^2 y s_2^2 son las varianzas muestrales de los dos grupos,
2. n_1 y n_2 son los tamaños de muestra de cada grupo.

17.6.3 Ejemplo hipotético (una varianza)

Supóngase que una empresa agrícola establece que la varianza máxima aceptable en el peso de frutos de tomate es de $\sigma_0^2 = 4$ g². Se toma una muestra de 10 frutos y se obtiene una varianza muestral de $s^2 = 5.8$ g². Se desea saber, con un nivel de significancia del 5%, si la variabilidad excede el estándar.

1. Planteamiento de hipótesis:

1. $H_0 : \sigma^2 = 4$ g² (la varianza cumple el estándar)
2. $H_a : \sigma^2 > 4$ g² (la varianza excede el estándar)

2. Cálculo del estadístico:

$$\chi^2 = \frac{(10-1) \times 5.8}{4} = \frac{52.2}{4} = 13.05$$

Decisión:

El valor crítico para $\alpha = 0.05$ y $gl = 9$ es $\chi^2\_0.05, 9 = 16.92$. Como $13.05 < 16.92$, no se rechaza H_0.

Código en R explicado:

```
# Aplicacion de la funcion personalizada
var_test_chi(n = 10, s2 = 5.8,
                                sigma0_2 = 4,
                                alternative = "greater",
                                alpha = 0.05)
```

```
Prueba de hipótesis para una varianza (Chi-cuadrado)
Datos: NULL

Hipótesis:
H0: sigma^2  4
Ha: sigma^2 > 4

Estadísticos de la muestra:
n = 10
s^2 = 5.8

Estadístico de prueba:
Chi-cuadrado = 13.05
Grados de libertad = 9

Valor crítico(s):
Chi^2 crítico = 16.919

Valor-p = 0.160357
```

```
Nivel de significancia = 0.05

Decisión: No rechazar H0
```

Para la solución de este problema se emplea la funcion personalizada `var_test_chique` permite realizar una prueba de hipotesis de la varianza para una muestra empleando los argumentos mencionados en la sección donde se presentó la función.

17.6.4 Ejemplo hipotético (dos varianzas)

Se desea comparar la varianza del rendimiento de dos tratamientos de riego.

Datos:

1. Tratamiento 1: 10, 12, 11, 13, 12, 11, 14, 13
2. Tratamiento 2: 9, 10, 11, 10, 12, 10, 11, 10

Planteamiento de hipótesis:

1. $H_0 : \sigma_1^2 = \sigma_2^2$ (las varianzas son iguales)
2. $H_a : \sigma_1^2 \neq \sigma_2^2$ (las varianzas son diferentes)

Cálculo y decisión en R:

```r
trat1 <- c(10, 12, 11, 13, 12, 11, 14, 13)
trat2 <- c(9, 10, 11, 10, 12, 10, 11, 10)

var.test(trat1, trat2,
         alternative = "two.sided", # prueba bilateral
         conf.level = 0.95)         # nivel de confianza
```

```
    F test to compare two variances

data:  trat1 and trat2
F = 2.0426, num df = 7, denom df = 7, p-value = 0.3667
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.408927 10.202368
sample estimates:
ratio of variances
          2.042553
```

1. **var.test**: realiza la prueba F para comparar varianzas.
2. **alternative**: define si la prueba es bilateral o unilateral.
3. **conf.level**: establece el nivel de confianza.

Capítulo XI

Regresión lineal y correlación

18 Análisis de correlación lineal simple

El análisis de correlación lineal simple permite cuantificar el grado de asociación lineal entre dos variables cuantitativas. En agronomía, este análisis es fundamental para evaluar relaciones como el diámetro y el peso de frutos, o el contenido de materia orgánica y calcio en suelos (López & González, 2018). La correlación no implica causalidad, pero sí proporciona una medida objetiva de la fuerza y dirección de la relación lineal entre dos variables (Moore et al., 2017).

18.1 Covarianza

La covarianza mide la tendencia conjunta de dos variables a aumentar o disminuir simultáneamente. Su valor puede ser positivo, negativo o cero, pero su magnitud depende de las unidades de las variables, lo que dificulta la comparación entre estudios.

