Análisis de Datos en Agricultura con R y RStudio

Oswaldo Navarrete Carreño

Indice de contenidos

A modo de introducción

En la actualidad es díficil pensar en una ciencia que evolucione sin el soporte de la estadística, las ciencias agrícolas no son la excepción. La investigación y el desarrollo en la agricultura usan métodos y procedimientos estadísticos necesarios para resolver diferentes aspectos y problemas que forman parte de algunas ramas de la actividad agrícola (Kibuuka (2009)).

La evolución de las técnicas y las metodologías estadísticas se debe en gran medida a la rapidez con la que han evolucionado las computadoras y su capacidad de procesar grandes volúmenes de datos. En la actualidad aprender estadística sin el soporte de algún programa o lenguaje de programación orientado a la estadística, puede limitar tanto la posibilidad de abstraer conceptos importantes como la oportunidad de saber como aplicar el análisis estadístico en contextos reales.

Uno de los objetivos de este libro es que el futuro profesional de agronomía adquiera conocimientos importantes de estadística y construya una base sólida desde el punto de vista teórico y práctico que le permita desarrollar de forma autónoma análisis estadísticos a datos provenientes de fuentes diversas y relacionadas con sus áreas de desempeño.

1 R y RStudio

1.1 ¿Qué es R y por qué lo vamos a usar?

R es un lenguaje de programación que tiene más de 30 años de historia, mientras que RStudio es uno de los tantos entornos de desarrollo integrados que existen para trabajar con R. Es común escuchar las preguntas *¿por qué trabajar con R cuando existen programas mucho más "sencillos" para hacer análisis de datos?*, *¿para qué complicar a los estudiantes enseñándoles a programar si además deben aprender estadística?*. Existen muchas respuestas a esas preguntas, sin embargo, a continuación se presenta una lista de razones que fortalecen la idea de porqué usar R, algunas de estas ideas son expuestas en Tucker et al. (2022) y Agwu y Bialas (2018).

1. R es un programa libre, gratuito y de desarrollo independiente disponible para la mayoría de sistemas operativos lo que reduce las barreras de acceso.
2. Es flexible y poderoso tanto para simulación como para el análisis de datos.
3. El hecho de que involucre escribir código permite reproducir los análisis hechos por cualquier persona sin ninguna condición o limitante.

Escribir código en R ayuda a que el pensamiento estadístico y los procesos de análisis estadístico sean más visibles y reproducibles; es decir que usar R, podría ofrecer una herramienta adicional de representación que permite construir el entendimiento de los conceptos (Tucker et al. (2022)).

Cuando se instala R por defecto vienen incorporados *paquetes* que permiten importar datos, ajustar y evaluar modelos, realizar gráficos. Mas, es posible añadir funcionalidades a R instalando y cargando nuevos *paquetes*, un *paquete* es una colección de funciones de R, datos y código compilado en un formato bien definido, creado para agregar alguna funcionalidad específica (Data Carpentry (2020)). En la actualidad existen más de 19000 paquetes (y creciendo) para todo tipo de análisis estadístico, gran parte de los nuevos resultados o metodologías estadísticas publicadas en revistas de investigación generalmente son desarrolladas en paquetes de R lo que hace fácil acceder a las últimas técnicas estadísticas.

1.2 RStudio.

La interfaz de R es muy básica y para personas interesadas (u obligadas) a aprender este lenguaje de programación que no están familiarizados con algún otro lenguaje, usar esta interfaz

puede resultar frustrante al punto de querer desistir de aprender R. Rstudio es un entorno de desarrollo integrado (IDE) gratuito y de código abierto para R. Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para la graficación, revisar el historial, gestionar conexiones, la depuración del código y la gestión del espacio o directorio de trabajo (Navarrete y Chávez (2019)).

En Ismay (2022) se hace una analogía interesante en la que se menciona que R se puede entender como el motor de un vehículo, mientras que Rstudio es el tablero del vehículo que permite tener a la mano todas las funcionalidades del mismo.

1.3 Poniendo todo a punto para empezar

1.3.1 Instalación de R y RStudio

Un flujo de trabajo recomendado es, primero descargar e instalar R. El programa puede ser descargado de la página https://cloud.r-project.org/. En la Figura 1.1 se muestra la pantalla de la página. Una vez descargado el programa, se procede a instalarlo.

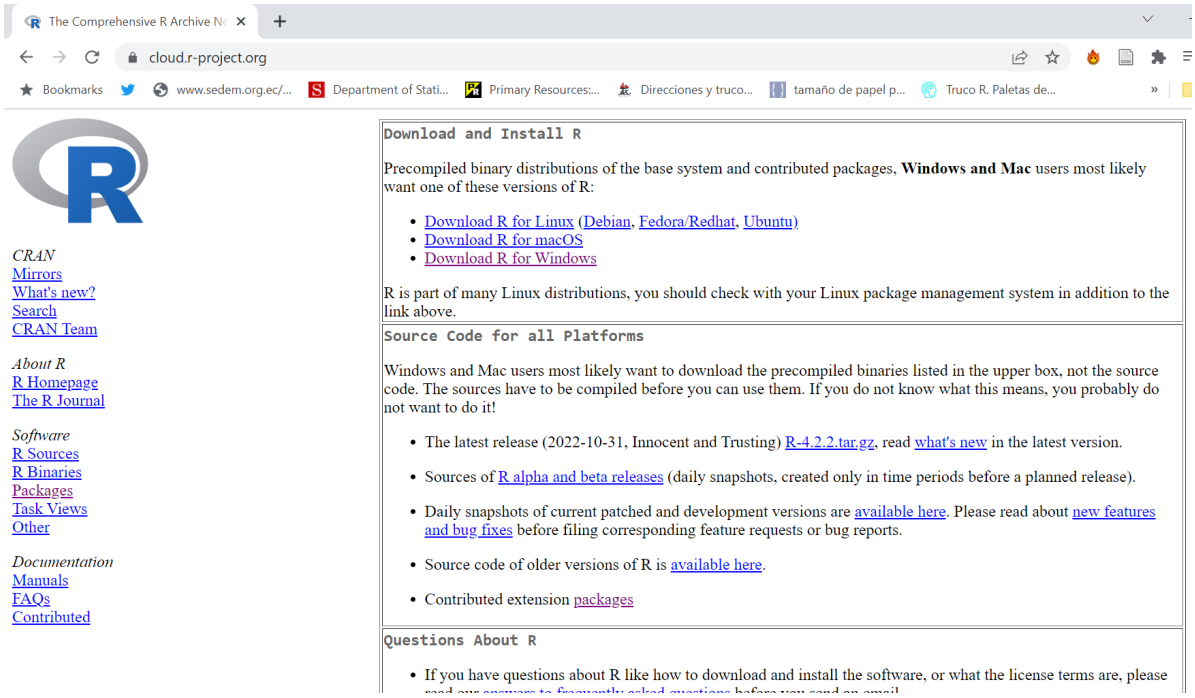


Figura 1.1: Página para descargar R

A continuación se procede a descargar RStudio de la página https://posit.co/downloads/. En la Figura 1.2 se muestra la pantalla de la página. Una vez descargado el programa se procede

a instalarlo.

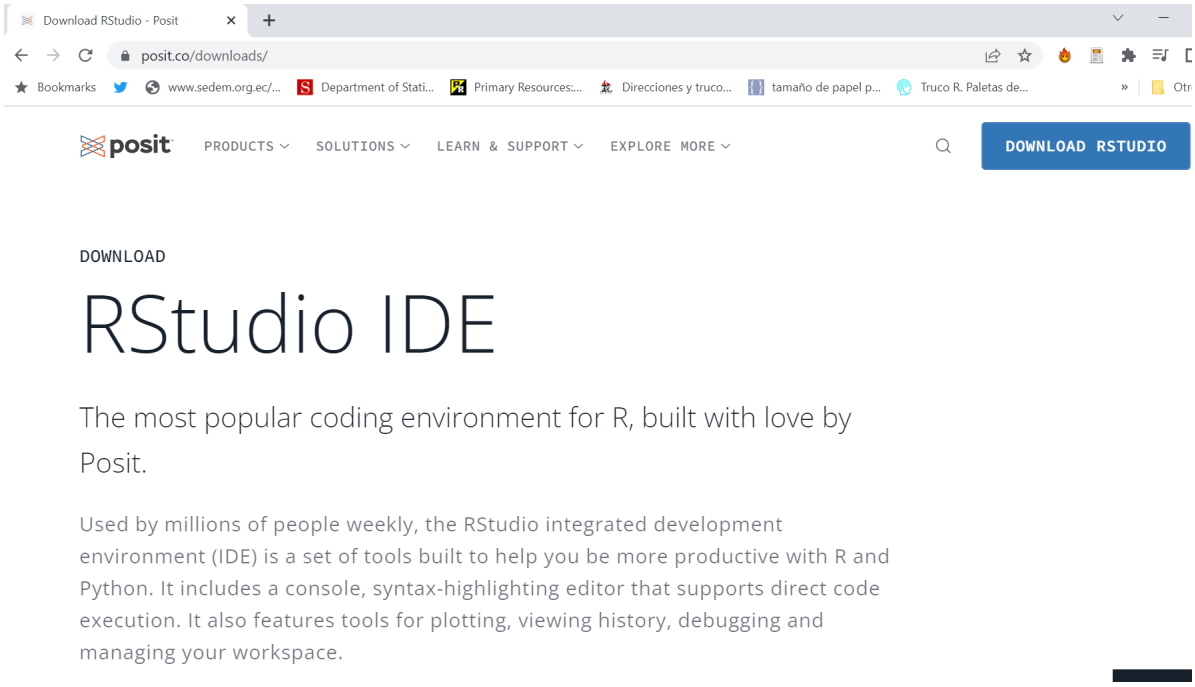


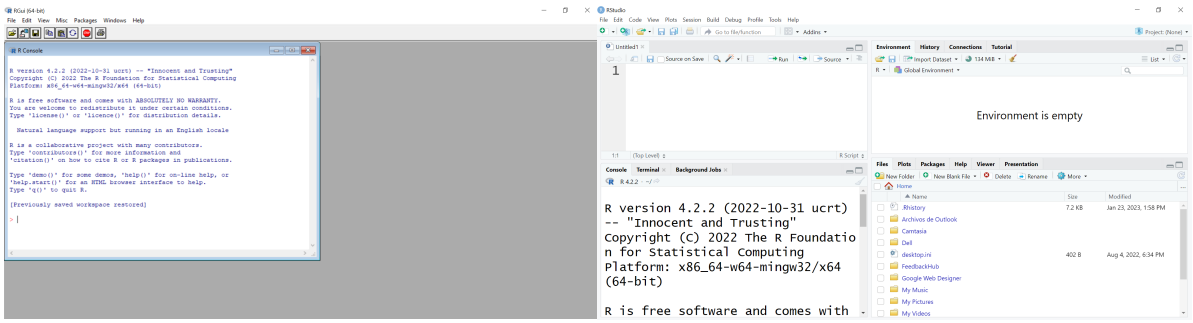
Figura 1.2: Página para descargar RStudio

En la Figura 1.3 se muestran las interfaces de R y RStudio respectivamente, a lo largo de este libro aprenderemos el lenguaje de programación R utilizando RStudio por lo que en la siguiente sección exploraremos la interfaz del segundo programa.

1.3.2 Conociendo la interfaz de RStudio

Cuando se abre RStudio por primera vez se observan 3 paneles. Sin embargo, cuando se comienza a usar el programa con frecuencia veremos 4 paneles. Por defecto en la parte superior izquierda aparece el *panel de fuente o scripts*, en la parte superior derecha el *panel de espacio de trabajo o ambiente*, en la parte inferior izquierda aparece la *consola* y en la parte inferior derecha aparece el *panel de archivos, gráficos, paquetes y ayuda*. En la Figura 1.4 se presenta la ubicación de estos cuatro paneles. A continuación una descripción de cada uno de estos paneles.

1. La consola de R, ubicada en la parte inferior izquierda, es donde se envían los comandos para ser ejecutados.
2. El panel de fuente o scripts, ubicado en la parte superior izquierda, es desde donde la mayoría de las veces se enviará el código a la consola donde R lo ejecutará. Para crear un script hay dos opciones. La primera es en la barra de herramientas principal escoger



(a) Interfaz de R (b) Interfaz de RStudio

Figura 1.3: Interfaces de R y RStudio

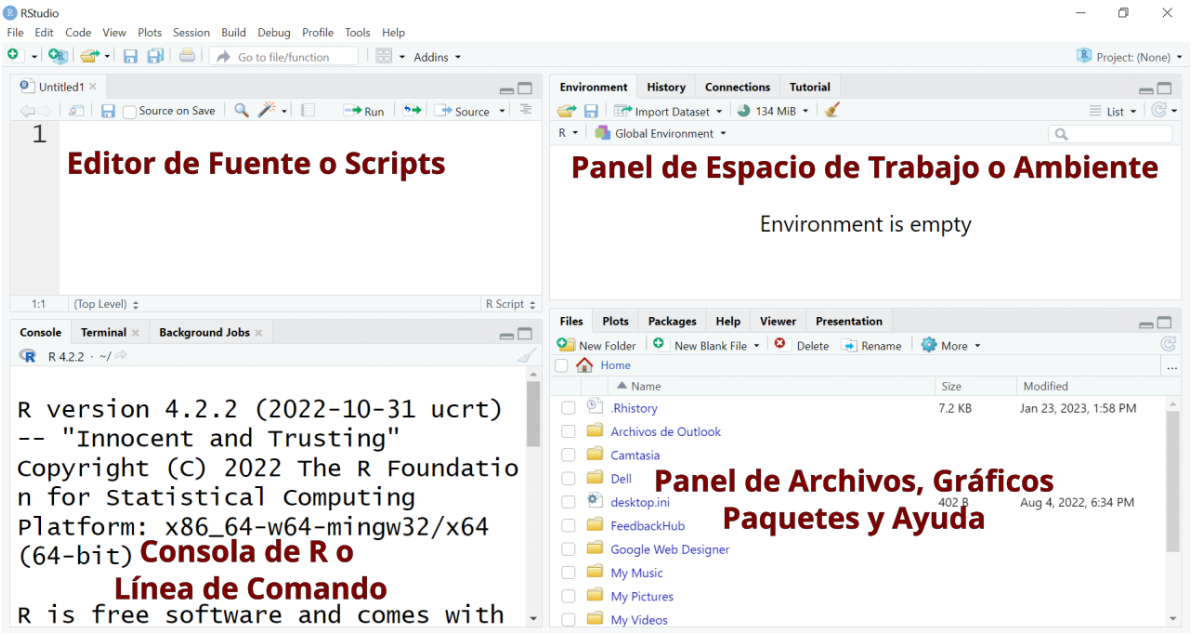


Figura 1.4: Paneles de RStudio

`File | New File | R Script`, la segunda opción es escribir la combinación de teclas (en Windows) `Ctrl + Shift + N` o `Cmd + Shift + N` (en Mac)

3. El panel de espacio de trabajo o ambiente, ubicado en la parte superior derecha, además del ambiente tiene pestañas para el historial, las conexiones y tutoriales. Sin embargo, solo detallaremos del ambiente y el historial. En el ambiente se van mostrando las variables, los objetos y los valores que se van creando o calculando. En el historial se muestran los comandos que se han enviado a R en la sesión de trabajo.

4. El panel de archivos, gráficos, paquetes y ayuda, ubicado en la parte inferior derecha, muestra los archivos que se encuentran en el directorio de trabajo, los gráficos que se van generando en la sesión y que no se almacenan como variables, los paquetes instalados tanto en la librería del sistema como del usuario y la ayuda para estos paquetes.

1.4 Instalar paquetes

A lo largo de este libro usaremos algunos paquetes. En esta sección aprenderemos a instalar paquetes. Existen 2 formas para instalar un paquete.

