Estadística con R (I)

Descripción general

Este material reúne **prácticas de iniciación al uso de R y RStudio**. Está orientado al **aprendizaje práctico del manejo del programa**, desde la instalación y exploración del entorno hasta la importación, manipulación y representación básica de datos. A través de ejercicios guiados y conjuntos de datos reales, el alumnado adquiere destrezas para trabajar con tipos y estructuras de datos, aplicar funciones esenciales y generar resultados reproducibles. El contenido incluye **enunciados, prácticas resueltas, archivos de datos y recursos complementarios** que facilitan el aprendizaje autónomo y la comprensión aplicada de la estadística con R.

Objetivo general	Descripción y competencias asociadas
1. Introducir al alumnado en el uso del entorno RStudio.	Familiarizarse con la interfaz, paneles de trabajo y herramientas básicas para ejecutar código y gestionar proyectos.
2. Promover el aprendizaje autónomo en el manejo de R.	Desarrollar destrezas prácticas para escribir, ejecutar y depurar scripts, entendiendo la lógica de programación de R.
3. Comprender y aplicar los tipos de datos básicos.	Identificar los tipos de objetos en R (numéricos, caracteres, lógicos, factores) y realizar operaciones elementales con ellos.
4. Conocer las principales estructuras de datos.	Manejar vectores, matrices, listas y data frames; acceder a elementos y modificar estructuras.
5. Aprender a instalar y utilizar paquetes.	Incorporar librerías externas y emplear funciones específicas de análisis y visualización de datos.
6. Dominar los procesos de importación y exportación de datos.	Leer y guardar datos desde distintos formatos (CSV, Excel, RData) y comprender las diferencias entre ellos.
7. Desarrollar habilidades básicas de análisis descriptivo.	Calcular medidas de posición, dispersión y forma; representar resultados mediante gráficos con R base y *ggplot2*.
8. Fomentar la reproducibilidad en el análisis de datos.	Comprender la importancia de los scripts bien documentados, la organización de proyectos y la trazabilidad del trabajo estadístico.

Autoras: Nerea González García · Ana B. Nieto Librero
Departamento de Estadística, Universidad de Salamanca
Curso académico: 2023–2024
Tipo de recurso: Material docente / prácticas resueltas
Repositorio institucional: GREDOS – Universidad de Salamanca

ESTRUCTURA DEL MANUAL

Práctica 1. Introducción a R y RStudio

- Familiarización con el entorno RStudio.
- Tipos y estructuras de datos.
- Instalación de paquetes.
- Importación y exportación de datos.

Práctica 2. Ordenación de datos, tablas de frecuencias y representaciones gráficas

- Resumen y ordenación de datos.
- Tablas de frecuencias.
- Representaciones gráficas básicas.
- Parámetros gráficos y función par().

Práctica 3. Análisis descriptivo: medidas de posición, dispersión y forma

- Cálculo e interpretación de medidas de síntesis.
- Análisis exploratorio de datos.
- Gráficos tipo *Box-Plot*.
- Uso de dplyr y psych para análisis descriptivo.

Práctica 4. Regresión lineal y no lineal

- Análisis de correlación.
- Modelos lineales simples y múltiples.
- Poder explicativo y Poder predictivo.
- Modelos no lineales (logarítmico, potencial, exponencial, parabólico).

Práctica 5. Estadística descriptiva y regresión (caso aplicado: cáncer de pulmón)

- Análisis exploratorio de datos biomédicos.
- Gráficos de dispersión y Box-Plots.
- Aplicación de regresión a datos reales.

Práctica 6. Distribuciones de probabilidad y simulación de datos

- Funciones (dxxx, pxxx, qxxx, rxxx).
- Distribuciones Binomial, Normal, Ji-Cuadrado, t-Student y F-Snedecor.
- Ejercicios de simulación y representación gráfica.

Práctica 7. Intervalos de confianza

- Cálculo e interpretación de intervalos para la media.
- Comparación de medias en dos poblaciones independientes.
- Intervalos de confianza gráficos con plotmeans().