La covarianza poblacional se define como:

$$\text{Cov}(X, Y) = \text{E}[(X - \mu_X)(Y - \mu_Y)]$$

El estimador muestral de la covarianza es:

$$\hat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Ejemplo paso a paso:

Considérese los siguientes datos de peso de padres (X) y peso de hijos (Y) en kilogramos:

x_i	78	65	86	68	83	68	75	80	82	66
y_i	60	52	68	53	65	57	58	62	65	53

1. Calcular las medias:

$$\bar{x} = \frac{78 + 65 + ... + 66}{10} = 75.1$$

$$\bar{y} = \frac{60 + 52 + ... + 53}{10} = 59.3$$

2. Calcular las diferencias respecto a la media y sus productos:

$$(x_i - \bar{x})(y_i - \bar{y})(xi\ - x)(yi\ - y\ )$$

Por ejemplo, para el primer par:

$$(78 - 75.1)(60 - 59.3) = 2.9 \times 0.7 = 2.03$$

Se repite para cada par y se suman los resultados:

$$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 386.7$$

3. Calcular la covarianza:

$$\hat{\text{Cov}}(X, Y) = \frac{386.7}{10 - 1} = 42.97$$

18.2 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (r) es una medida adimensional que cuantifica la fuerza y dirección de la relación lineal entre dos variables. Su valor oscila entre -1 y 1.

Definición poblacional:

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Estimador muestral:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Cálculo paso a paso:

1. Calcular las sumas de cuadrados:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

2. Calcular el numerador y denominador:

$$\text{Numerador} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 386.7$$

$$\text{Denominador} = \sqrt{S_{xx} \cdot S_{yy}}$$

Supóngase que $S_{xx} = 546.9$ y $S_{yy} = 288.1$:

$$\text{Denominador} = \sqrt{546.9 \times 288.1} = \sqrt{157,561.89} = 396.9$$

3. Calcular r:

$$r = \frac{386.7}{396.9} = 0.974$$

Interpretación: Un valor de $r = 0.974$ indica una asociación lineal positiva muy fuerte entre las variables.

18.3 Prueba de significancia para el coeficiente de correlación

Para determinar si la correlación observada es estadísticamente significativa, se utiliza la siguiente hipótesis:

1. $H_0 : \rho = 0$ (no hay correlación lineal)
2. $H_1 : \rho \neq= 0$ (existe correlación lineal)

El estadístico de prueba es:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Este estadístico sigue una distribución t de Student con $n - 2$ grados de libertad.

Ejemplo:

Con $r = 0.974$ y $n = 10$:

$$
\begin{aligned}
t &= \frac{0.974\sqrt{10-2}}{\sqrt{1-0.974^2}} \\
&= \frac{0.974 \times 2.828}{\sqrt{1-0.949}} \\
&= \frac{2.754}{\sqrt{0.051}} \\
&= \frac{2.754}{0.226} \\
&= 12.19
\end{aligned}
$$

Se compara el valor calculado con el valor crítico de t para $n-2=8$ grados de libertad y el nivel de significancia deseado (por ejemplo, $\alpha = 0.05$. Si $|t| > t_{crtico}$, se rechaza H_0.

18.4 Uso de funciones en R

18.4.1 Función cov()

La función `cov()` permite calcular la covarianza muestral entre dos vectores numéricos. Su sintaxis general es:

```
cov(x,
    y,
    use = "everything",
    method = "pearson")
```

Argumentos principales:

Los argumentos principales son los siguientes:

1. **x, y:** vectores numéricos que contienen los datos de las dos variables a comparar.
2. **use:** especifica el método para el tratamiento de valores faltantes. Por ejemplo, `"everything"` utiliza todos los datos, mientras que `"complete.obs"` excluye las observaciones con valores faltantes.
3. **method:** indica el tipo de covarianza a calcular. El valor por defecto es `"pearson"`, que corresponde a la covarianza clásica.

18.4.2 Función cor()

La función `cor()` se utiliza para calcular el coeficiente de correlación entre dos vectores numéricos. La sintaxis básica es:

```
cor(x,
    y,
    method = "pearson")
```

Argumentos principales:

1. **x, y:** vectores numéricos que representan las variables de interés.
2. **method:** define el tipo de correlación a calcular. Puede tomar los valores `"pearson"` (por defecto, para correlación lineal), `"spearman"` (para correlación de rangos) o `"kendall"` (para correlación de concordancia).