1. Desde la consola o un script
2. Desde la pestaña de paquetes o desde la barra de menú.

Antes de instalar los paquetes vamos a describir brevemente algunos de los paquetes que vamos a usar en este curso:

1. ***tidyverse***: El paquete **tidyverse** es considerado un metapaquete porque es un paquete que carga paquetes. Los paquetes que se cargan con el paquete **tidyverse** son:
    a. ***dplyr***: implementa una gramática para la manipulación de datos.
    b. ***forcats***: ofrece herramientas para trabajar con variables categóricas.
    c. ***ggplot2***: crea visualizaciones de datos elegantes utilizando una gramática de gráficos.
    d. ***lubridate***: permite trabajar con fechas de una forma sencilla.
    e. ***purrr***: herramientas de programación funcional.
    f. ***readr***: ayuda a leer datos de texto rectangulares.
    g. ***stringr***: ofrece funciones para manejo de cadenas de carácteres.
    h. ***tibble***: provee una versión de *data frame* que facilita el trabajo con *tidyverse*.
    i. ***tidyr***: herramientas para crear datos ordenados (*tidy data*)
2. ***agricolae***: procedimientos agrícolas para investigación agrícola.
3. ***readxl***: leer archivos de Excel.
4. ***writexl***: guardar conjuntos de datos y tablas en archivos de Excel.
5. ***jtools***: análisis y presentación de datos sociales y científicos.
6. ***DescTools***: herramientas para estadística descriptiva.

7. ***cowplot***: provee varias características que ayudan a la creación de figuras de alta calidad con ***ggplot2***.

1.4.1 Desde la consola o un script

Para instalar paquetes desde la consola o desde un script se debe utilizar la función `install.packages("nombrepaquete")` así si se quiere instalar el paquete **tidyverse** se escribe en la consola o en un script `install.packages("tidyverse")`, cuando se escribe en un script se debe ejecutar el comando con la combinación de teclas `Ctrl + Enter` si está trabajando en Windows o `Cmd + Enter` si está trabajando en Mac. En la Sección 3.3 se darán más detalles de como crear un script y ejecutar códigos desde un script, por lo pronto puede instalar paquetes desde la consola.

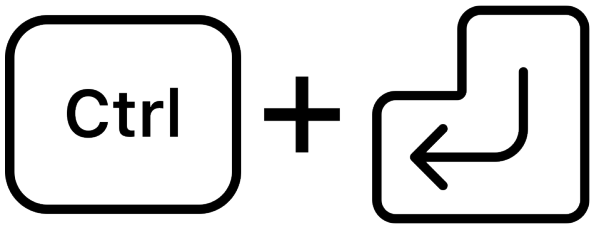


Figura 1.5: Combinación de teclas para ejecutar código desde un script (Windows)

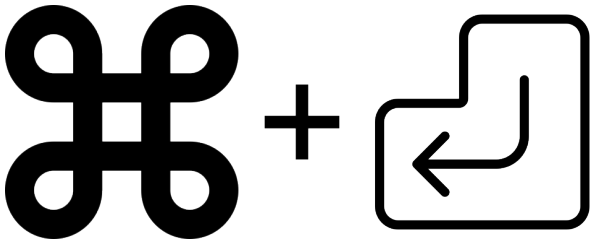


Figura 1.6: Combinación de teclas para ejecutar código desde un script (Mac)

1.4.2 Desde la pestaña de paquetes o desde la barra de menú

En el panel de archivos, gráficos, paquetes y ayuda se escoge la pestaña `Packages` y luego se da clic en la opción `Ìnstall`, aparece una ventana como la que se muestra en la Figura 1.7. En caja de texto titulada `Packages (separate multiple with space or comma)` se deben escribir el nombre del o de los paquetes que se van a instalar separados por coma o por un espacio como se muestra en la Figura 1.8. A medida que se escribe el nombre del paquete aparecen posibles paquetes a ser instalados. Otra forma de acceder a esta venta es en la barra de menú `Tools -> Install Packages...`.

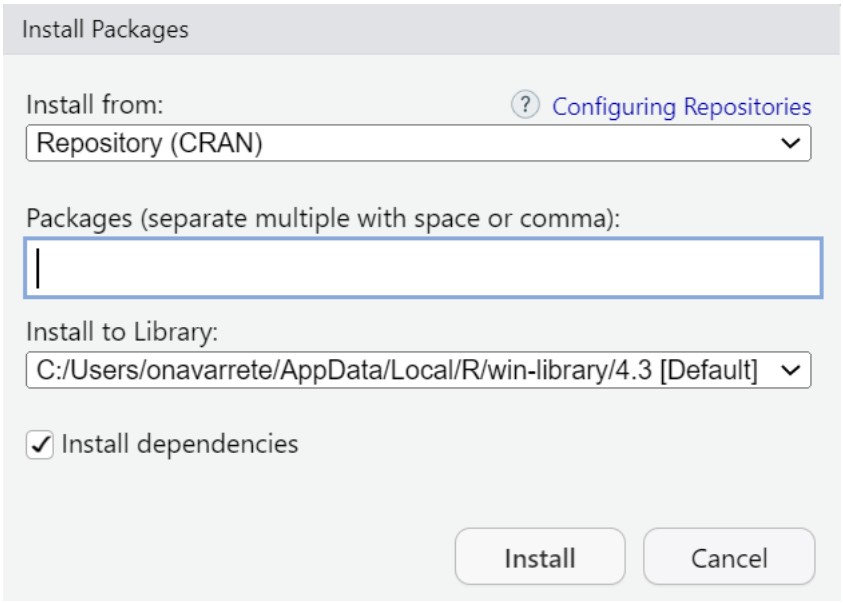


Figura 1.7: Ventana de Instalación de Paquetes

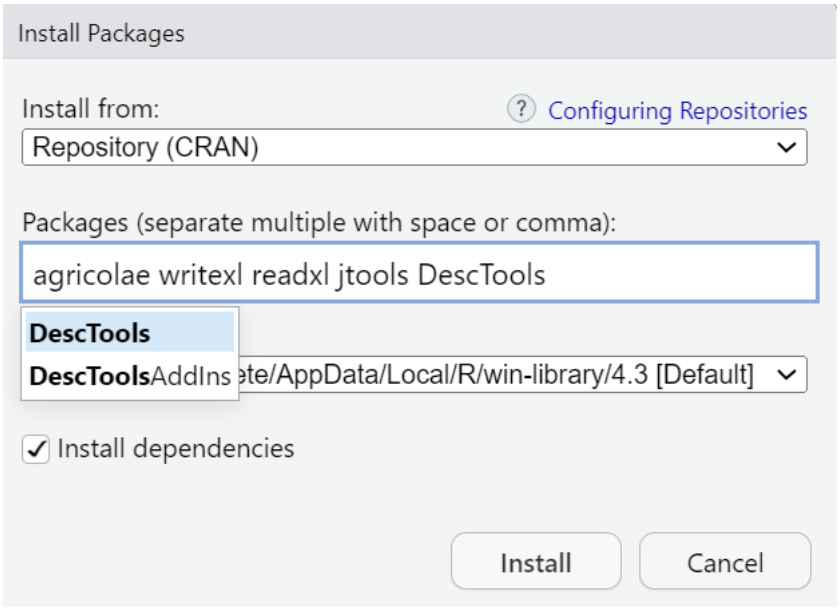


Figura 1.8: Ventana de Instalación de Paquetes

1.5 Los básicos del lenguaje R

1.5.1 Sintaxis de R

En la consola de R podemos escribir las siguientes expresiones

```
> x <- 6 # Crear la variable x con un valor igual a 6
```

```
> print(x) # Impresión explicita
```

```
[1] 6
```

```
> x # Auto-impresión
```

```
[1] 6
```

El símbolo `#` sirve para poner comentarios a nuestro código. Todo lo que se escriba a la derecha del numeral (incluido el símbolo) no se toma en cuenta. En R no se pueden poner comentarios multilíneas.

Cuando se escribe un comando en el **prompt**, se evalúa el comando y se obtiene el resultado del comando evaluado.

El resultado se puede **auto-imprimir** que es lo que ocurre cuando se escribe el nombre del objeto. El `[1]` indica que `x` es un vector y el `6` es el primer elemento. De manera general la **impresión explícita** con la función `print()` no se usa con frecuencia. Esta es una práctica que a veces es necesaria cuando se escriben scripts, funciones o programas largos.

Cuando se presenta un vector en R podremos ver que un índice del vector se imprime en corchetes `[]` en el lado izquierdo. En el siguiente ejemplo podemos fijarnos en esto. Los números que se muestran en los corchetes no son parte del vector, únicamente son parte del resultado impreso.

```r
x <- seq(20,200, by = 5)
x
```

```
 [1]  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95 100 105 110
[20] 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200
```

1.5.2 Objetos de R

R tiene cinco clases de objetos básicos, algunos autores también los llaman objetos atómicos.

1. Caracter. (character, chr)
2. Numérico. (numeric, num)
3. Entero. (integer, int)

4. Complejo. (complex, cplx)
5. Lógico. (logical, logi)

El objeto más básico de R es el **vector**. Para crear vectores vacíos se utiliza la función `vector()`. La regla única para crear vectores es: *Un vector debe tener solamente objetos de la misma clase*, la excepción de la regla son las listas que son un tipo especial de vectores que pueden contener elementos de diferentes clases. Las listas son un importante tipo de datos en R. La creación de una lista se lo hace con la función `list()`.

1.5.3 Vectores

Para crear vectores se utiliza la función `c()` que concatena o une los elementos.

```
v1 <- c(3, 5, 2.3) # Numérico
v2 <- c(TRUE, FALSE, FALSE, TRUE, TRUE) # Lógico
v3 <- c(T, F, F, T, T) # Lógico
v4 <- c(3+4i, 7-2i, 3i) # Complejo
v5 <- c(3, 2, 7) # Entero
```

Las letras `T` y `F` son abreviaturas para especificar `TRUE` (verdadero) o `FALSE` (falso). Con la función `vector()` se pueden inicializar vectores.

```
y <- vector("complex", length = 5)
y
```

```
[1] 0+0i 0+0i 0+0i 0+0i 0+0i
```

1.5.4 Coerción implícita y explícita

Cuando en un vector se mezclan objetos de distintas clases, ocurre la *coerción implícita* de tal forma que todos los elementos en el vector se convierten a elementos de la misma clase. Para verificar la clase de un vector usamos la función `class()`.

```
v6 <- c(3.14, "x")
class(v6)
```

```
[1] "character"
```

```
v7 <- c(T, 4, F)
class(v7)
```

```
[1] "numeric"
```

```
v8 <- c(F,"x")
class(v8)
```

```
[1] "character"
```

Por su parte la *coerción explicita* ocurre cuando se utiliza la función `as.*`, el `*` puede ser reemplazado por `numeric`, `character`, `logical`, `integer` y `complex`.

```
p <- seq(0, 14, 1.4)
class(p)
```

```
[1] "numeric"
```

```
as.integer(p)
```

```
 [1]  0  1  2  4  5  7  8  9 11 12 14
```

```
as.logical(p)
```

```
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
as.character(p)
```

```
 [1] "0"    "1.4"  "2.8"  "4.2"  "5.6"  "7"    "8.4"  "9.8"  "11.2" "12.6"
[11] "14"
```

```
as.complex(p)
```

```
 [1]  0.0+0i  1.4+0i  2.8+0i  4.2+0i  5.6+0i  7.0+0i  8.4+0i  9.8+0i 11.2+0i
[10] 12.6+0i 14.0+0i
```

En ciertas ocasiones no es posible para R coercionar un objeto en la clase que se desea y esto resulta en que se produzcan `NAs`.

```r
q <- c("R","S","T")
class(q)
```

```
[1] "character"
```

```r
as.numeric(q)
```

```
Warning: NAs introducidos por coerción
```

```
[1] NA NA NA
```

```r
as.complex(q)
```

```
Warning: NAs introducidos por coerción
```

```
[1] NA NA NA
```

```r
as.integer(q)
```

```
Warning: NAs introducidos por coerción
```

```
[1] NA NA NA
```

```r
as.logical(q)
```

```
[1]   NA   NA TRUE
```

1.5.5 Listas

Las listas son un tipo especial de vectores que pueden contener elementos de distintas clases. Las listas se pueden crear con la función `list( )`. Con la función `vector( )` se puede crear una lista vacia.

```r
r <- list(2.71, 2, "m", 3+2i, F)
r
```

```
[[1]]
[1] 2.71

[[2]]
[1] 2

[[3]]
[1] "m"

[[4]]
[1] 3+2i

[[5]]
[1] FALSE
```

```r
s <- vector("list", length = 5)
s
```

```
[[1]]
NULL

[[2]]
NULL

[[3]]
NULL

[[4]]
NULL

[[5]]
NULL
```

1.5.6 Matrices

Las matrices son vectores que tienen un atributo de dimensión, este atributo es en si mismo un vector entero de longitud 2 (número de filas y columnas).

```r
m <- matrix(nrow = 3, ncol = 4)
m
```

```
     [,1] [,2] [,3] [,4]
[1,]   NA   NA   NA   NA
[2,]   NA   NA   NA   NA
[3,]   NA   NA   NA   NA
```

```r
dim(m)
```

```
[1] 3 4
```

```r
attributes(m)
```

```
$dim
[1] 3 4
```

Las matrices son construidas por columna, de tal forma que los elementos de la matriz pueden comenzar a ser llenados desde la esquina superior izquierda recorriendo las columnas.

```r
m <- matrix(1:12, nrow = 3, ncol = 4)
m
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

Otra forma de crear matrices es a partir de vectores añadiéndole los atributos de dimensión.

```r
m <- 1:12
m
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```r
dim(m) <- c(3,4)
m
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

Una forma común de crear matrices es uniendo filas o columnas con las funciones `rbind` o `cbind`.

```r
s1 <- 4:8
s2 <- 16:20
cbind(s1, s2) # unión por columnas
```

```
     s1 s2
[1,]  4 16
[2,]  5 17
[3,]  6 18
[4,]  7 19
[5,]  8 20
```

```r
rbind(s1, s2) # unión por filas
```

```
   [,1] [,2] [,3] [,4] [,5]
s1    4    5    6    7    8
s2   16   17   18   19   20
```

1.5.7 Data frames

Los *data frames* (marcos de datos, cuadros de datos) son usados para guardar datos tabulares en R. Son un tipo importante de objetos en R y son usados en muchas situaciones de modelación y análisis de datos. Los *data frames* a diferencia de las matrices pueden guardar diferentes clases de objetos en cada columna. Las matrices tienen un atributo de nombres de columnas (`colnames`) los *data frames* tienen además un atributo de nombres de filas (`row.names`).

Los *data frames* pueden ser leídos con funciones como `read.csv()` o `read.table()`. También pueden ser creados con la función `data.frame()`, el metapaquete `tidyverse` tiene funciones que permiten trabajar de forma más eficiente con conjuntos de datos. El equivalente de

`data.frame()` en el paquete `dplyr` de `tidyverse` es `tibble()`. Con la función `str()` se puede conocer la estructura de un *data frame*, la función de *dplyr* equivalente a *str()* es *glimpse( )*

```
df1 <- data.frame(
  Variable = c("Var1","Var2","Var3","Var4"),
  Valor = c(12, 14, 13, 16)
 )
df1
```

```
  Variable Valor
1     Var1    12
2     Var2    14
3     Var3    13
4     Var4    16
```

```
str(df1)
```

```
'data.frame':   4 obs. of  2 variables:
 $ Variable: chr  "Var1" "Var2" "Var3" "Var4"
 $ Valor   : num  12 14 13 16
```

```
glimpse(df1)
```

```
Rows: 4
Columns: 2
$ Variable <chr> "Var1", "Var2", "Var3", "Var4"
$ Valor    <dbl> 12, 14, 13, 16
```

```
nrow(df1)
```

```
[1] 4
```

```
ncol(df1)
```

```
[1] 2
```

```r
df2 <- tibble(
  Variable = c("Var1","Var2","Var3","Var4"),
  Valor = c(12, 14, 13, 16)
 )
df2
```

```
# A tibble: 4 x 2
  Variable Valor
  <chr>    <dbl>
1 Var1        12
2 Var2        14
3 Var3        13
4 Var4        16
```

```r
str(df2)
```

```
tibble [4 x 2] (S3: tbl_df/tbl/data.frame)
 $ Variable: chr [1:4] "Var1" "Var2" "Var3" "Var4"
 $ Valor   : num [1:4] 12 14 13 16
```

```r
glimpse(df2)
```

```
Rows: 4
Columns: 2
$ Variable <chr> "Var1", "Var2", "Var3", "Var4"
$ Valor    <dbl> 12, 14, 13, 16
```

```r
nrow(df2)
```

```
[1] 4
```

```r
ncol(df2)
```

```
[1] 2
```

2 Estadística. Conceptos Básicos.

El término *estadística* tiene su origen en la palabra alemana *statistik* acuñada por el economista aleman Gottfried Achenwall, la palabra tuvo su génesis en el latín *statisticus* que se podría traducir como *relativo al estado*, pues lo que hoy conocemos como estadística estuvo incialmente relacionado con datos relativos a los gobiernos o estados, dichos datos podían provenir de censos, registros de salud, pago de impuestos, etc. En los siglos XVI y XVII hubo un fuerte interés por las probabilidades, más que nada por el deseo de obtener ventaja en juegos de cartas y otras formas de apuesta.