Material complementario

- Práctica de autoevaluación y conjunto de datos.
- Cuestionario de autoevaluación.
- Conjuntos de datos utilizados.

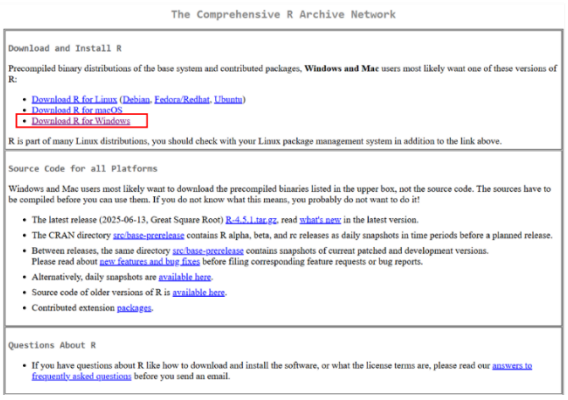
1.Introducción a R. Manejo de datos

OBJETIVOS
Descarga e instalación de R y RStudio
Familiarizarse con el entorno de RStudio
Conocer:
1. Tipos de datos básicos y realizar operaciones con ellos.
2. Estructuras de datos y aprender su manejo.
Instalación de paquetes.
Importar bases de datos desde ficheros externos.
Exportar bases de datos a ficheros externos.

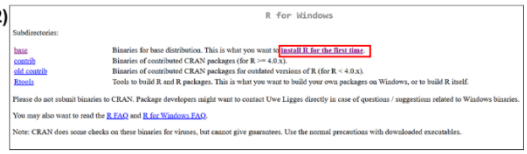
1. INSTRUCCIONES DE INSTALACIÓN DE R

Acceda a la página oficial de R https://cran.r-project.org/ y seleccione el correspondiente sistema operativo (Window, Linux o Mac). Por ejemplo, para Windows:

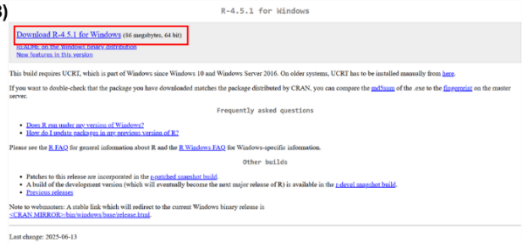
1)



2)

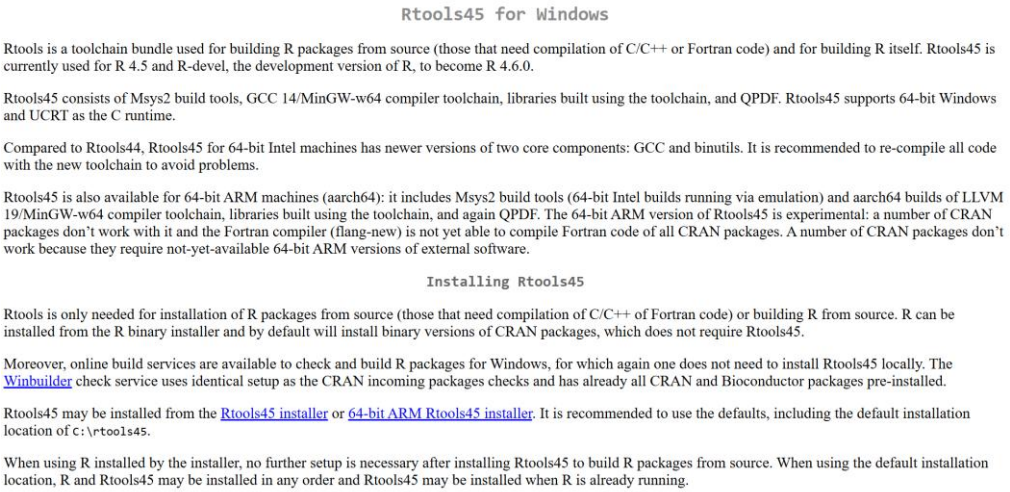


3)



Una vez descargado el ejecutable, siga las instrucciones indicadas en el Asistente de Instalación.