18.4.3 Función cor.test()

La función `cor.test()` realiza una prueba de hipótesis para el coeficiente de correlación entre dos variables. Su sintaxis general es:

```
cor.test(x,
         y,
         alternative = "two.sided",
         method = "pearson",
         conf.level = 0.95)
```

Argumentos principales:

1. **x, y:** vectores numéricos que contienen los datos de las variables a analizar.
2. **alternative:** especifica la hipótesis alternativa. Puede ser `"two.sided"` (prueba bilateral), `"less"` (prueba unilateral para correlación negativa) o `"greater"` (prueba unilateral para correlación positiva).
3. **method:** determina el tipo de correlación a evaluar. Puede ser `"pearson"`, `"spearman"` o `"kendall"`.
4. **conf.level:** establece el nivel de confianza para el intervalo del coeficiente de correlación, siendo el valor por defecto 0.95 (95%).

18.4.4 Resolución del ejemplo en R

```
# Importación de los valores
# Datos del ejemplo: peso de padres (X) y peso de hijos (Y) en kilogramos
datos <- data.frame(
  Peso_Padres = c(78, 65, 86, 68, 83, 68, 75, 80, 82, 66),
  Peso_Hijos = c(60, 52, 68, 53, 65, 57, 58, 62, 65, 53)
)

# Calculo de la suma de cuadrados (Sxx)
sum((datos$Peso_Padres-mean(datos$Peso_Padres))^2)
```

```
[1] 546.9
```

```
# Calculo de la suma de cuadrados (Syy)
sum((datos$Peso_Hijos-mean(datos$Peso_Hijos))^2)
```

```
[1] 288.1
```

```r
# Calculo de la covarianza
cov(datos$Peso_Padres,datos$Peso_Hijos)
```

```
[1] 42.96667
```

```r
# Calculo del coeficiente de correalción
cor(datos$Peso_Padres,datos$Peso_Hijos)
```

```
[1] 0.974201
```

```r
# Test de correlación
cor.test(datos$Peso_Padres, datos$Peso_Hijos,
         alternative = "two.sided",
         method = "pearson", conf.level = 0.95)
```

```
    Pearson's product-moment correlation

data:  datos$Peso_Padres and datos$Peso_Hijos
t = 12.209, df = 8, p-value = 1.879e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8912550 0.9940775
sample estimates:
     cor
0.974201
```

Interpretación: El coeficiente de correlación de Pearson calculado es $r = 0.974$, lo que indica una asociación lineal positiva muy fuerte entre las variables analizadas. El valor del estadístico de prueba es $t = 12.209$ con 8 grados de libertad, y el valor p asociado es 1.879×10^{-6}. Este valor p es considerablemente menor que el nivel de significancia convencional ($\alpha = 0.05$), lo que proporciona evidencia estadísticamente significativa para rechazar la hipótesis nula de ausencia de correlación lineal ($H_0 : \rho = 0$).

El intervalo de confianza al 95% para el coeficiente de correlación se encuentra entre 0.891 y 0.994, lo que refuerza la conclusión de que la verdadera correlación poblacional es positiva y muy alta.

18.5 Visualización Gráfica en el Análisis de Correlación Lineal Simple

La representación gráfica constituye una herramienta fundamental en el análisis de correlación lineal simple, ya que permite visualizar la naturaleza y fuerza de la relación

entre dos variables cuantitativas (López & González, 2018). Los gráficos facilitan la interpretación de los resultados estadísticos y proporcionan una comprensión intuitiva de los datos antes de proceder con los cálculos formales del coeficiente de correlación de Pearson.

18.5.1 Preparación de los Datos

Antes de generar los gráficos, es necesario extraer los datos del conjunto de datos y organizarlos en vectores individuales para facilitar su manipulación:

```
# Extraer los datos del dataframe a vectores
x <- datos$Peso_Padres
y <- datos$Peso_Hijos
```

Esta separación permite un manejo más eficiente de las variables y facilita la aplicación de las funciones gráficas de R.