La ciencia estadística como tal evolucionó al punto que hoy encontramos estadísticas y estadística en muchos aspectos de nuestra vida cotidiana. Cuando hablamos de estadística nos referimos a la ciencia (que definiremos formalmente un poco más adelante), mientras que al usar el termino estadísticas nos referimos a datos, o a la presentación de hechos, gráficas o información. Por ejemplo cuando en una hacienda cacaotera se presenta la información relativa a la producción en kilogramos de los híbridos usados en la plantación estamos hablando de estadísticas, mientras que el problema de analizar, y establecer las diferencias en el rendimiento medido en kilogramos por hectárea para diferentes híbridos con el fin de aumentar las ganancias de la finca puede ser resuelto usando estadística.

Una definición clásica de estadística es **ciencia que se encarga de la recolección, clasificación, resumen, organización, presentación, análisis e interpretación de información numérica** (Cleff (2013)). En Agresti et al. (2023), de manera corta definen a la estadística como **el arte y la ciencia de aprender de los datos**.

> **!** Importante
>
> **Definición de Estadística** ciencia que trata de la recolección, presentación y análisis de datos para resolver problemas, tomar decisiones y diseñar productos y procesos.

2.1 Etapas del proceso estadístico.

El proceso estadístico tiene como principal entrada a los datos. Sin embargo, el proceso estadístico no empieza en los datos. De manera general las etapas del proceso estadístico son:

1. **Identificación y delimitación del problema**: Un problema en investigación debe ser entendido como una pregunta o cuestión sobre un tema o aspecto que no se conoce. Ejemplos de problema en agronomía pueden ser: comparación de producción lechera en vacas sometidas a diferentes dietas, la identificación de características fenotípicas de variedades de banano resistentes a determinada enfermedad, evaluación del rendimiento de un cultivar utilizando biocarbón en diferentes concentaciones, etc. La delimitación implica recortar el tema dentro de los límites de espacio, tiempo y área de estudio que caracteriza al problema. Por ejemplo, en el problema de la comparación de la producción lechera en vacas sometidas a diferentes dietas la delimitación puede incluir aspectos como, la época del año en que se va a recoger la información, las razas de animales a ser incluidas en el estudio, el lugar dónde se llevará a cabo el estudio. Dentro de esta fase se incluye la definición de los objetivos del estudio.
2. **Planificación**: Definir bien los objetivos ayuda a realizar una planificación adecuada. A partir de los objetivos se puede identificar los datos que se necesitan, el tipo de datos, la fuente de los datos, la frecuencia de recolección de los datos, la forma de recolección y almacenamiento de los datos y los programas o paquetes estadísticos que serán usados en el análisis.
3. **Recolección de datos**: la recolección puede ser realizada de fuentes existentes como bases de datos de organismos como la FAO, ministerios de agricultura, cámaras de productores, etc. O los datos pueden provenir de experimentos diseñados para el propósito específico de cumplir con los objetivos planteados.
4. **Revisión de los datos**: una vez que los datos son recogidos, es necesario revisarlos para verificar inconsistencias o problemas potenciales. Por ejemplo, al registrar las longitudes de hojas de cacao de ciertas accesiones es posible que ocurran cosas como valores mal ingresados. En estudios donde los datos son obtenidos de organismos como la FAO es probable que los datos hayan sido ingresados con un formato no apropiado.
5. **Tabulación** la tabulación permite dar una primera mirada a los datos y ayuda a obtener información sobre los datos. Esta tabulación puede ser presentada ya sea como tablas o como gráficos, lo que será discutido más adelante.
6. **Análisis estadístico**: en este paso se obtienen medidas de tendencia central, dispersión, posición, o asociación. Se hacen pruebas de hipótesis, o se realizan modelos con el fin de contestar las preguntas o cumplir los objetivos del estudio.
7. **Inferencia**: Basados en los resultados del análisis estadístico se hacen inferencias acerca de los objetivos del estudio. Aunque los datos provengan de una muestra, utilizando teorías estadísticas se puede llegar a conclusiones sobre la población de la que fueron tomadas las muestras.

2.2 Definiciones importantes

- **Población**: una población es el conjunto de todos los sujetos u objetos de interés en una investigación o análisis. Por ejemplo, en un estudio que tenga como objetivo determinar

características fenotípicas de variedades de banano resistentes a cierta enfermedad en el trópico húmedo, la poblacion son todas las plantas de banano sembradas en el trópico húmedo.

- **Muestra**: es la parte de la población que es analizada, dicho de otra forma una muestra es un subconjunto de la población. Sigamos con el ejemplo de las características fenotípicas de variedades de banano resistentes a cierta enfermedad en el trópico húmedo, el investigador no va a tomar todas las plantas sino que va a escoger un grupo de plantas de las variedades de interés. La muestra debe representar lo mejor posible a la población. La parte de la estadística que comprende los métodos estadísticos para obtener muestras representativas de una población se llama *muestreo*
- **Parámetro**: un parámetro es una cantidad numérica que caracteriza a una población.
- **Estadístico**: un estadístico es una cantidad numérica que caracteriza a una muestra.

2.3 Clasificación de la estadística.

De manera general, la estadística se clasifica en dos grandes ramas : estadística descriptiva y estadística inferencial.

> **!** Importante
>
> **Estadística Descriptiva** se refiere a los métodos para resumir los datos recogidos, estos datos pueden provenir de una muestra o de una población. Los resumenes de los datos pueden ser tablas, gráficos, y números como promedios o porcentajes.

> **!** Importante
>
> **Estadística Inferencial** se refiere a los métodos para sacar conclusiones, hacer predicciones o tomar decisiones sobre una población, basándose en datos de una muestra perteneciente a una población.

En la Figura 2.1 se muestra la relación entre algunos conceptos revisados hasta ahora. Supongamos que en una finca lechera se desea hacer un estudio sobre el efecto de una dieta en la producción de leche en una determinada raza de vacas lecheras. El conjunto de todas las vacas representa a toda la **población**. El objetivo de estudio es determinar el rendimiento promedio de todas las vacas, esta medida de toda la población es un **parámetro**. Debido a ser un producto de prueba el estudio no es aplicado a toda la población, por lo que se utiliza técnicas de **muestreo** para escoger una **muestra** representativa, con **estadística descriptiva** se determina el rendimiento promedio de la muestra, este valor es un **estadístico**. Con técnicas de **estadística inferencial** se estima el valor del parámetro de interés.

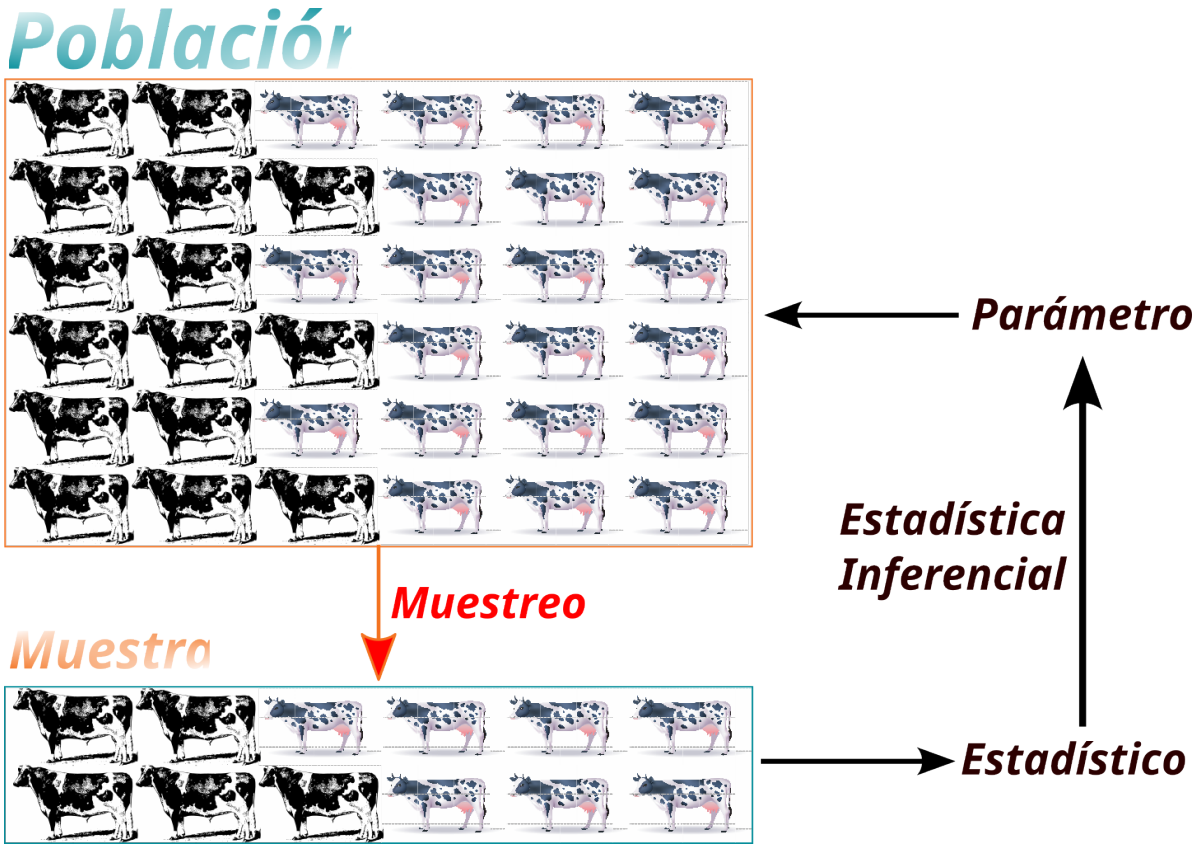


Figura 2.1: Relación entre algunos conceptos estadísticos

2.4 Tipos de estudio

Supongamos que un investigador desea analizar las tendencias de las exportaciones de maíz durante la última década en Latinoamérica y el Caribe, puede obtener esta información de la página FAOSTAT https://www.fao.org/faostat/en/#home. En este caso los datos no son recogidos directamente por el investigador sino que usa una fuente de información secundaria. Ahora supongamos que el mismo investigador quiere conocer la impresión de las medidas tomadas por el gobierno para ayudar a los pequeños productores de maíz, para registrar esto puede realizar una encuesta a una muestra de pequeños productores de maíz. Tanto al tomar datos de una fuente secundaria o al realizar una encuesta el investigador está observando los datos como aparecen y no está interviniendo en el problema que se investiga. Este tipo de estudios recibe el nombre de **observacional**

> **!** Importante
>
> **Estudio observacional** se observan los datos como aparecen y no se interviene en el problema o fenómeno investigado. Un estudio observacional puede ser *analítico* o *descriptivo*

Ahora imaginemos que se desea evaluar los parámetros nutricionales y fermentativos de un tipo de pasto, así como la inclusión de diferentes niveles de maíz molido en el proceso de ensilaje. El investigador diseña un experimento utilizando microsilos experimentales para evaluar el efecto de incluir diferentes porcentajes de maíz molido en el ensilaje del pasto, el diseño de este experimento incluye la determinación del número de tratamientos y repeticiones. En este caso el investigador está manipulando los porcentajes de maíz molido para evaluar el efecto sobre los parámetros nutricionales y fermentativos del pasto. Este tipo de estudios recibe el nombre de **experimental**

> **!** Importante
>
> **Estudio experimental** se interviene de manera intencionada y programada para manipular una o más variables con el fin de analizar los efectos de estas manipulaciones en otras variables

2.5 Variables, clasificación y niveles de medición.

En un estudio sobre características fenotípicas de girasoles cultivadas en el trópico húmedo se registra información sobre el número de planta, el número de parcela donde se encuentra la planta, la variedad de girasol observada, el número de fila, la altura de las plantas, la circunferencia del tallo, el color de las flores, el número de pétalos de las flores y número de

hojas. Estas características, medidas o valores de interés para la persona que investiga reciben el nombre de **variables**.

En la variable **Variedad de Girasol** se puede almacenar respuestas como Almonte, Ariadna, Candela, etc. Mientras que en la variable **Color de las flores** se puede registrar valores como Marfil, Amarillo pálido, Naranja, Púrpura, Rojo o Multicolor. Estas dos variables son **variables cualitativas**, pues describen una cualidad o categoría.

En las variables **altura de las plantas**, **circunferencia del tallo**, **número de pétalos de las flores** y **número de hojas** se almacenan respuestas numéricas. Estas variables son ejemplo de **variables cuantitativas**. Para la altura de las plantas en centímetros ejemplos de respuestas son 75.23, 92.15, etc. Mientras que para el número de pétalos o para el número de hojas las respuestas pueden ser, por ejemplo, 10, 5, 20, etc. Para el caso de la altura, como se observa, aceptamos respuestas con decimales, en este caso la variable es **cuantitativa continua**. La variable número de hojas es un ejemplo de variable **cuantitativa discreta** pues es una variable que solo acepta números enteros. En la Figura 2.2 se presenta un resumen de la clasificación de las variables.

Las variables además presentan **niveles de medición**, estos pueden ser:

1. Ordinal: toman valores que se ordenan o clasifican de forma lógica, ejemplos de variables con este nivel de medición pueden ser **fila en que se encuentra una planta**, **nivel de aceptación de un producto (muy alto, alto, medio, bajo, muy bajo)**
2. Nominal: toman valores que no se ordenan o clasifican de forma lógica, ejemplos de variables con este nivel de medición pueden ser **variedad de semilla de maíz sembrada**, **raza de ganado sometida a una dieta**.
3. Intervalo: existe diferencia significativa entre valores pero el cero no representa la ausencia de la característica, es decir que el cero no es significativo, ejemplos de este nivel de medición son **número de cajas de papaya exportadas en una semana**, **calificación obtenida en un examen**.
4. Razón: el 0 es significativo y la razón entre dos números es significativa, ejemplos de este nivel de medición son **edad**, **altura de una planta**, **circunferencia del tallo de una planta**.

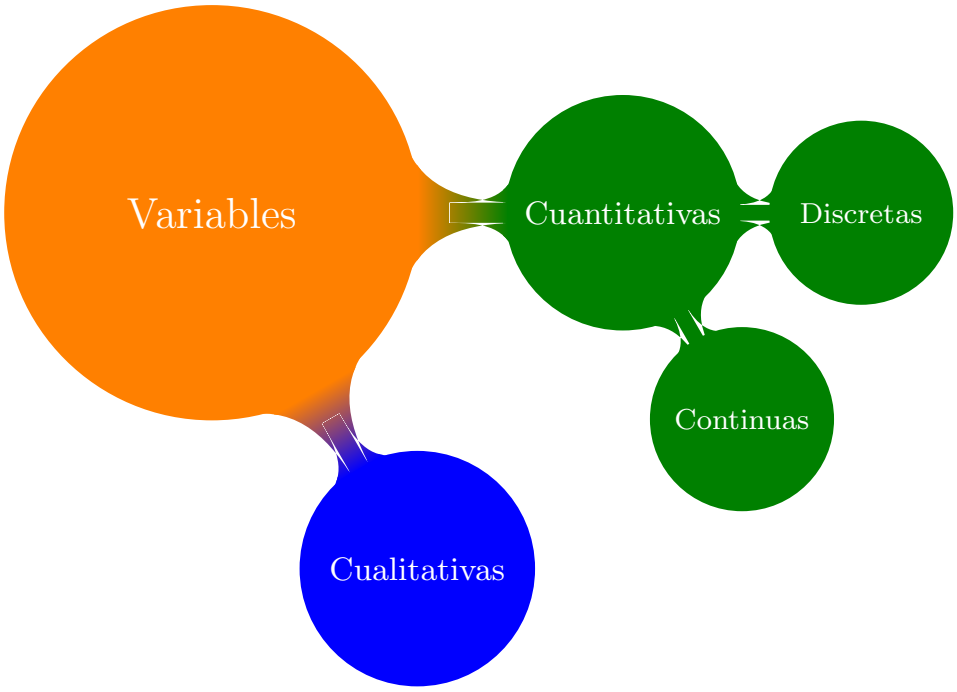


Figura 2.2: Clasificación de las variables

Ejercicios

Ejercicio 2.1. Clasifique las siguientes variables en cuantitativas o cualitativas

a. Diámetro del tallo de una planta de banano.
b. Número de pétalos de una flor de girasol.
c. Altura de una planta de tomate.
d. Perímetro toráxico de un bovino
e. Razas de ganado existentes en una finca lechera.
f. Variedad de semillas usadas en un cultivo de banano.
g. Cantidad de fertilizante utilizado en una plantación.
h. Cajas de banano vendidas por semana.
i. Precipitación en milímetros de una región por año.
j. Región en la que se siembra cierta variedad de piña.
k. Fila en la que se encuetra una planta dentro de una parcela.
l. Temperatura medida en grados Celsius
m. Temperatura medida en grados Kelvin

Ejercicio 2.2. Clasifique las variables cuantitativas del Ejercicio 2.1 como continuas o discretas.