Descargue también RTools:



Rtools45 for Windows

Rtools is a toolchain bundle used for building R packages from source (those that need compilation of C/C++ or Fortran code) and for building R itself. Rtools45 is currently used for R 4.5 and R-devel, the development version of R, to become R 4.6.0.

Rtools45 consists of Msys2 build tools, GCC 14/MinGW-w64 compiler toolchain, libraries built using the toolchain, and QPDF. Rtools45 supports 64-bit Windows and UCRT as the C runtime.

Compared to Rtools44, Rtools45 for 64-bit Intel machines has newer versions of two core components: GCC and binutils. It is recommended to re-compile all code with the new toolchain to avoid problems.

Rtools45 is also available for 64-bit ARM machines (aarch64): it includes Msys2 build tools (64-bit Intel builds running via emulation) and aarch64 builds of LLVM 19/MinGW-w64 compiler toolchain, libraries built using the toolchain, and again QPDF. The 64-bit ARM version of Rtools45 is experimental: a number of CRAN packages don't work with it and the Fortran compiler (flang-new) is not yet able to compile Fortran code of all CRAN packages. A number of CRAN packages don't work because they require not-yet-available 64-bit ARM versions of external software.

Installing Rtools45

Rtools is only needed for installation of R packages from source (those that need compilation of C/C++ of Fortran code) or building R from source. R can be installed from the R binary installer and by default will install binary versions of CRAN packages, which does not require Rtools45.

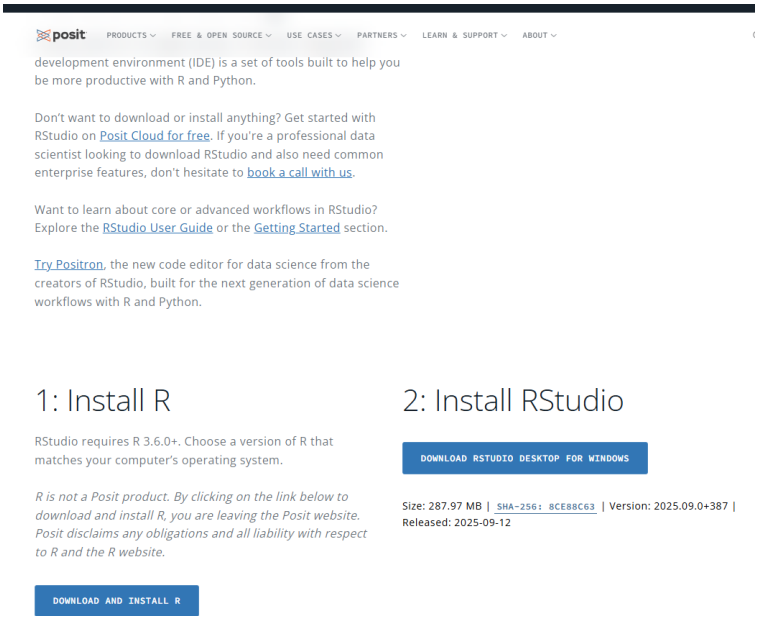
Moreover, online build services are available to check and build R packages for Windows, for which again one does not need to install Rtools45 locally. The Winbuilder check service uses identical setup as the CRAN incoming packages checks and has already all CRAN and Bioconductor packages pre-installed.

Rtools45 may be installed from the Rtools45 installer or 64-bit ARM Rtools45 installer. It is recommended to use the defaults, including the default installation location of `C:\rtools45`.

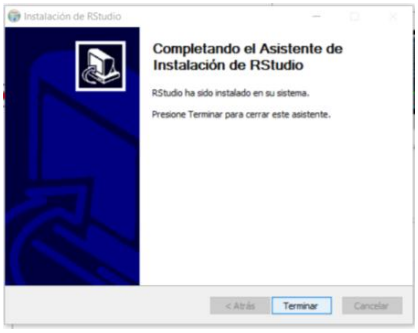
When using R installed by the installer, no further setup is necessary after installing Rtools45 to build R packages from source. When using the default installation location, R and Rtools45 may be installed in any order and Rtools45 may be installed when R is already running.