18.5.2 Gráfico de Dispersión con Línea de Regresión

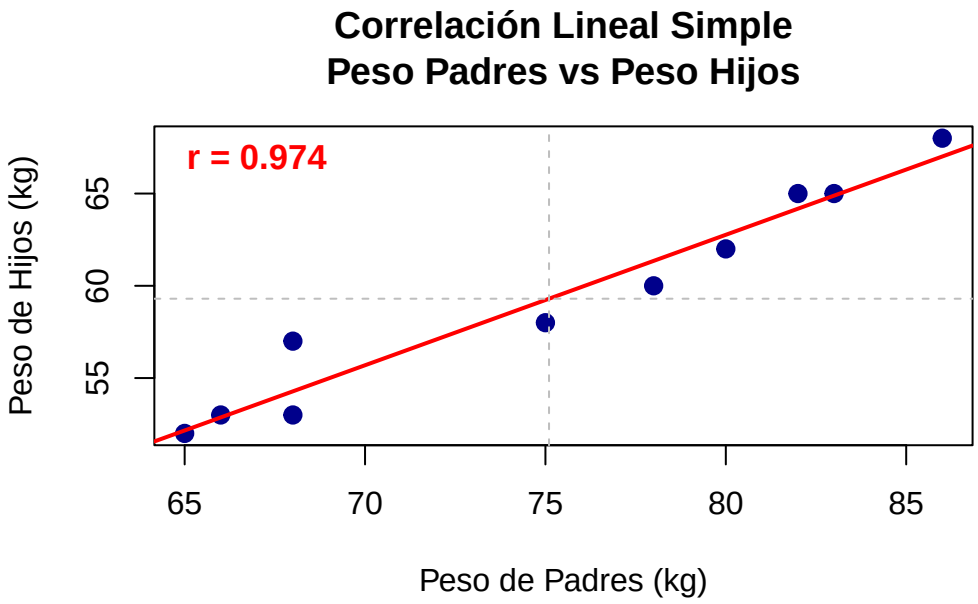
El diagrama de dispersión representa la herramienta visual más importante para evaluar la correlación lineal, ya que permite observar directamente el patrón de asociación entre las variables (López & González, 2018).

```
# 1. Gráfico de dispersión básico con línea de regresión
plot(x, y,
     main = "Correlación Lineal Simple\nPeso Padres vs Peso Hijos",
     xlab = "Peso de Padres (kg)",
     ylab = "Peso de Hijos (kg)",
     pch = 19,
     col = "darkblue",
     cex = 1.2)

# Agregar línea de regresión
abline(lm(y ~ x), col = "red", lwd = 2)

# Agregar líneas de las medias
abline(v = mean(x), col = "gray", lty = 2, lwd = 1)
abline(h = mean(y), col = "gray", lty = 2, lwd = 1)

# Agregar texto con estadísticas
text(67, 67, paste("r =", round(cor(x,y), 3)),
     col = "red", font = 2, cex = 1.1)
```



Elementos Explicativos:

1. **Puntos de dispersión**: Cada punto representa un par de observaciones (xi, yi)
2. **Línea de regresión**: Muestra la tendencia lineal de los datos
3. **Líneas de medias**: Indican los valores promedio de cada variable
4. **Coeficiente de correlación**: Cuantifica la fuerza de la asociación lineal

18.5.3 Gráfico con Intervalos de Confianza

Este gráfico avanzado incorpora bandas de confianza que indican la incertidumbre asociada con la línea de regresión estimada.

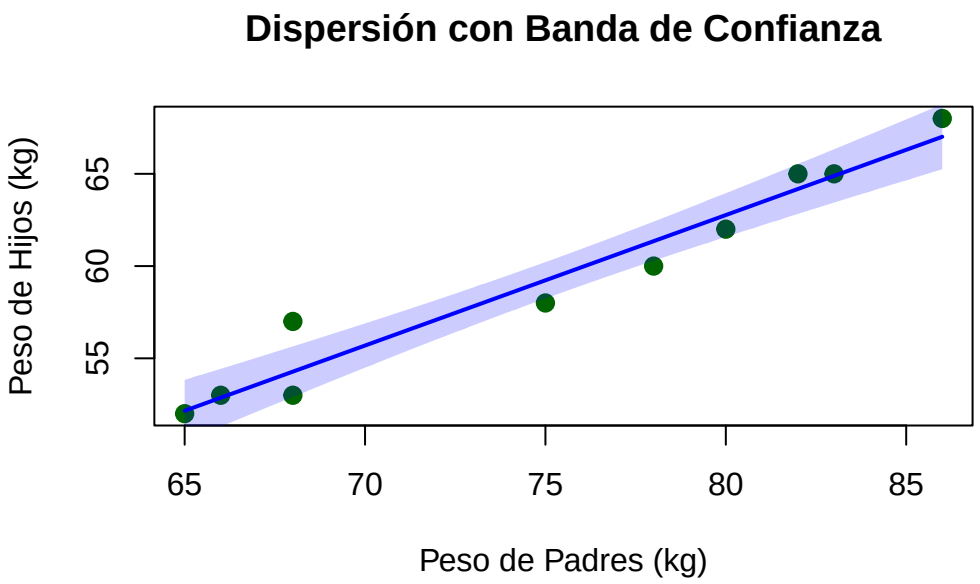
```
# 2. Gráfico con intervalos de confianza
plot(x, y,
     main = "Dispersión con Banda de Confianza",
     xlab = "Peso de Padres (kg)",
     ylab = "Peso de Hijos (kg)",
     pch = 19,
     col = "darkgreen",
     cex = 1.2)

# Crear secuencia para línea suave
x_seq <- seq(min(x), max(x), length.out = 100)
modelo <- lm(y ~ x)
predicciones <- predict(modelo, newdata = data.frame(x = x_seq),
                        interval = "confidence")

# Agregar banda de confianza
```

```r
polygon(c(x_seq, rev(x_seq)),
        c(predicciones[,"lwr"], rev(predicciones[,"upr"])),
        col = rgb(0, 0, 1, 0.2), border = NA)

# Línea de regresión
lines(x_seq, predicciones[,"fit"], col = "blue", lwd = 2)
```