Ejercicio 2.3. Determine el nivel de medición de las variables cuantitativas del Ejercicio 2.1.

3 Estadística Descriptiva

3.1 Tablas y gráficos para resumir datos

Un primer paso en el análisis de datos es resumir los datos. En esta sección exploraremos dos formas de resumir la información usando tablas o gráficos.

Las tablas y los gráficos se escogen de acuerdo al tipo de variable que está siendo analizada. Empezaremos por revisar las tablas para variables cualitativas y cuantitativas.

3.1.1 Tablas para variables cualitativas

Una tabla o tabla de frecuencias para variables cualitativas o categóricas generalmente tiene 3 columnas:

1. **Categoría**: muestra el nombre de cada categoría. Vamos a suponer que en el conjunto de datos para nuestra variable de interés existen k categorías.
2. **Frecuencia**: (f_i) la frecuencia o frecuencia absoluta corresponde al número de observaciones correspondientes a la categoría, la suma de todas las frecuencias absolutas debe ser igual al número de observaciones n. $\sum_{i=1}^{k} f_i = n$

> **!** Importante
>
> El símbolo $\sum$ se utiliza para representar la sumatoria de un conjunto de elementos. En este caso $\sum_{i=1}^{k} f_i$ se lee sumatoría desde i igual a 1 hasta k de f_i.

3. **Porcentaje**: el porcentaje se lo obtiene multiplicando la frecuencia relativa h_i por 100. La frecuencia relativa es igual a la frecuencia absoluta f_i dividida por el número n de observaciones. La frecuencia relativa también recibe el nombre de proporción. $h_i = \frac{f_i}{n}$

Supongamos que en un estudio se utilizaron 50 semillas de 3 variedades de tomate. En la Tabla 3.1 se muestra un resumen del número de semillas usadas de cada variedad, esta tabla es un ejemplo de una tabla para variables cualitativas. De la tabla se deduce que la menor proporción de semillas utilizadas corresponde a la variedad 2, mientras que la mayor proporción de semillas corresponde a la variedad 1.

Tabla 3.1: Ejemplo de tabla para variables cualitativas

Categoría	Frecuencia	Porcentaje
Variedad 1	20	40
Variedad 2	14	28
Variedad 3	16	32

3.1.2 Tablas para variables cuantitativas

Una tabla de frecuencias para variables cuantitativas tiene 6 columnas:

1. Clase: una clase es un intervalo semi abierto o semicerrado con la forma

$$[\text{Límite Inferior}, \text{Límite Superior})$$

2. Marca de Clase X_i: es un valor igual al promedio de los dos límites de la clase.

$$X_i = \frac{\text{Límite Superior} + \text{Límite inferior}}{2} \tag{3.1}$$

3. Frecuencia f_i: la frecuencia es igual al número de observaciones de la variable que están dentro del intervalo o clase. También se la conoce como frecuencia absoluta.
4. Frecuencia relativa h_i: la frecuencia relativa se la calcula como la frecuencia dividida para el total de valores de la variable.

$$h_i = \frac{f_i}{n} \tag{3.2}$$

5. Frecuencia acumulada F_i: se la calcula sumando las frecuencias desde la primera clase hasta la clase en consideración.

$$F_i = \sum_{j=1}^{i} f_j \tag{3.3}$$

6. Frecuencia Relativa acumulada H_i: se la calcula sumando las frecuencias relativas desde la primera clase hasta la clase en consideración.

$$F_i = \sum_{j=1}^{i} h_j \tag{3.4}$$

Exiten expresiones matemáticas que nos permiten conocer el número de clases necesarias y la amplitud de las clases sin embargo, por el enfoque de este texto en algunas ocasiones dejaremos eso a criterio del investigador o trabajaremos con las clases que por defecto obtengamos de las funciones que aprenderemos a usar de R.

El ingeniero Yorky Gil realizó un experimento con el fin de estimar una curva fenológica para el cultivo de girasol en condiciones del trópico humedo. Recogió datos durante 8 semanas de la altura en centímetros y el número de hojas de las plantas de girasol de determinada variedad. En la Tabla 3.2 se muestra la distribución de frecuencias de la altura para la semana 8. Este es un ejemplo de una tabla de frecuencias para variables cuantitativas. En secciones posteriores discutiremos paso a paso la elaboración de esta tabla.

De la interpretación de esta tabla se obtienen datos importantes. Si nos fijamos en la sexta clase podemos afirmar que en la semana 8, 78 observaciones es decir el 32.4% de las plantas tenían una altura entre 130 y 154 centímetros. De la misma clase podemos decir que 196 plantas o el 81.3% tienen una altura menor a 154 centímetros. En la última clase se observa que la frecuencia absoluta acumulada es de 241, este número corresponde al 100% de las observaciones.

Tabla 3.2: Ejemplo de tabla de frecuencia para variables cuantitativas

Clase	Marca de Clase	Frec.	Frec. Rel.	Frec. Acu.	Rel. Acu.
[10,34)	22	2	0.008	2	0.008
[34,58)	46	6	0.025	8	0.033
[58,82)	70	21	0.087	29	0.120
[82,106)	94	39	0.162	68	0.282
[106,130)	118	50	0.207	118	0.490
[130,154)	142	78	0.324	196	0.813
[154,178)	166	29	0.120	225	0.934
[178,202)	190	13	0.054	238	0.988
[202,226)	214	3	0.012	241	1.000

3.1.3 Gráficos

3.1.3.1 Diagrama de barras

Un gráfico de barras es la representación visual de una tabla de distribución de frecuencias. En el eje de las x se colocan los niveles de la variable cualitativa y en el eje de las y los valores de las frecuencias o de los porcentajes. En la Figura 3.1 se presenta el gráfico de barras para la Tabla 3.1. Mientras que en la Figura 3.2 se muestra la relación entre estos dos objetos. En general cuando se tiene una variable categórica **única** que se quiere desglosar y cuantificar por cada categoría este es el tipo de gráfico adecuado.

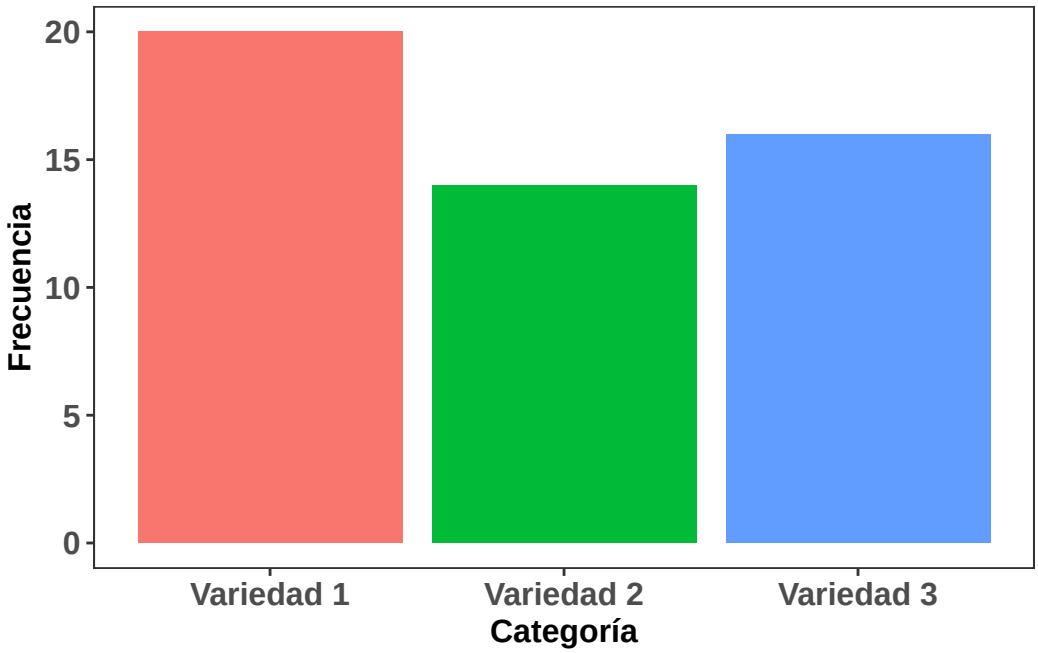


Figura 3.1: Gráfico de Barras para la Tabla 3.1

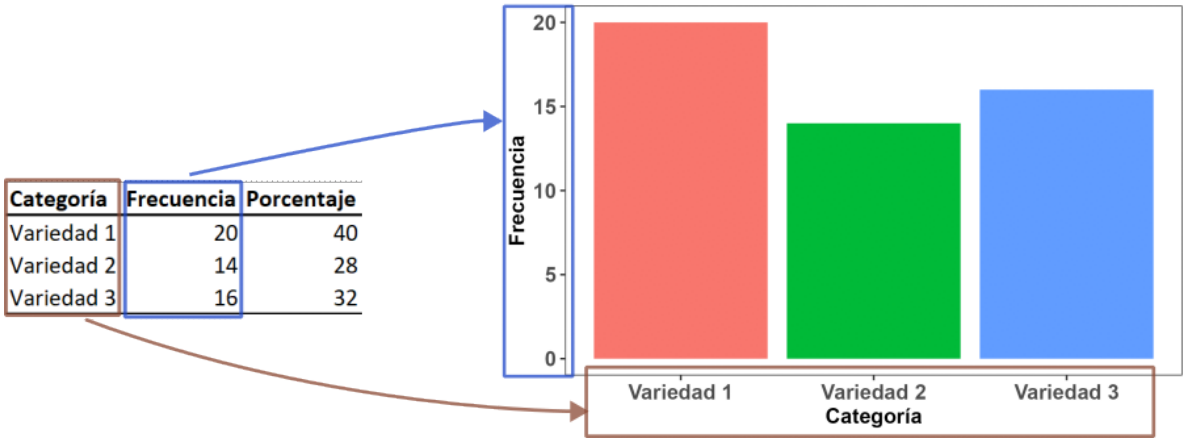


Figura 3.2: Relación entre la Tabla 3.1 y la Figura 3.1

3.1.3.2 Diagrama de barras agrupadas

En un diagrama de barras agrupadas, al igual que en un diagrama de barras, se muestran los conteos para un grupo, pero además se desglosa para una variable cuantitativa adicional.

Por ejemplo en la Figura 3.3 se muestran los porcentajes de nitrógeno, fosforo y potasio para cuatro variedades diferentes de maíz.

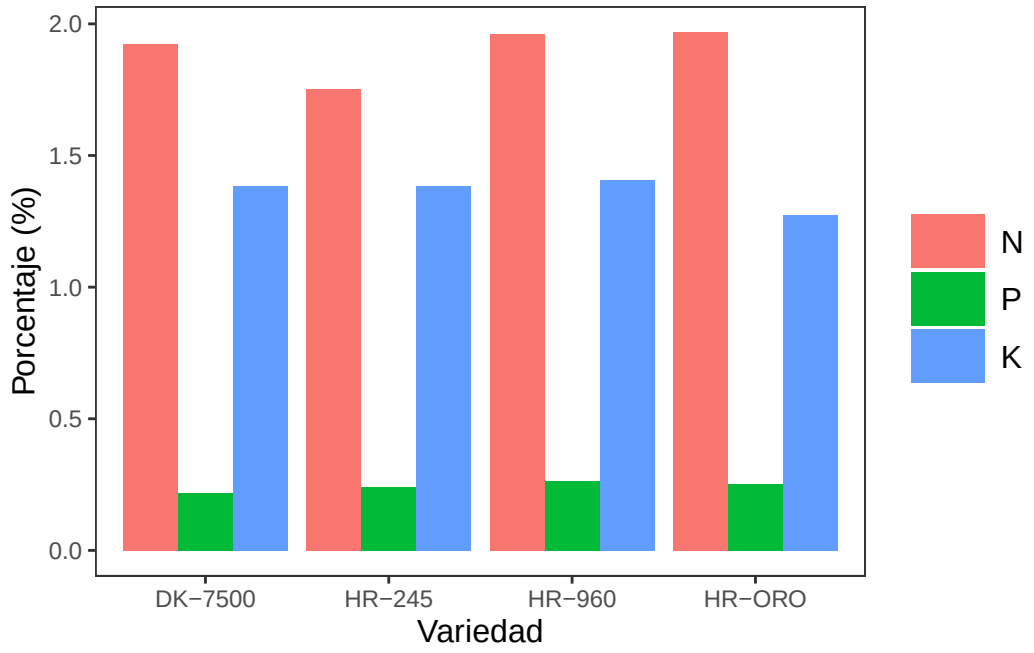


Figura 3.3: Diagrama de Barras Agrupadas

3.1.3.3 Diagrama de barras apiladas

Una variación muy útil de los diagramas de barras son los diagramas de barras apiladas. A diferencia de los diagramas de barras agrupadas en este caso se apilan las barras una sobre otra. Esto es útil cuando las barras suman un 100%. En la Figura 3.4 se muestra la relación Humedad, Materia Seca para cuatro variedades de maíz.

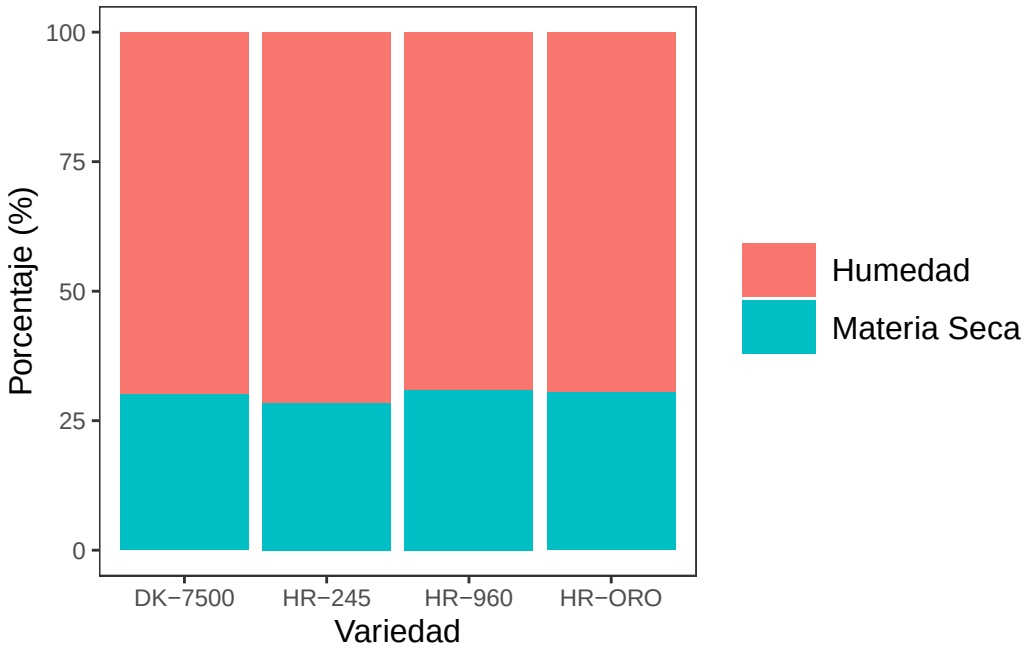


Figura 3.4: Diagrama de Barras Apiladas

3.1.3.4 Histograma

Un histograma es un gráfico que utiliza barras para representar las frecuencias absolutas o las frecuencias relativas de los posibles resultados de una variable cuantitativa y ayuda a entender los valores que se tienen en un conjunto de datos, cuando hablamos de un histograma o histograma de frecuencias nos referimos a las frecuencias absolutas. Con un histograma se puede conocer la forma, el centro y la variabilidad de la distribución. La altura de las barras corresponde a las frecuencias absolutas.