2. INSTRUCCIONES DE INSTALACIÓN DE RSTUDIO

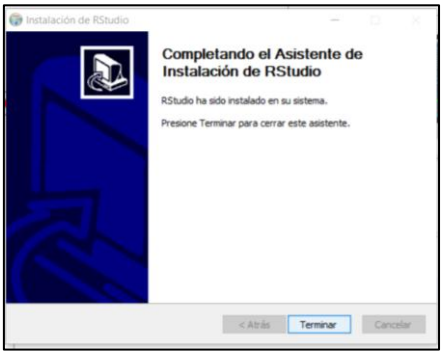
Acceda a la página web: https://www.rstudio.com/ y seleccione la versión de RStudio correspondiente (en este caso RStudio Desktop) y el sistema operativo adecuado.



Abra el ejecutable y siga las instrucciones indicadas en el Asistente de Instalación.



…



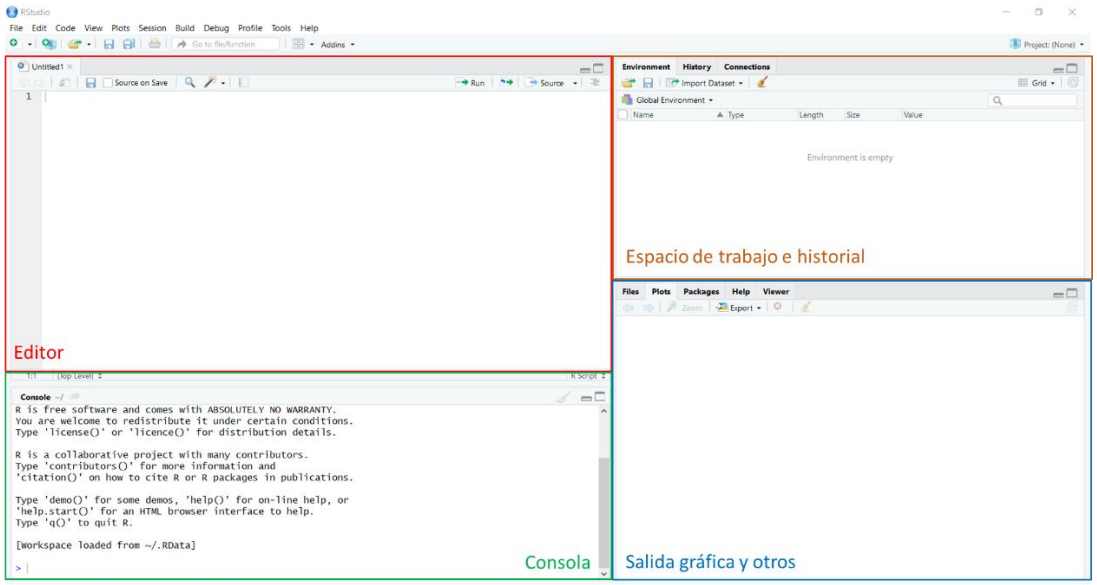
3. RSTUDIO: ENTORNO DE TRABAJO

3.1. Paneles: editor, consola, entorno y salida gráfica.

Al iniciar RStudio, además de la barra de menús, se muestra su interfaz de usuario, compuesta por 4 paneles diferentes:

- **Editor**. Permite escribir líneas de código y guardarlas en un fichero de texto o en un *script* de R. El código no se evalúa automáticamente; debe ser ejecutado en la consola.
- **Consola**. En esta pestaña se evalúan secuencialmente las expresiones de código en R. Los cálculos aquí realizados no se almacenan.
- **Entorno**. Espacio de trabajo e historial. En la pestaña *Environment* quedan recogidos todos los objetos creados en una sesión de R. El código evaluado se almacena en la pestaña *History*.
- **Salida gráfica.** Está formada por cinco pestañas diferentes: *Files* (administrador de ficheros), *Plots* (pestaña de visualización de gráficos), *Packages* (asistente para la instalación de paquetes), *Help*

(documentación de ayuda), *Viewer* (visualizador de documentos HTML)