Interpretación: La banda sombreada representa el intervalo de confianza del 95% para la línea de regresión, indicando el rango de valores donde se espera que se encuentre la verdadera relación poblacional.

18.6 Simulación interactiva del coeficiente de Pearson

A continuación, se presenta una aplicación interactiva diseñada para visualizar cómo los diferentes valores del coeficiente de correlación de Pearson (r) se reflejan en un gráfico de dispersión junto con su respectiva línea de tendencia. Esta herramienta facilita la comprensión visual y práctica del concepto de correlación lineal simple, permitiendo observar de manera dinámica cómo varía la relación entre dos variables a medida que cambia el valor de r.

Para explorar la aplicación y experimentar con distintos escenarios de correlación, se recomienda acceder al siguiente enlace: https://ludwing-mj.shinyapps.io/pearson/.

19 Regresión Lineal Simple usando R

La regresión lineal simple es una técnica estadística fundamental para analizar la relación entre dos variables cuantitativas, permitiendo modelar y predecir el comportamiento de una variable dependiente a partir de una variable independiente. En el contexto de la agronomía, esta herramienta resulta esencial para comprender fenómenos como la relación entre el peso en cultivos o animales, entre otros ejemplos relevantes (Montgomery et al., 2021; López & González, 2018).

19.1 Fundamentos Teóricos

El modelo de regresión lineal simple se expresa mediante la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

En esta expresión:

1. Y_i representa el valor observado de la variable dependiente para el individuo i.
2. X_i es el valor observado de la variable independiente para el individuo i.
3. β_0 es el intercepto o constante, que indica el valor esperado de Y cuando $X = 0$.
4. β_1 es la pendiente, que representa el cambio promedio en Y por cada unidad de cambio en X.
5. ε_i es el término de error aleatorio, que recoge la variabilidad no explicada por el modelo.

El objetivo de la regresión es estimar los valores de β_0 y β_1 que mejor se ajustan a los datos observados. Para ello, se utiliza el método de mínimos cuadrados, que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (Montgomery et al., 2021).

Las fórmulas para los estimadores de mínimos cuadrados son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde $\bar{x}$ y $\bar{y}$ son las medias de las variables X y Y respectivamente.

19.2 Supuestos del Modelo

Para que los resultados de la regresión lineal simple sean válidos, es necesario que se cumplan los siguientes supuestos (López & González, 2018):

1. **Linealidad:** La relación entre la variable independiente y la dependiente debe ser lineal. Esto significa que el efecto de X sobre Y es constante a lo largo de todo el rango de valores.
2. **Normalidad de los errores:** Los residuos (diferencias entre los valores observados y los predichos) deben seguir una distribución normal.
3. **Homocedasticidad:** La varianza de los errores debe ser constante para todos los valores de X.
4. **Independencia:** Las observaciones deben ser independientes entre sí, es decir, el valor de una observación no debe influir en el valor de otra.

El incumplimiento de estos supuestos puede llevar a conclusiones erróneas o a una interpretación incorrecta de los resultados.