Es importante tener en cuenta que un histograma no debe ser usado para hacer inferencias ya que nos dan una rápida visión de la distribución de los datos y solo sugieren información sobre ciertas características. En la Figura 3.5 se aprecia el histograma de la altura de los girasoles. En la figura Figura 3.6 se esquematiza la relación entre la tabla de frecuencias y el histograma de frecuencias absolutas.

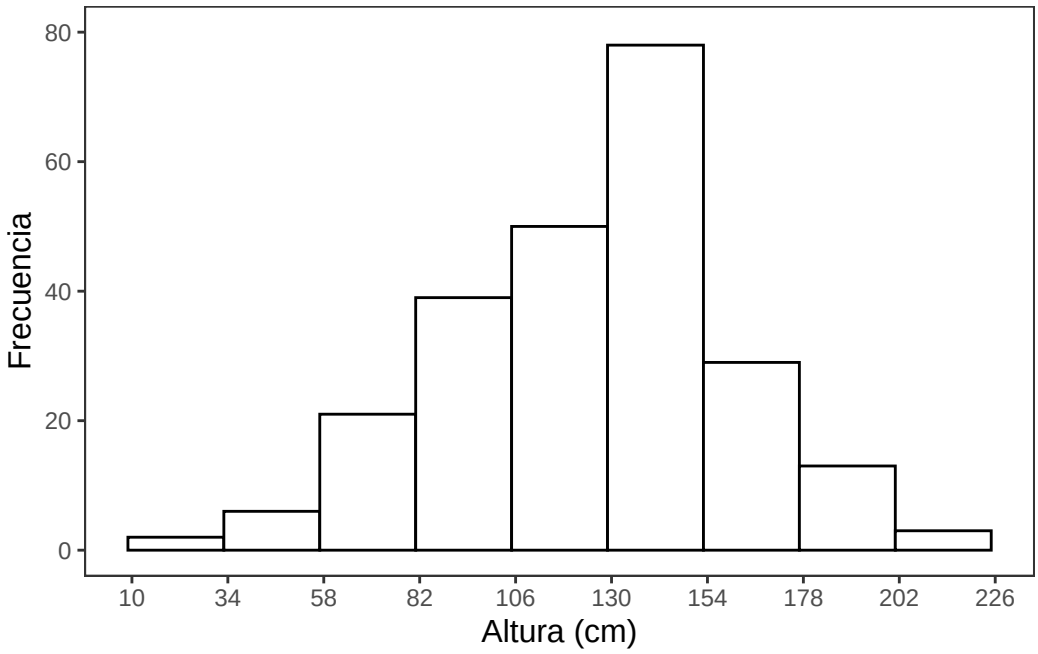
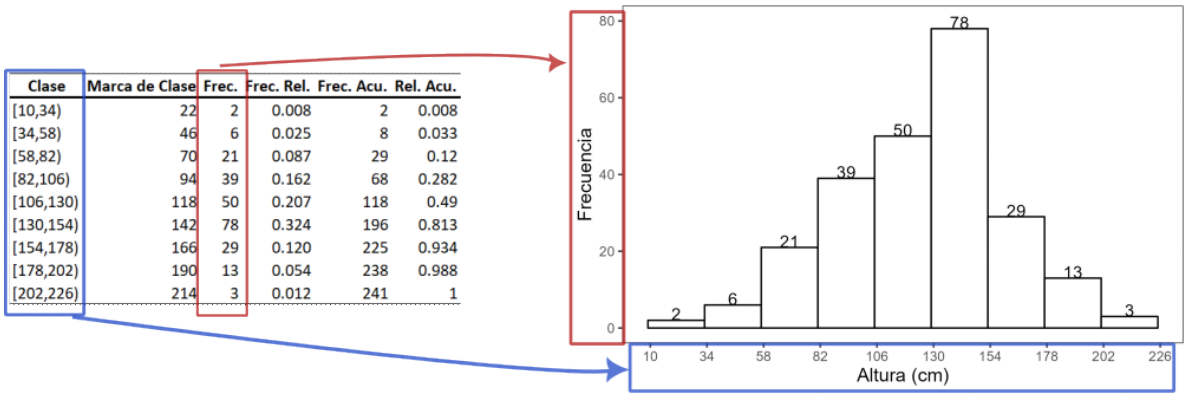


Figura 3.5: Histograma de la altura de girasoles



Clase	Marca de Clase	Frec.	Frec. Rel.	Frec. Acu.	Rel. Acu.
[10,34)	22	2	0.008	2	0.008
[34,58)	46	6	0.025	8	0.033
[58,82)	70	21	0.087	29	0.12
[82,106)	94	39	0.162	68	0.282
[106,130)	118	50	0.207	118	0.49
[130,154)	142	78	0.324	196	0.813
[154,178)	166	29	0.120	225	0.934
[178,202)	190	13	0.054	238	0.988
[202,226)	214	3	0.012	241	1

Figura 3.6: Relación entre la Tabla 3.2 y la Figura 3.5

En algunas ocasiones es de interés graficar las frecuencias relativas. El histograma de las frecuencias relativas tiene la misma forma que el histograma de las frecuencias absolutas, pero las alturas ahora corresponden a las frecuencias relativas. En la Figura 3.7 se muestra el histograma de frecuencias relativas para las alturas de los girasoles. Se puede apreciar que este histograma tiene la misma forma que el histograma de la Figura 3.5.

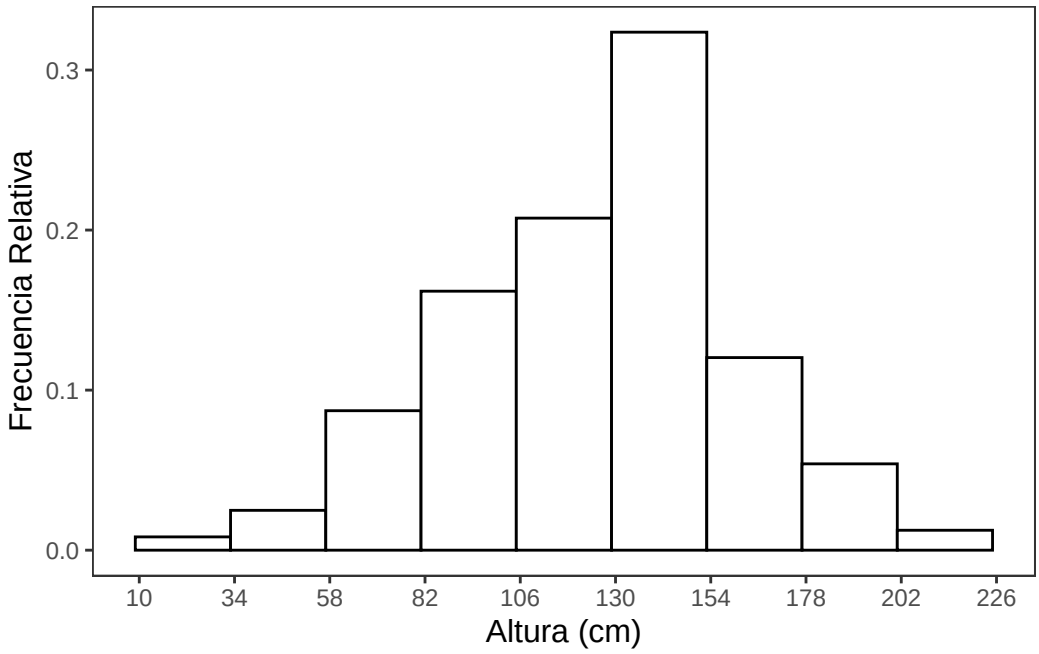
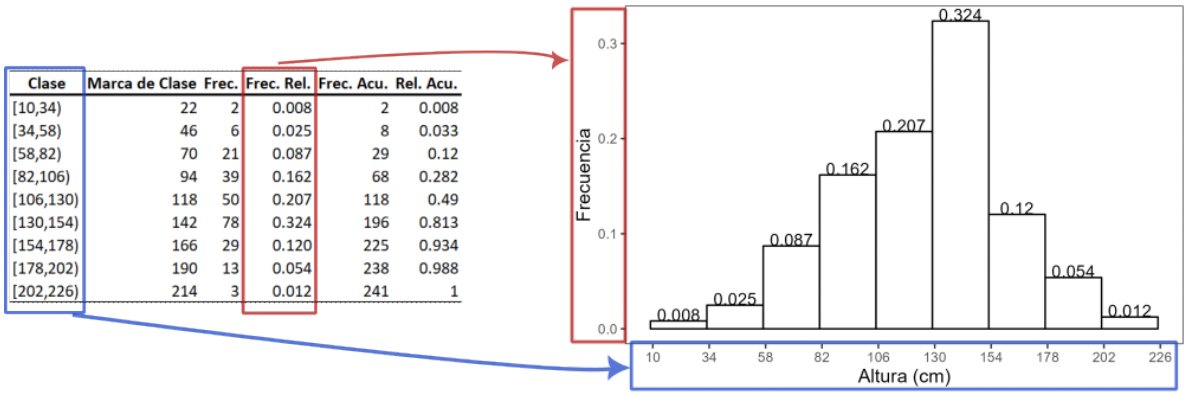


Figura 3.7: Histograma de frecuencias relativas para las alturas de los girasoles



Clase	Marca de Clase	Frec.	Frec. Rel.	Frec. Acu.	Rel. Acu.
[10,34)	22	2	0.008	2	0.008
[34,58)	46	6	0.025	8	0.033
[58,82)	70	21	0.087	29	0.12
[82,106)	94	39	0.162	68	0.282
[106,130)	118	50	0.207	118	0.49
[130,154)	142	78	0.324	196	0.813
[154,178)	166	29	0.120	225	0.934
[178,202)	190	13	0.054	238	0.988
[202,226)	214	3	0.012	241	1

Figura 3.8: Relación entre la Tabla 3.2 y la Figura 3.7

3.1.3.5 Gráficos de densidad

Un gráfico de densidad es una representación de la distribución de una variable numérica. Es una versión suavizada del histograma. Los gráficos de densidad son usados para estudiar la distribución de una variable. Los picos de un gráfico de densidad muestran donde los se concentran los valores en el rango de la variable. Una ventaja que tienen los gráficos de densidad

sobre los histogramas es que son mejores para determinar la forma de la distribución de los datos porque no se ven afectados por el número de barras usadas.

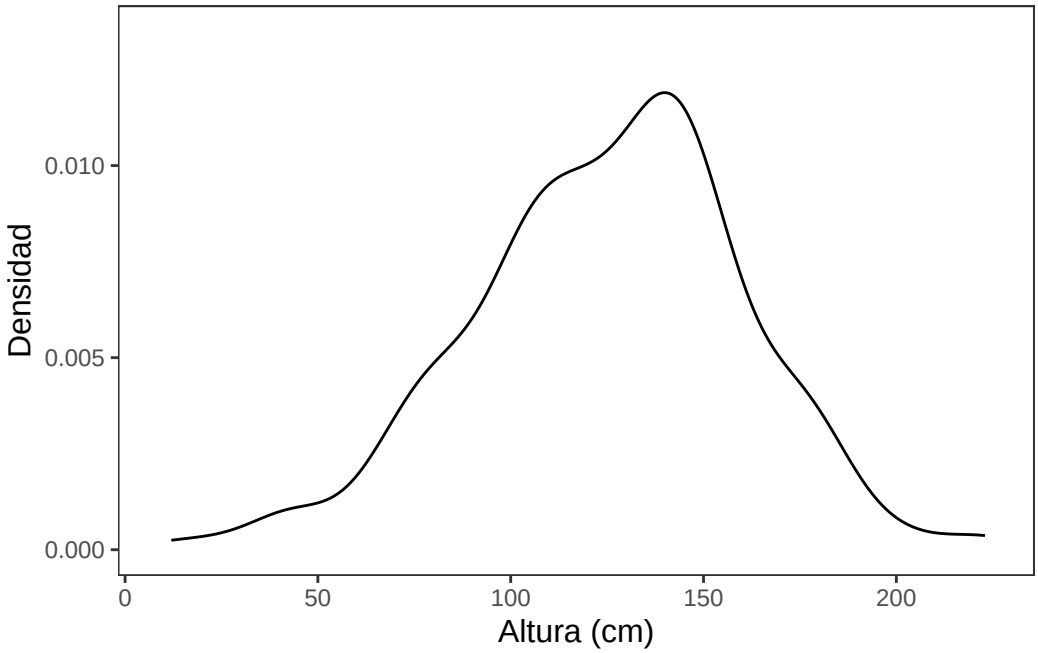


Figura 3.9: Gráfico de Densidad para la altura de los girasoles

3.1.3.6 Diagramas de caja

Un diagrama de caja resume los valores numéricos de una variable categórica, pero no se limitan solamente a la comparación de los valores sino que además ofrece una idea del rango de valores que puede tomar cada categoría dentro de la variable. En la Figura 3.10 se muestra el diagrama de caja para la altura de plantas de caña de azúcar que fueron sometidas a 3 tratamientos diferentes de riego y fertilización, estas mediciones corresponden a los 149 días después de la siembra.

Para la interpretación de un diagrama de caja principalmente hay que fijarse en la línea central de la caja. Este valor corresponde a la mediana, en secciones posteriores desarrollaremos este concepto, por el momento podemos decir que el 50% de los valores son menores a esta cantidad. Cuando la distribución de los datos es simétrica, la mediana se ubicará justo en el centro de la caja. Cuando la distribución de los datos es sesgada, la mediana estará posicionada más cerca de la parte superior o inferior de la caja. Los puntos que se ven sobre o bajo la caja son considerados valores atípicos. En secciones subsiguientes se desarrollarán en detalle los conceptos de sesgo, simetría y valores atípicos.

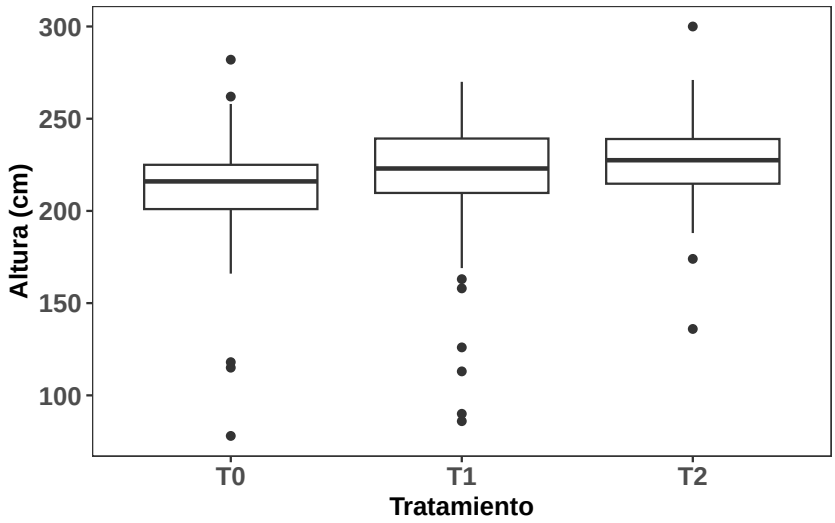


Figura 3.10: Diagrama de Caja para la Altura de plantas de caña de azúcar 149 días después del día de siembra

3.1.3.7 Gráficos de líneas

Estos gráficos son útiles para mostrar la tendencia de una variable cuantitativa en el tiempo. En la Figura 3.11 se muestra la evolución de la altura de las plantas de caña separadas por tratamiento.