4. TIPOS BÁSICOS DE DATOS

En R existen distintos tipos de datos, que se engloban de manera general en:

- Numeric. Datos de tipo numérico.
    - Integer. Datos de tipo entero.
    - Double. Datos de tipo real.
    - Complex. Números complejos.
- Logical. Datos de tipo lógico o binario (TRUE, FALSE)

logic = TRUE

- Character. Caracteres o cadenas de caracteres.

char1 = “Hola”

char2 <- “1 - 3”

Para conocer el tipo de dato se utilizará la función *class() o la función is.* seguida del tipo de objeto que quiera analizar. El objetivo de ambas funciones es diferente.

a) ¿Qué tipo de datos es el número 5?

class(5)

b) ¿Qué tipo de dato es el objeto "verde"?

class("verde")

is.character("verde")

Números e, pi e infinito y valores perdidos

El número e se representa en R como $exp(1)$, el número pi se escribe como pi y el infinito se muestra como Inf. Cuando no se dispone de un dato, este se representa como NA (Not available). Cuando no existe se expresa como NaN (not a number).

5. FUNCIONES MATEMÁTICAS

Algunas de las funciones básicas más utilizadas son:

Función	Significado
$\boldsymbol{log(x)}$	Logaritmo en base e de x
$\mathbf{exp(x)}$	Exponencial de x
$\mathbf{sqrt(x)}$	Raíz cuadrada de x
$\mathbf{abs(x)}$	Valor absoluto de x
$\mathbf{sin(x)}$ $\mathbf{cos(x)}$, $\mathbf{tan(x)}$	Funciones trigonométricas
$\mathbf{floor(x)}$	Mayor entero $< x$
$\mathbf{ceiling(x)}$	Menor entero $> x$
$\mathbf{round(x, digits = 2)}$	Aproximación de x a 2 decimales
$\mathbf{runif(n)}$	Genera n números aleatorios entre 0 y 1 de una distribución uniforme
Operadores aritméticos	
$+ \ - \ * \ \hat{} \ /$	

EJERCICIOS PROPUESTOS. **Use R para realizar los siguientes cálculos**

1) $\left|4^8 + 7^4\right|$
2) e^4
3) $(3 \cdot 7)^{7} + ln(9 \cdot 5) - \sin(\pi/4)$

6. OPERADORES

6.1 Operador de asignación "<-"

Está compuesto de un < ("menor") y un - ("menos" o "guion") escritos juntos. Se lee como "toma": la variable "pruebas" toma el valor c(2,5,1,6,5,5,4,1). Puede utilizarse también "=".

```
p <- 58
p
n=58
n
class(58)
class(a)
log(a)
int = as.integer(5)
f <- "Hola mundo"
f
class(f)
```

6.2 Operador de comparación (>, <, ==)

```
(2+2) <5
s<- (2+2==5)
class(s)
```

7. ESTRUCTURAS DE DATOS

Las principales estructuras de datos que pueden encontrarse en R son:

- **Vector**. Los vectores son variables con uno o más de un valor del mismo tipo: lógico, entero, real, complejo o caracter. Un vector es una matriz de una sola dimensión. Los escalares son variables de un solo elemento; sin embargo, en R, estos son considerados como vectores de longitud 1. Generación de vectores: la función *c()*.

```
y<- 4.3
x<- c()
a<- c(2,4,6)
class(a)
b<- c("rojo", "azul", "verde")
class(b)
l <- c(TRUE, FALSE, FALSE)
class(l)
x<-1:3
class(x)
```

Generación de secuencias

Las dos funciones más utilizadas para la generación de secuencias son: $seq()$ y $rep()$.

```
c <- 1:10
seq(from = 0, to = 10, by = 2)
seq(from = 0, to = 10, length.out = 15)
rep(1:3, each = 4, len = 20)
rep(1:3, each = 4, times =2)
```