19.3 Análisis Práctico en R

19.3.1 Instalación y carga de paquetes

El análisis inicia con la carga de los paquetes especializados y la exploración de los datos:

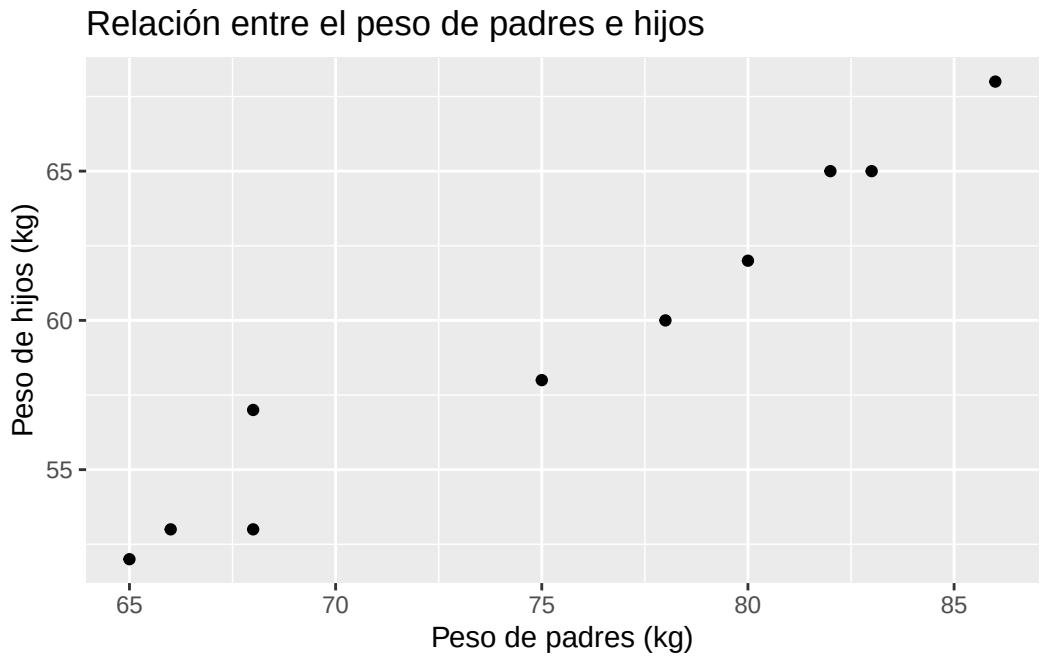
```
# Instalación y carga de paquetes necesarios
if (!require(tidyverse)) install.packages("tidyverse")
if (!require(car)) install.packages("car")
if (!require(lmtest)) install.packages("lmtest")
if (!require(nortest)) install.packages("nortest")
```

Se recomienda siempre inspeccionar los datos antes de analizarlos. En este ejemplo, se utiliza un conjunto de datos ficticio sobre el peso de padres e hijos empleado para explicar el análisis de correlación lineal:

```
# Datos del ejemplo: peso de padres (X) y peso de hijos (Y) en kilogramos
datos <- data.frame(
  Peso_Padres = c(78, 65, 86, 68, 83, 68, 75, 80, 82, 66),
  Peso_Hijos = c(60, 52, 68, 53, 65, 57, 58, 62, 65, 53)
)
```

Es recomendable graficar los datos para observar la posible relación lineal:

```r
# Gráfico de dispersión
ggplot(datos, aes(x = Peso_Padres, y = Peso_Hijos)) +
  geom_point() +
  labs(title = "Relación entre el peso de padres e hijos",
       x = "Peso de padres (kg)",
       y = "Peso de hijos (kg)")
```



19.3.2 Ajuste del Modelo

Para ajustar el modelo, se utiliza la función `lm()`, cuya sintaxis general es:

```r
modelo <- lm(Y ~ X, data = datos)
```

En este caso:

```r
modelo <- lm(Peso_Hijos ~ Peso_Padres, data = datos)
```

Para obtener un resumen detallado del modelo, se emplea:

```r
summary(modelo)
```

```
Call:
lm(formula = Peso_Hijos ~ Peso_Padres, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.35052 -1.11314 -0.02222  0.64948  2.72024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.19857    4.37024   1.418    0.194
Peso_Padres  0.70708    0.05791  12.209 1.88e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.354 on 8 degrees of freedom
Multiple R-squared:  0.9491,    Adjusted R-squared:  0.9427
F-statistic: 149.1 on 1 and 8 DF,  p-value: 1.879e-06
```