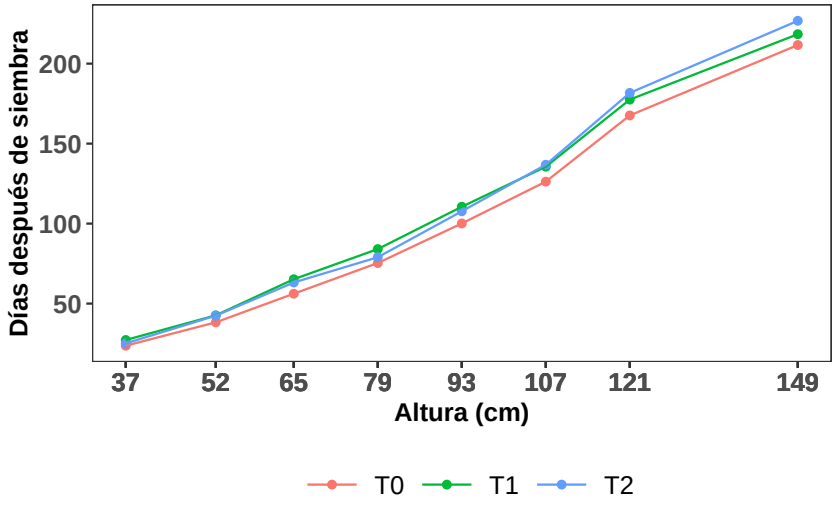


Figura 3.11: Evolución de la altura de las plantas de caña de azúcar

3.2 Medidas numéricas de resumen

Una medida numérica de resumen es un número único que se calcula a partir de una muestra que transmite una característica específica de toda la muestra. Las medidas numéricas sirven para medir la tendencia central, la dispersión o la posición de los datos.

3.2.1 Medidas de tendencia central

Las medidas de tendencia central indican alrededor de qué valor se centra, agrupa o aglutina la mayoría de datos. Existen algunas medidas de tendencia central. Como se mencionó antes, en todas se combina la información de una muestra en un único número y cada medida tiene ventajas y desventajas en su uso.

3.2.1.1 Media

Definición 3.1. La media o media aritmética es simplemente el promedio de los valores observados. Es decir, para los n valores de una muestra x_1, $x_2, \ldots, x_n$ la media muestral $\bar{x}$ se define como la suma de todos los valores dividida para el número de valores:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.5}$$

La media poblacional es denotada por la letra griega μ, en este caso se divide para N que representa el tamaño poblacional.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.6}$$

> **!** Importante
>
> **Propiedades de la media**
>
> 1. Si a todos los valores $\{x_1, x_2, \ldots, x_n\}$ de una variable x con media $\bar{x}$ se les suma una constante k, la nueva media es $\bar{x} + k$
> 2. Si todos los valores $\{x_1, x_2, \ldots, x_n\}$ de una variable x con media $\bar{x}$ se los multiplica por una constante k, la nueva media es $\bar{x}k$
> 3. Si todos los valores $\{x_1, x_2, \ldots, x_n\}$ de una variable x con media $\bar{x}$ se los divide por una constante k diferente de 0, la nueva media es $\dfrac{\bar{x}}{k}$

Tip

Ventajas de la media

1. Es fácil de entender y calcular.
2. Depende de todos los valores.
3. Es susceptible de cálculos aritméticos posteriores.
4. No se ve afectada por las fluctuaciones producto del muestreo

Desventajas de la media

1. Es susceptible a valores extremos.
2. Al depender de todos los valores de un conjunto de datos, si existe uno o varios valores perdidos no es posible calcularla.
3. No se usa para variables cualitativas.

3.2.1.2 Media geométrica

Definición 3.2. La media geométrica MG de un conjunto de n observaciones $\{x_1, x_2, \dots, x_n\}$ se define como la raíz enésima del producto de todas las observaciones. Entonces la media geométrica MG viene dada por:

$$MG = \sqrt[n]{x_1\, x_2\, \dots\, x_n} = \sqrt[n]{\Pi_{i=1}^{n} x_i} \tag{3.7}$$

Importante

Propiedades de la media geométrica

1. La media geométrica es menor que la media aritmética.
2. El producto de los valores de un conjunto de datos se mantiene si cada término es reemplazado por la media geométrica.

$$x_1 \times x_2 \times \dots \times x_n = \overbrace{GM_x \times \cdots \times GM_x}^{n\ veces} \tag{3.8}$$

3. La media geométrica del cociente de las observaciones correspondientes en dos series es igual a los cocientes de sus medias geométricas.
4. La media geométrica del producto de las observaciones correspondientes en dos series es igual a los productos de sus medias geométricas.

Tip

¿Cuándo usar la media aritmética o la media geométrica? La media $\bar{x}$ es usada en la mayoría de las situaciones, sin embargo, la media geométrica GM es preferida cuando los cambios en los valores de una distribución ocurren de forma multiplicativa. Es decir que se la puede usar para promediar datos que siguen progresiones geométricas, por ejemplo razones, interés compuesto, tasas de depreciación, crecimientos de bacterias en microbiología. La media geométrica es muy útil para construir índices

Los **valores atípicos** son valores que están muy lejos del resto de los datos. Cuando un conjunto de datos tiene valores atípicos la media podría no ser la mejor medida de tendencia central para describirlo. Los valores atípicos tienen mucha influencia sobre la media y tienden a arrastrar la media en dirección hacia ellos.

3.2.1.3 Mediana

Definición 3.3. La mediana muestral denotada con $\tilde{x}$ de n observaciones $x_1, x_2, \ldots, x_n$ es el valor central cuando las observaciones están ordenadas en forma ascendente. Cuando el número de observaciones es impar, la mediana es el único valor que está en la mitad de los datos. Mientras que cuando el número de observaciones es par, la mediana es el promedio de los dos valores centrales. En términos matemáticos, supongamos que $\{X\}$ denota al conjunto de datos X ordenado de forma ascendente y X_i representa el i-ésimo elemento del conjunto ordenado. La mediana $\tilde{x}$ es:

$$\tilde{x} = \begin{cases} \{X\}_{\frac{n+1}{2}} & ; n \quad \text{impar} \\ \dfrac{\{X\}_{\frac{n}{2}} + \{X\}_{\frac{n}{2}+1}}{2} & ; n \quad \text{par} \end{cases} \tag{3.9}$$

Tip

Ventajas de usar la mediana

1. Es fácil de calcular y comprender
2. Solo existe una mediana para un conjunto de datos.
3. No se ve afectada por valores extremos
4. Se puede determinar para escalas ordinales, nominales, de razón e intervalo

Desventajas de usar la mediana

1. No toma en cuenta el valor exacto de cada dato y por tanto no usa toda la información disponible.

2. Si se agrupan los valores de dos grupos, la mediana de cada grupo no puede ser expresada en términos del grupo agrupado. Dicho de otra forma si se calculan las medianas de subconjuntos de un conjunto de datos, estas medianas no pueden ser combinadas para calcular la mediana de todo el conjunto.

3.2.1.4 Moda

Definición 3.4. La moda denotada por Mo es el valor que tiene la mayor frecuencia absoluta. Hay conjuntos de datos que no tienen moda, también existen conjuntos de datos con más de una moda. A un conjunto de datos con 2 modas se los llama bimodal, cuando existen más de 2 modas un conjunto recibe el nombre de multimodal.

Tip

Ventajas de usar la moda

1. Fácil de determinar.
2. Se puede usar para datos con escala nominal u ordinal.

Desventajas de usar la moda

1. Debido a que la moda no está definida algebraicamente no se acostumbra a usarla en análisis estadístico.

3.2.2 ¿Cómo escoger la medida de tendencia central adecuada?

En la mayoría de situaciones se prefiere la media como la medida de tendencia central que se reporta, sin embargo, en algunas situaciones se recomienda usar la mediana por ejemplo:

1. Hay algunos valores extremos en la distribución.
2. Algunas observaciones tienen valores no determinados.
3. Los datos se miden en una escala ordinal

Cuando los datos se encuentran en una escala nominal, se prefiere a la moda.

3.2.3 Simetría y sesgo

La forma de una distribución unimodal, puede ser simétrica o sesgada. Una distribución es **simétrica** si hacia la derecha y la izquierda de un valor central tiene la misma forma como se

observa en el histograma de la Figura 3.14, además en esta figura se muestra un diagrama de caja y un gráfico de densidad para una distribución simétrica.

Una distribución es **sesgada a la derecha** si la cola derecha es más larga que la izquierda, en la Figura 3.12 se muestran el histograma, el diagrama de caja y el gráfico de densidad para una distribución sesgada a la derecha. Finalmente, cuando la cola izquierda es más larga que la derecha la distribución es **sesgada a la izquierda** en la Figura 3.13 se aprecia una distribución sesgada a la izquierda.

Entre el sesgo y las medidas de tendencia central existen las siguientes relaciones.

1. Cuando la distribución es sesgada a la derecha se cumple $Mo < \tilde{x} < \bar{x}$. En el gráfico de densidad de la Figura 3.12 la línea vertical de color azul representa la moda, la línea vertical de color rojo representa a la mediana y la línea vertical de color naranja representa la media

2. Cuando la distribución es sesgada a la izquierda se cumple que $\bar{x} < \tilde{x} < Mo$. Con los mismos códigos de colores anteriores podemos observar la ubicación de estos valores en la Figura 3.13.
3. Cuando la distribución es simétrica se cumple que $\bar{x} = \tilde{x} = Mo$.

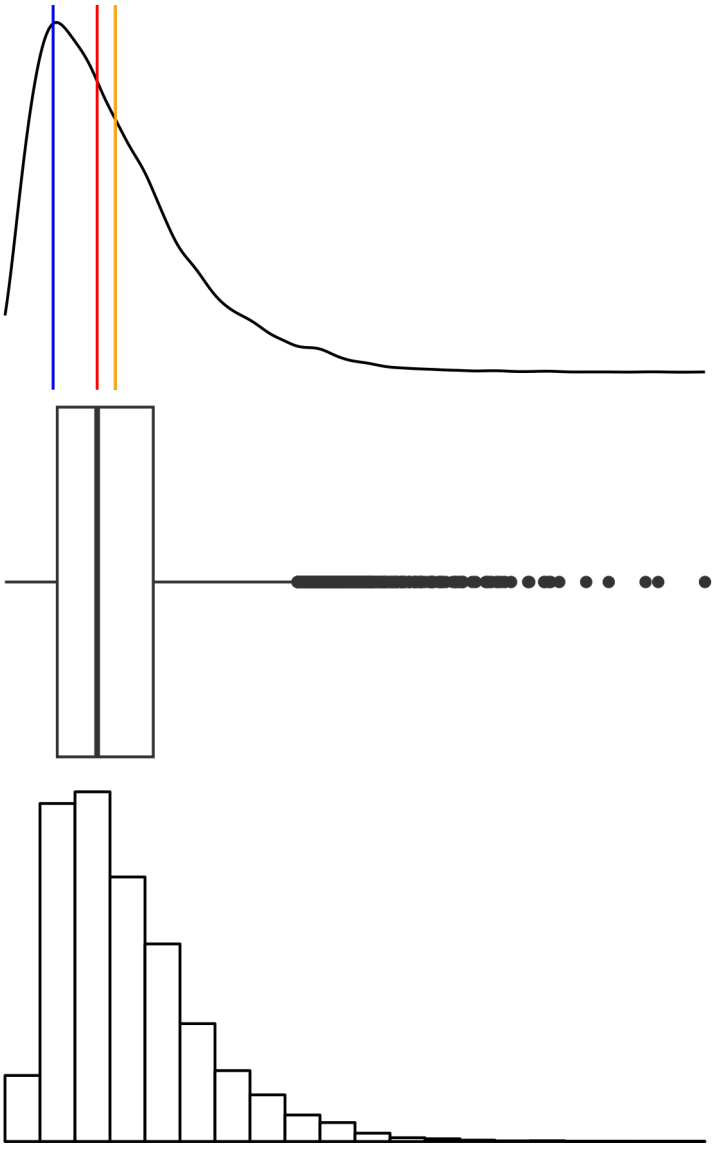


Figura 3.12: Distribución sesgada a a la derecha

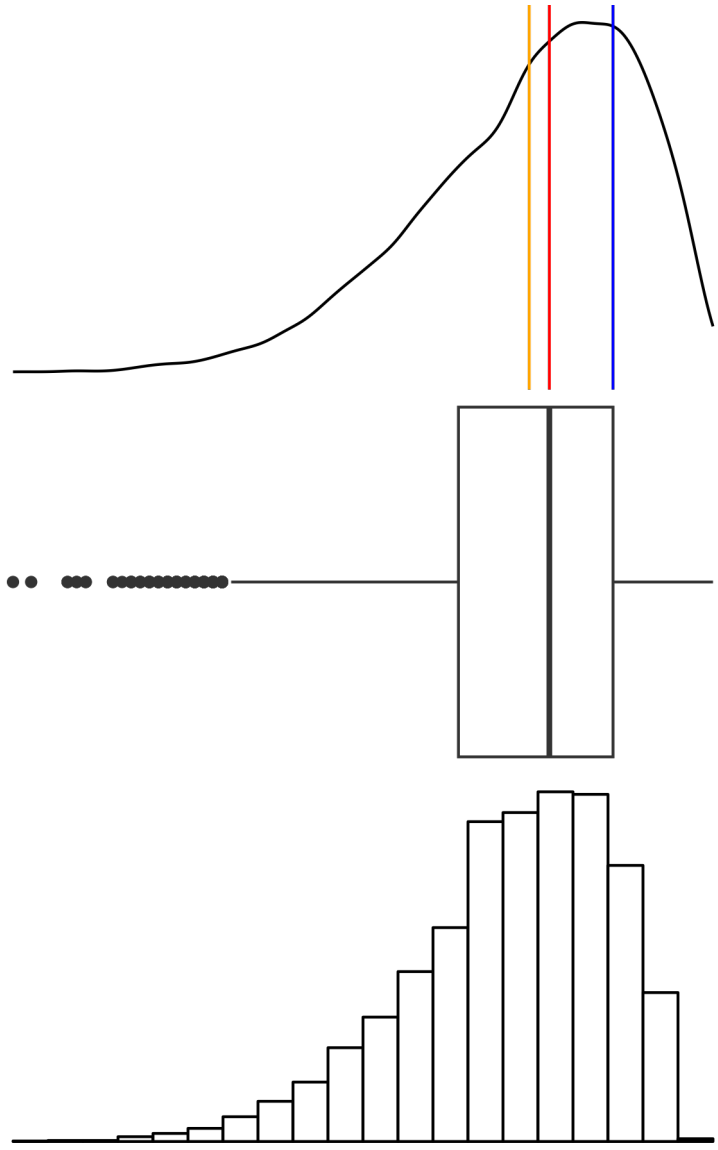


Figura 3.13: Distribución sesgada a a la izquierda

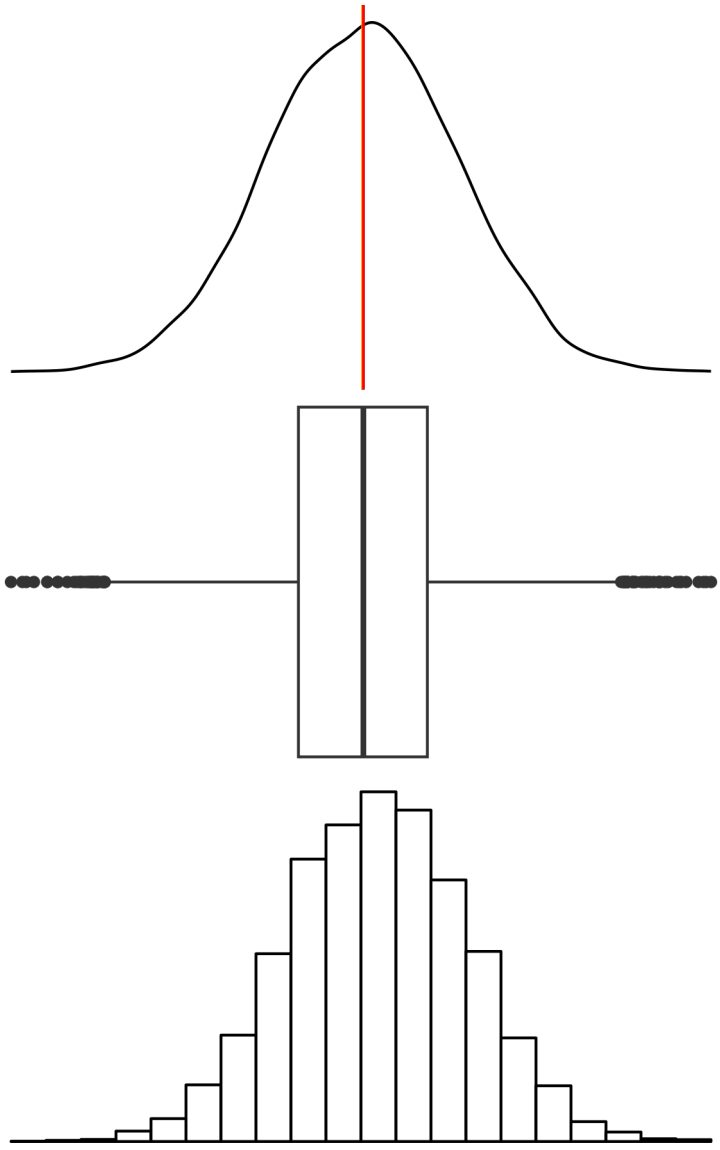


Figura 3.14: Distribución Simétrica

3.2.4 Medidas de dispersión

Supongamos que en una empresa distribuidora de agroquímicos se forman dos equipos de ventas, la empresa ofrece un bono al equipo que tiene el mejor desempeño en las ventas mensuales promedio. En la Tabla 3.3 se muestra el mínimo, la media, la mediana, el máximo y el total para las ventas mensuales de ambos equipos. Se observa que ambos equipos vendieron la misma cantidad total durante el mes por lo que tuvieron la misma media para las ventas mensuales. Es decir, que si se quisiera escoger al mejor equipo por el total de ventas mensuales o por el promedio de ventas mensuales no hay un equipo que haya tenido un mejor desempeño. Por otro lado el equipo 2 tiene un valor mínimo menor que el equipo 1 y en cuanto al máximo el equipo 2 tiene un valor máximo mayor al del equipo 1, esto nos llevaría a pensar que aunque ambos equipos tienen la misma media mensual la distribución de los datos para ambos equipos es diferente.

En la Figura 3.15 se muestra un gráfico de puntos de las ventas por equipo, se ha incluido una línea punteada de color negro en la media de ambos conjuntos (20000.00), con este gráfico se puede notar que las ventas del equipo 2 están más *dispersas* que las ventas del equipo 1. El histograma que se muestra en la Figura 3.16 permite visualizar la distribución de las ventas, y finalmente en el diagrama de caja de la Figura 3.17 se aprecia que el equipo 2 tuvo valores atípicos en sus ventas.

Tabla 3.3: Estadística descriptiva de las ventas por Equipo

Equipo	Mínimo	Media	Mediana	Máximo	Total
Team1	18174.56	20000.00	19946.35	21620.26	400000.00
Team2	16947.96	20000.00	20033.14	22918.67	400000.00

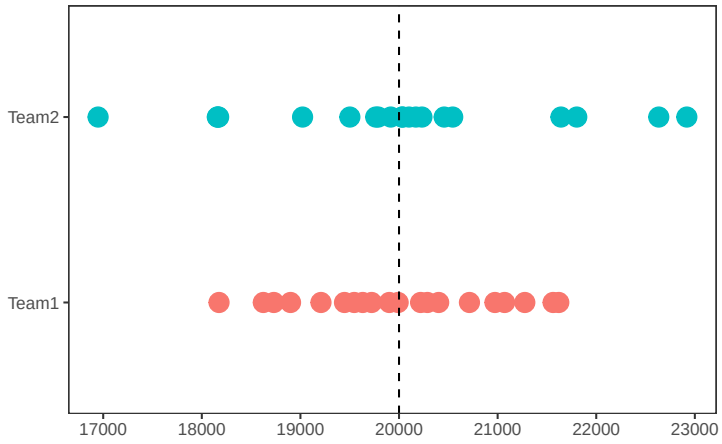


Figura 3.15: Gráfica de puntos de las ventas por equipo

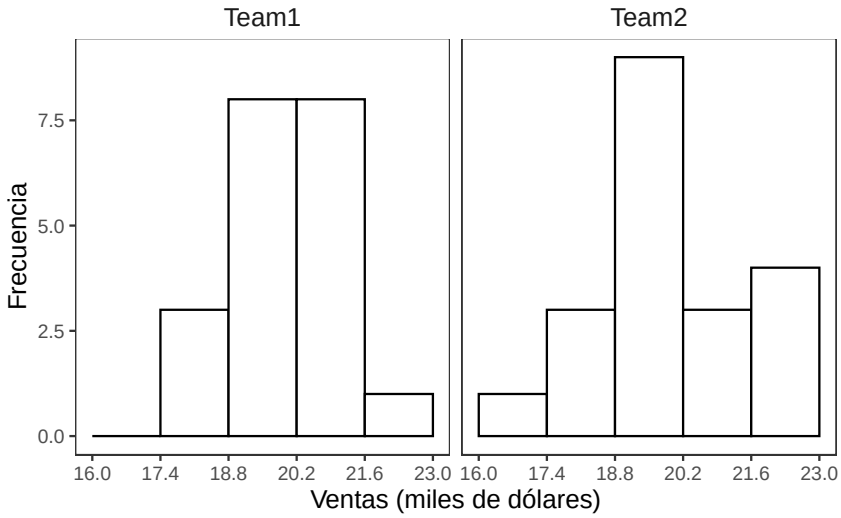


Figura 3.16: Histograma de las ventas por equipo

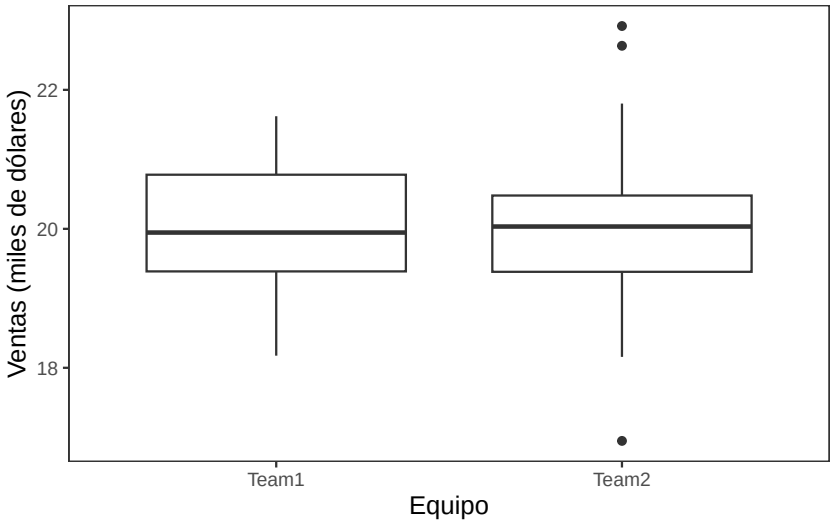


Figura 3.17: Diagrama de caja de las ventas por equipo

Regresando a nuestra pregunta inicial ¿qué equipo tiene un mayor desempeño en sus ventas mensuales? Para contestar esta pregunta vamos a cuantificar la **dispersión** de los datos, esta cuantificación la haremos con las tres medidas de dispersión más usadas.

3.2.4.1 Rango

El rango es la medida de dispersión más fácil de calcular, resulta de la diferencia entre el máximo y el mínimo de un conjunto de datos.

$$Rango = \text{Máximo} - \text{Mínimo}$$

3.2.4.2 Varianza poblacional y muestral

Una forma de definir la dispersión es "la desviación de los datos respecto a la media". La desviación de una observación respecto a la media se la calcula como la diferencia entre la observación y la media, supongamos un conjunto x de n observaciones x_1, x_2, ... , x_n la desviación i-ésima respecto a la media es igual a $x_i - \bar{x}$. La media puede ser interpretada como un punto de balance por lo que en un conjunto cualquiera de n observaciones las desviaciones respecto a la media positivas $(x_i - \bar{x} > 0)$ hacen contrapeso a las desviaciones respecto a la media negativas $(x_i - \bar{x} < 0)$, es decir que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Por esta razón las medidas de dispersión respecto a la media utilizan las desviaciones cuadráticas o el valor absoluto de estas.

Definición 3.5. La **varianza poblacional** que se denota con σ^2 se la calcula como el promedio de las desviaciones cuadráticas respecto a la media:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{3.10}$$

Definición 3.6. La **varianza muestral** que se denota con s^2 se la calcula con la expresión:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{3.11}$$

3.2.4.3 Desviación poblacional y muestral

Definición 3.7. La varianza poblacional o muestral está expresada en unidades cuadráticas, para nuestro ejemplo de las ventas la varianza se expresaría en dólares al cuadrado. Es más fácil interpretar la raíz cuadrada de la varianza. Esta raíz cuadrada se llama **desviación estándar**. La desviación estándar poblacional se la calcula con:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{3.12}$$

La desviación estándar muestral se la calcula con:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{3.13}$$

3.2.5 Medidas de posición

Las medidas de posición no central permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre las medidas de posición no central más importantes están los cuantiles. El término cuantil fue usado por primera vez por Kendall en 1940.

El cuantil de orden p de una distribución con $0 < p < 1$ es el valor x_i de la variable X que marca un corte de modo que una proporción p o un porcentaje $100p\%$ de valores de la población es menor o igual que x_i Por ejemplo el cuantil de orden 0.35 dejaría un 35% de valores por debajo de él.

3.2.5.1 Tipos de Cuantiles

- *Cuartiles*: son 3 valores (Q_1, Q_2, Q_3) que dividen a la distribución en 4 partes iguales.
- *Quintiles*: son 4 valores (K_1, K_2, K_3, K_4) que dividen a la distribución en 5 partes iguales.
- *Deciles*: son 9 valores $(D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9)$ que dividen a la distribución en 10 partes iguales.
- *Percentiles*, son 99 valores $(P_1, P_2, \ldots P_{99})$ que dividen a la distribución en 100 partes iguales.

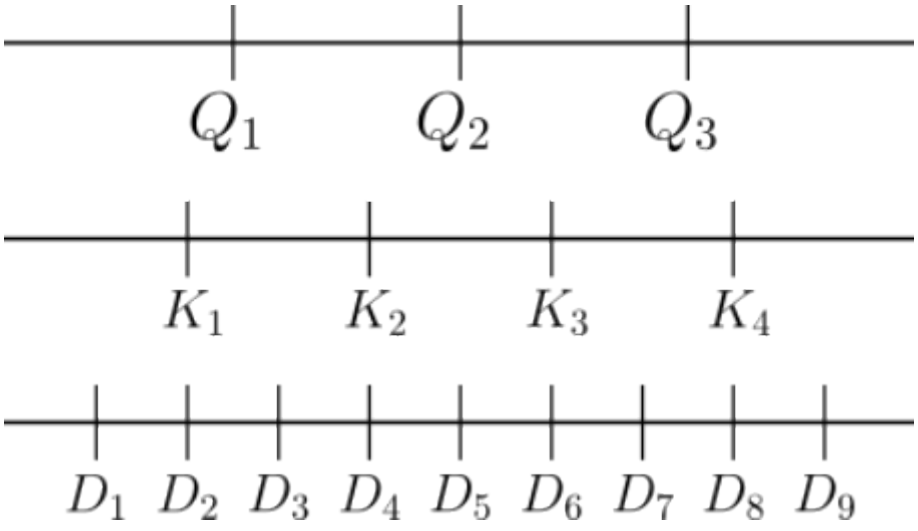


Figura 3.18: Cuantiles

3.2.5.2 Cálculo de cuantiles

Es fácil darse cuenta que existen equivalencias importantes entre los cuantiles, algunos ejemplos de estas equivalencias:

- $D_5 = Q_2 = P_{50}$
- $D_4 = K_2 = P_{40}$
- $D_3 = P_{30}$

Se deduce entonces que no es necesario tener una expresión para cada tipo de cuantiles, basta con conocer una expresión para calcular percentiles. Para esto debemos conocer dos cosas:

1. La posición del percentil en nuestro conjunto de datos.
2. El valor del percentil tomando en cuenta su posición.

Para calcular la posición del percentil i que acumula el $100p\%$ en un conjunto de datos no agrupado X, de tamaño n y ordenado en forma ascendente primero determinamos la posición del percentil con la expresión:

$$\text{Posición} = p(n-1) + 1 \tag{3.14}$$

Para determinar el valor $X_{i.a}$ utilizamos la expresión:

$$X_{i.a} = X_i + 0.a(X_{i+1} - X_i) \tag{3.15}$$

3.2.5.3 Relación entre los cuartiles y el diagrama de caja.

En la Sección 3.1.3.6 se dijo que la línea central de la caja corresponde a la mediana de los datos, ahora que conocemos a los cuartiles es útil saber cuál es la relación entre los cuartiles y los diagramas de caja. En la Figura 3.19 se muestran las partes del diagrama de cajas. El límite inferior y el límite superior se calculan en función del **rango intercuartílico** (IQR por sus siglas en inglés) que es igual a la diferencia entre el tercer y primer cuartil.

$$IQR = Q_3 - Q_1 \tag{3.16}$$

El límite superior (LS) y el límite inferior (LI) se calculan con las siguientes expresiones, respectivamente:

$$LS = Q_3 + 1.5IQR \tag{3.17}$$

$$LI = Q_1 - 1.5IQR \tag{3.18}$$

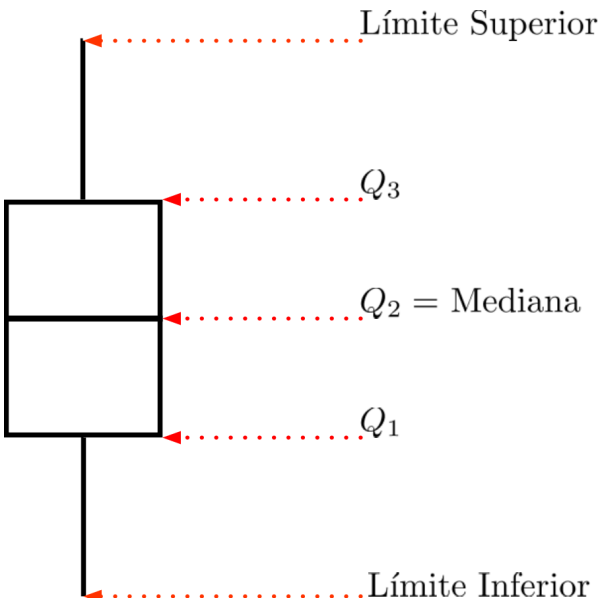


Figura 3.19: Partes de un diagrama de caja

3.3 Primeros pasos en RStudio

En la Sección 1.5, se presentaron algunos conceptos básicos de R utilizando la consola de R. Cuando se trabaja con conjuntos de datos, o cuando se desea guardar y compartir el trabajo realizado es conveniente usar **scripts**, un script es un archivo de texto que contiene los mismos comandos que ingresaría en la línea de comandos de R.

Una vez abierto RStudio, para crear un script hay dos opciones. La primera es en la barra de herramientas principal escoger `File | New File | R Script`, la segunda opción es escribir la combinación de teclas (en Windows) `Ctrl + Shift + N` o `Cmd + Shift + N` (en Mac).

Cuando el script está creado se puede escribir código como se visualiza a la derecha de la Figura 3.20. Para ejecutar el código se puede ubicar en cualquier parte de la línea y presionar la combinación de teclas `Ctrl + Entero` `Cmd + Enter`, otra opción es dar clic en el botón `Run` ubicado en la parte superior derecha del script (encerrada en rectángulo rojo de la Figura 3.21). Si solo es una línea de código no es necesario seleccionarla para poder ejecutarla. Al momento de ejecutar el código, el código se ejecuta en la consola (rectángulo azul de la Figura 3.21) y además se visualiza el objeto en el ambiente (rectángulo negro de la Figura 3.21).