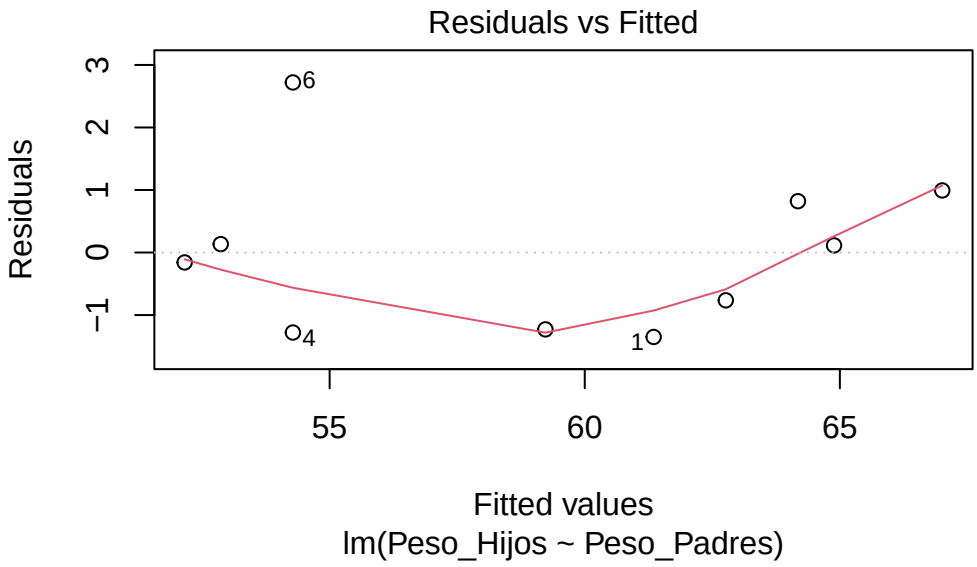
El resumen incluye los coeficientes estimados, sus errores estándar, valores t y p, así como el coeficiente de determinación (R^2), que indica la proporción de la variabilidad de Y explicada por X.

19.3.3 Evaluación Crítica de Supuestos

19.3.3.1 Supuesto de Linealidad

Se evalúa mediante el gráfico de residuos vs valores ajustados. Si los residuos se distribuyen aleatoriamente alrededor de cero, el supuesto se considera cumplido.

```r
plot(modelo, which = 1)  # Residuals vs Fitted
```

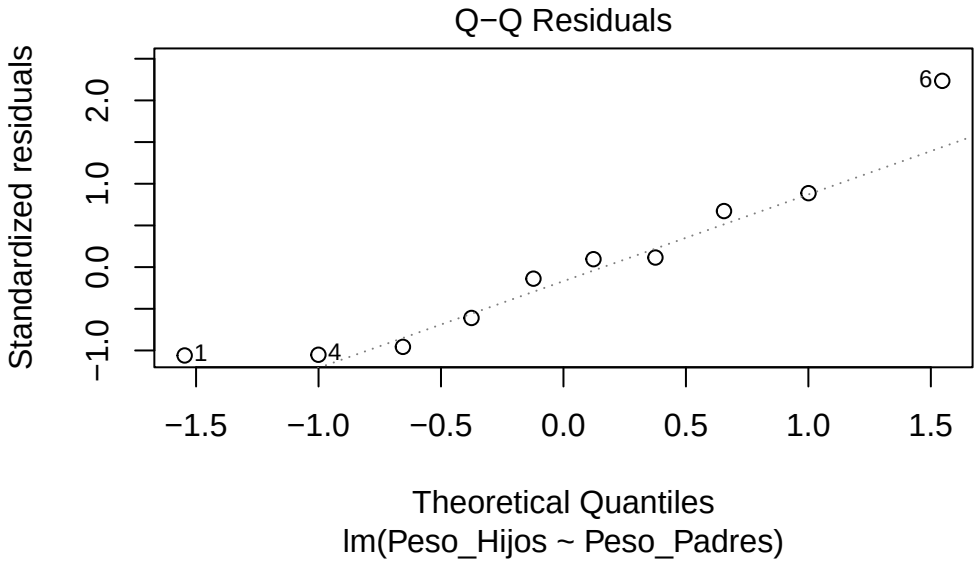


19.3.3.2 Supuesto de Normalidad

Se puede evaluar visualmente con un gráfico Q-Q y mediante pruebas estadísticas como Shapiro-Wilk y Anderson-Darling:

Gráfico Q-Q:

```
# Gráfico Q-Q
plot(modelo, which = 2)  # Normal Q-Q
```



Prueba de Shapiro-Wilk:

1. H_0 : Los residuos siguen distribución normal
2. H_a: Los residuos no siguen distribución normal

```
shapiro.test(residuals(modelo))
```

```
    Shapiro-Wilk normality test

data:  residuals(modelo)
W = 0.9049, p-value = 0.2478
```

Prueba de Anderson-Darling (más potente para muestras grandes):

```
ad.test(residuals(modelo))
```

```
    Anderson-Darling normality test

data:  residuals(modelo)
A = 0.36544, p-value = 0.3604
```

19.3.3.3 Supuesto de Homocedasticidad

Se evalúa con la **Prueba de Breusch-Pagan:**

1. H_0: Varianza constante (homocedasticidad)
2. H_a: Varianza no constante (heterocedasticidad)

```
bptest(modelo)
```

```
    studentized Breusch-Pagan test

data:  modelo
BP = 0.71286, df = 1, p-value = 0.3985
```

19.3.3.4 Supuesto de independencia

En estudios experimentales, la independencia suele garantizarse mediante un diseño adecuado. En estudios observacionales, se recomienda analizar el contexto y, si es posible, realizar pruebas adicionales.

19.4 Predicción con el modelo ajustado

Una vez ajustado el modelo, se pueden realizar predicciones para nuevos valores de la variable independiente:

```
# Nuevos valores de Peso_Padres
nuevos_pesos <- data.frame(Peso_Padres = c(60, 75, 80))

# Predicción con intervalos de predicción
predicciones <- predict(modelo, nuevos_pesos, interval = "prediction")
predicciones
```

```
       fit      lwr      upr
1 48.62315 44.77666 52.46963
2 59.22929 55.95375 62.50484
3 62.76467 59.42443 66.10492
```

El resultado incluye el valor predicho y los límites inferior y superior del intervalo de predicción para cada nuevo valor.

19.5 Interpretación de Resultados

19.5.1 Coeficientes del Modelo

1. **Intercepto** ($\hat{\beta}_0$): Valor esperado de Y cuando X = 0
2. **Pendiente** ($\hat{\beta}_1$): Cambio promedio en Y por unidad de cambio en X

19.5.2 Bondad de Ajuste

El **coeficiente de determinación** (R^2) indica la proporción de variabilidad explicada:

$$R^2 = \frac{SC_{Regresin}}{SC_{Total}}$$

1. $R^2 > 0.7$: Ajuste bueno
2. $-0.5 < R^2 < 0.7$: Ajuste moderado
3. $R^2 < 0.5$: Ajuste pobre

19.5.3 Significancia Estadística

La **prueba F global** evalúa:

1. H_0: $\beta_1 = 0$ (no hay relación lineal)
2. H_a: $\beta_1 \neq 0$ (existe relación lineal)

19.5.4 Criterios de Decisión para los supuestos

Supuesto	Prueba	Criterio de Aceptación
Normalidad	Shapiro-Wilk	p-valor > 0.05
Homocedasticidad	Breusch-Pagan	p-valor > 0.05
Linealidad	Gráfico residuos	Patrón aleatorio
Independencia	Contexto experimental	Diseño adecuado

19.5.5 Pasos para una Interpretación Integral y Conclusiones

1. **Evaluar significancia del modelo** (prueba F global)
2. **Verificar supuestos** mediante pruebas estadísticas y gráficos
3. **Interpretar coeficientes** en el contexto del problema
4. **Evaluar bondad de ajuste** (R^2 y R^2 ajustado)
5. **Identificar limitaciones** del modelo
6. **Formular recomendaciones** prácticas

Capítulo XII

Referencias

Referencias

Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to mathematical statistics* (8th ed.). Pearson.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314.

López, E., & González, B. (2018). *Notas de Estadística General* (Edición marzo 2018). Guatemala: Universidad de San Carlos de Guatemala.

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

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The definitive guide* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781138359444

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

Johnson, R. A., & Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson.

Montgomery, D. C. (2017). *Design and analysis of experiments* (9th ed.). Wiley.

Montgomery, D. C., & Runger, G. C. (2018). *Applied Statistics and Probability for Engineers* (7th ed.). Wiley.

Ross, S. M. (2014). *Introduction to probability and statistics for engineers and scientists* (5th ed.). Academic Press.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2014). *Mathematical statistics with applications* (7th ed.). Cengage.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists* (2nd ed.). Wiley.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability and Statistics for Engineers and Scientists* (9th ed.). Pearson.