En el proceso de análisis de datos, adicional a los scripts, la mayoría de ocasiones es necesario tener los datos almacenados en algún directorio de la computadora (en algunas ocasiones los datos se leen desde internet). Las tablas y los gráficos que se generen, también pueden ser almacenados en la computadora. Una forma de manejar esto de forma eficiente es trabajando con *directorios de trabajo*, un directorio de trabajo es una ruta dentro de la computadora que

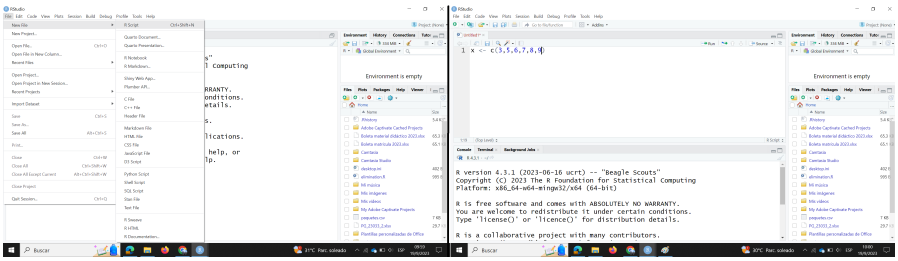


Figura 3.20: Creación de un script

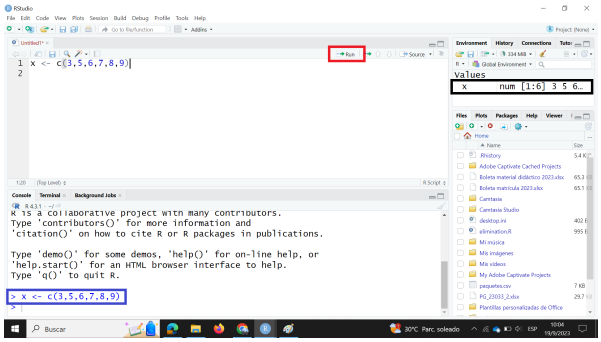


Figura 3.21: Ejecución de comandos desde un script

especifica la ubicación predeterminada de los archivos que leamos en R o que guardemos desde R.

> **Tip**
>
> - Para conocer el directorio de trabajo actual se usa la función `getwd()`.
> - Para fijar el directorio de trabajo se usa la función `setwd(dir = "directorio")`

3.3.1 Creación de proyectos en RStudio

RStudio tiene una forma más eficiente de manejar los directorios de trabajo, utilizando **proyectos**. Un proyecto es simplemente un directorio de trabajo designado con un archivo de extensión `.Rproj`. Al momento de abrir un proyecto, se configura automáticamente el directorio de trabajo como el directorio donde se encuentra el archivo `.Rproj`.

Una buena práctica es crear un proyecto en RStudio por cada investigación, tarea o proyecto de análisis que se desee trabajar. Dentro de la carpeta del proyecto es recomendable tener subcarpetas que permitan almacenar los datos que van a ser leídos y producidos como parte de nuestro análisis. Una sugerencia es tener cuatro carpetas como mínimo:

1. `datos`: en esta carpeta se encuentran los datos que van a ser analizados.
2. `graficos`: para guardar los gráficos producidos.
3. `resultados`: para guardar las tablas resultado del análisis de datos.
4. `scripts`: en esta carpeta se guardan todos los scripts utilizados en el análisis.

En este texto se utilizará un solo proyecto al que llamaremos **r4agro**. En cualquier ubicación de nuestra computadora vamos a crear la carpeta llamada **r4agro** y dentro de esta carpeta las carpetas **datos**, **graficos**, **resultados** y **scripts** como se muestra en la Figura 3.22.

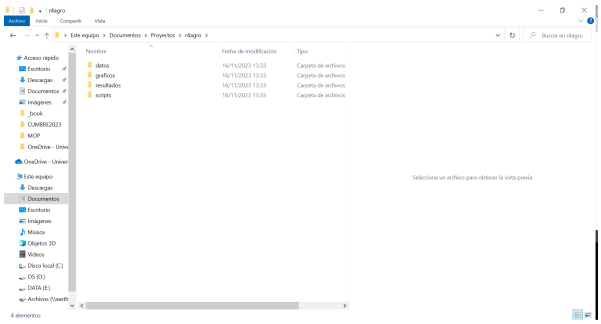


Figura 3.22: Carpeta del proyecto r4agro

Para crear un proyecto, se debe dar clic en `File -> New Project` y aparecerá la ventana mostrada en la Figura 3.23.

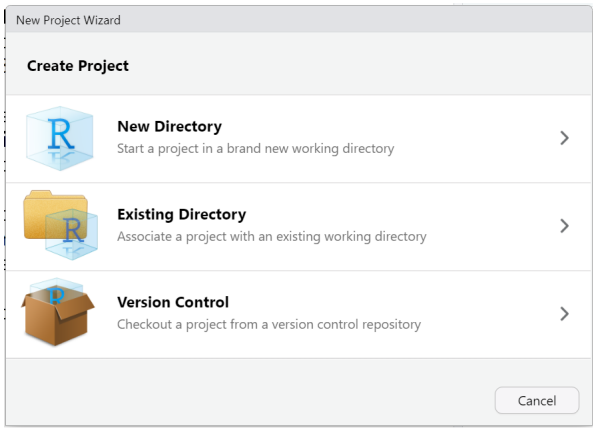


Figura 3.23: Crear un nuevo proyecto

Escogemos la opción de acuerdo a nuestras necesidades:

- **New Directory**: escogemos esta opción si vamos a crear un nuevo directorio desde 0. Cuando se da clic, aparece una nueva ventana como se muestra en la Figura 3.24

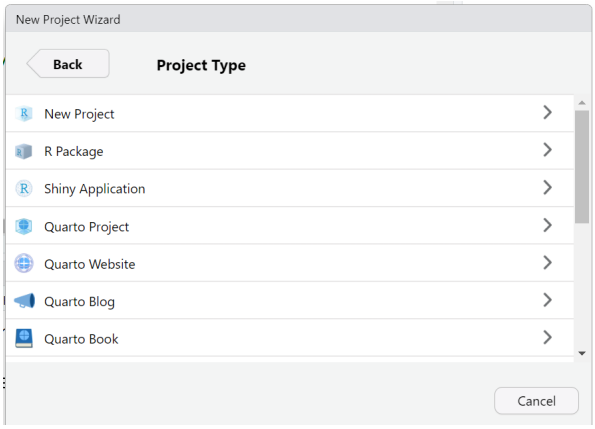


Figura 3.24: Opción New Directory

- **Existing Directory**: escogemos esta opción si ya tenemos creada la carpeta de nuestro proyecto. En este libro vamos a trabajar siempre escogiendo esta opción. Más adelante explicaremos en detalle como crear un proyecto escogiendo esta opción.
- **Version Control**: El control de versiones ayuda a los equipos de desarrollo de software a controlar y gestionar los cambios en el código fuente a lo largo del tiempo. Los programas y plataformas de control de versiones mantienen un registro de las modificaciones hechas al código. Si se comete un error, los desarrolladores pueden comparar versiones anteriores del código para ayudar a corregir el error al tiempo que se minimizan las molestias para todos los miembros del equipo. RStudio trabaja con dos sistemas de código abierto para el control de versiones: **Git** y **Subversion**.

Creación de proyectos (Existing Directory)

- `File -> New Project`
- `Existing Directory`
- Se abre una ventana como la de la Figura 3.25 y damos clic en `Browse`.
- Se abre una ventana en la que buscamos la ubicación de la carpeta de nuestro proyecto y damos clic en `Open` (Figura 3.26).
- Posteriormente, escogemos la opción `Create New Project` (Figura 3.27).
- Finalmente, el proyecto se crea. Identificamos que el proyecto ha sido creado porque ahora aparece el nombre del proyecto en las esquinas superior izquierda y derecha de nuestra ventana de RStudio (Figura 3.28).

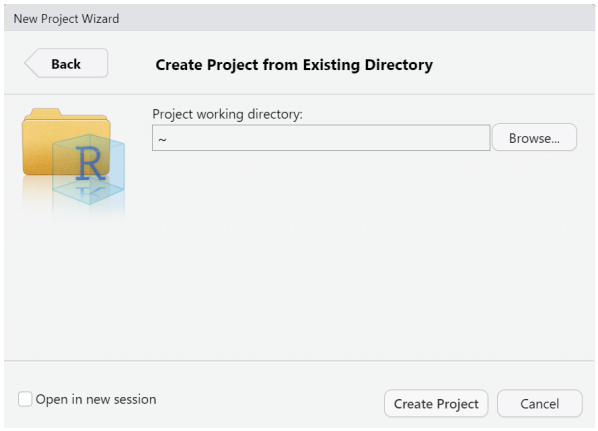


Figura 3.25: Existing Directory

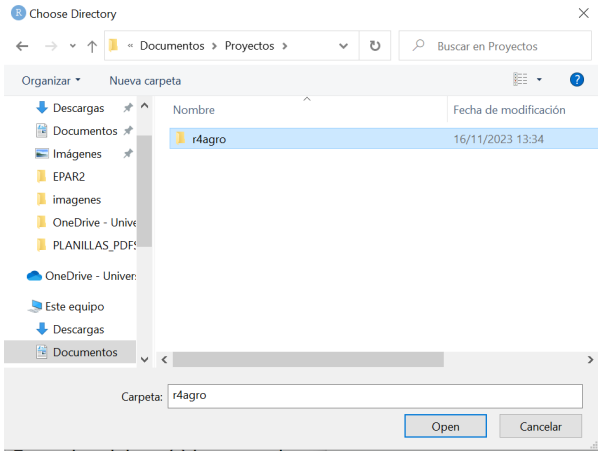


Figura 3.26: Escoger carpeta

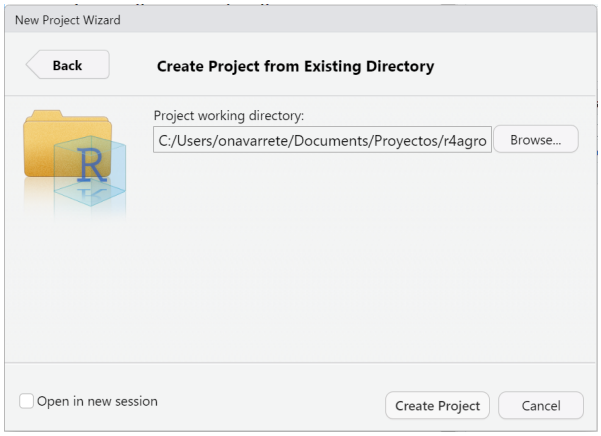


Figura 3.27: Crear proyecto

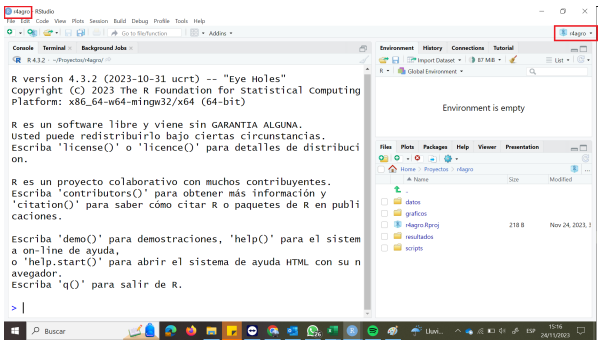


Figura 3.28: Proyecto creado

3.3.2 Creando nuestros primeros gráficos.

En esta sección trabajaremos con el conjunto de datos `alt_almacigo.csv`, este conjunto de datos corresponde a las alturas de 105 plantas de tomate de 5 variedades, 20 días después de la siembra en los almácigos.

El primer paso es guardar nuestros datos en la carpeta `datos` de nuestro proyecto, creado anteriormente, `r4agro`. Posteriormente creamos un script, en esta oportunidad guardaremos nuestro script en la carpeta `scripts` con el nombre `01_visualizacion.R`.

Existen muchas formas de trabajar en un script, sin embargo, se recomienda seguir el siguiente flujo:

> **Tip**
>
> 1. Cargar los paquetes que sean necesarios para el análisis que se va a trabajar.
> 2. Cargar los datos que se van a utilizar en el análisis.
> 3. Trabajar con los datos, esto incluye realizar gráficos, analizar los datos o crear modelos.

En este primer script trabajaremos solamente con los paquetes **dplyr** y **ggplot2** descritos en la Sección 1.4. En las primeras líneas escribiremos las siguientes líneas de código.

```
library(dplyr)
library(ggplot2)
```

Una vez cargados los paquetes procedemos a leer los datos. El archivo con el que vamos a trabajar es un archivo de valores separados por comas (**c**omma ¨**s**eparted **v**alues). Existen muchas funciones que podemos usar para leer este tipo de archivos, sin embargo, vamos a trabajar con la función `read.csv`. Para no cometer errores al momento de la lectura de los datos es importante identificar si nuestros valores efectivamente están separados por coma (,)

y si el archivo utiliza el punto (.) como separador de decimales. El archivo de nuestros datos tiene como separador de los valores el punto y coma, y como separador de decimales la coma.

Importante

1. Un archivo `csv` no necesariamente, tiene sus valores separados por coma. Dependiendo del sistema operativo o la región geográfica, los valores se guardan separados por coma (,) o por punto y coma (;).
2. Si los valores están separados por coma, generalmente el separador de decimales es el punto (.).
3. Cuando los valores están separados por punto y coma, el separador de decimales es la coma.

Para leer los datos escribimos la siguiente línea de código.

```
datos <- read.csv("datos/alt_almacigo.csv",sep = ";", dec = ",")
```

Entendiendo el código

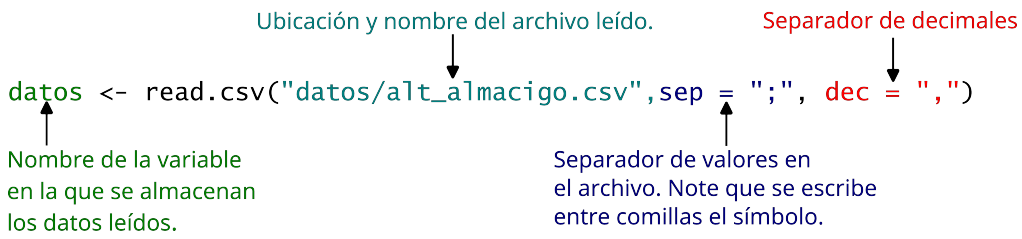


Figura 3.29: Leer datos con la función `read.csv()`

4 Distribuciones de probabilidad

4.1 Definiciones

4.2 Distribución Normal

4.3 Distribución t de Student

5 Muestreo

5.1 Técnicas de muestreo

6 Estadística Inferencial

6.1 Estimación puntual y por intervalos

6.2 Intervalos de confianza

6.3 Pruebas de hipótesis

6.3.1 Para la media (Prueba t)

6.3.1.1 1 sola muestra

6.3.1.2 2 muestras independientes

6.3.1.3 2 muestras pareadas

6.3.2 Para la proporción

6.3.3 Para la varianza

6.4 Pruebas de hipótesis no paramétricas

7 Correlación y Regresión

7.1 Gráficos de dispersión

7.2 Análisis de correlación

7.3 Regresión

7.3.1 Regresión lineal simple

7.3.2 Regresión lineal múltiple

7.3.3 Regresión logística

8 Diseño de experimentos

8.1 Principios básicos

8.2 Prueba ANOVA

8.3 Diseño completo al azar

8.4 Diseño de bloques completos al azar

8.5 Diseño de cuadro latino

8.6 Pruebas de comparaciones múltiples paramétricas

8.7 Pruebas no paramétricas equivalentes al análisis de varianza

8.8 Pruebas de comparaciones múltiples no paramétricas

8.9 Diseño factorial

Referencias

Agresti, A., Franklin, C., y Klingenberg, B. (2023). *Statistics. The Art and Science of Learning from Data* (5a. ed.). Pearson.

Agwu, N., y Bialas, P. (2018). *Using R/R Studio In An Introductory Statistics Course*. *10*(3).

Cleff, T. (2013). *Exploratory Data Analysis in Business and Economics* (1a. ed.). Springer. https://link.springer.com/book/10.1007/978-3-319-01517-0

Data Carpentry. (2020). Packages and libraries. En *Introduction to R*. Harvard Chan Bioinformatics Core (HBC). https://hbctraining.github.io/Intro-to-R-flipped/lessons/04_introR_packages.html

Ismay, A. Y., Chester and Kim. (2022). *Statistical Inference via Data Science*. https://moderndive.com/index.html

Kibuuka, E. (2009). *Formulating a country's agricultural statistics strategy: The South African experience*. 10.

Navarrete, O., y Chávez, M. (2019). *Estadística para Contadores y Auditores con R*. Abya Yala.

Tucker, M. C., Shaw, S. T., Son, J. Y., y Stigler, J. W. (2022). Teaching Statistics and Data Analysis with R. *Journal of Statistics and Data Science Education*, 1-15. https://doi.org/10.1080/26939169.2022.2089410