Funciones con vectores

```
x<-c(8,3,5,7,6,6,2,4,9,11,2,3,7)
sum() #Devuelve la suma de los elementos de un vector
prod() #Devuelve el producto de los elementos de un vector
sort(x, decreasing=FALSE) #Devuelve el vector ordenado de menor a mayor
order(x) #Devuelve los índices de los elementos ordenados
```

```
x>5 #Vector lógico que informa de si un elemento cumple o no la condición.
which(x>5) #Índice de los elementos del vector que lo cumplen
```

Longitud de un vector:

```
length(x)
```

Acceso a los elementos de un vector

```
x[3]
x[which(x>5)]
```

- **Matriz.** Dos dimensiones. En R el objeto matriz se crea a partir de vectores, transformándolos en objetos con estructura de matriz. Por defecto, se completa añadiendo elementos por columnas.

```
mat1 <- matrix(c(1,2,3,4,5,6,7,8,9,10), nrow = 5, ncol = 2)
```

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

```
matrix(c(1,2,3,4,5,6,7,8,9,10), nrow = 5, ncol = 2, byrow=TRUE)
```

Creación de una matriz a partir de vectores:

```
x1<- c(1,2,3)
x2<- c(4,5,6)
x3<- c(0,0,0)
X<- cbind(x1,x2,x3) #Pega los vectores por columnas
Y<- rbind(x1,x2,x3) #Pega los vectores por filas
X
Y
is.matrix(X) #Determina si el argumento es un objeto de clase matriz
```

Acceso a los elementos de una matriz:

```
mat1[1,]
mat1[,1]
mat1[1:3,]
```

Funciones matriciales. Atributos de una matriz:

Función	Significado
$nrow(A)$	Número de filas
$\mathbf{ncol(A)}$	Número de columnas
$\mathbf{dim(A)}$	Dimensión
$\mathbf{length(A)}$	Tamaño
$\mathbf{mode}(A)$	Tipo de datos que contiene
$\mathbf{diag(A)}$	Diagonal
$\mathbf{rownames(A)}$	Nombre de filas
$\mathbf{colnames(A)}$	Nombre de columnas
$\mathbf{t(A)}$	Matriz traspuesta
$\det(A)$	Determinante
%*%	Producto matricial
$solve(A)$	Inversa de una matriz. También resuelve un sistema de ecuaciones lineales
Operadores aritméticos	
$+ \quad -$	

EJERCICIOS PROPUESTOS. **Use R para realizar los siguientes cálculos**

1) Genere las secuencias:
   i) "Do" "Re" "Mi" "Do" "Re" "Mi"
   ii) 2 4 6 2 4 6 2 4 6
   iii) 8.00000 8.04545 8.09091 8.13636 8.18182 8.22727 8.27273 8.31818 8.36364 8.40909 8.45455 8.50000
2) ¿Qué obtiene al sumar un vector y un escalar? Explíquelo.
3) Calcule el producto elemento a elemento de los siguientes vectores:
   i) $(\mathbf{2,5,6,7}), (\mathbf{-1,3,-1,-1})$
   ii) $(\mathbf{2,5,6}), (\mathbf{-1,3,-1,-1})$
4) Define las matrices $\boldsymbol{A} = \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{2} \\ \mathbf{2} & \mathbf{1} & \mathbf{2} & \mathbf{6} \\ \mathbf{3} & \mathbf{7} & \mathbf{5} & \mathbf{4} \end{bmatrix}$ y $\boldsymbol{B} = \begin{bmatrix} \mathbf{0} & \mathbf{2} & \mathbf{3} & \mathbf{5} \\ \mathbf{4} & \mathbf{3} & \mathbf{4} & \mathbf{1} \\ \mathbf{2} & \mathbf{1} & \mathbf{4} & \mathbf{6} \\ \mathbf{6} & \mathbf{5} & \mathbf{3} & \mathbf{1} \end{bmatrix}$
5) Calcule los productos matriciales: $\boldsymbol{AB^{-1}}$ y $\boldsymbol{BA^{T}}$

- **Array.** Tres o más dimensiones

```
arr1 <-array(c(1,2,3,4,5,6,7,8), dim = c(2,2,2))
arr1
```

- **Factor**. Los factores son un tipo de datos que sirven para almacenar una variable categórica.

```
fac <- factor(c("H","M","M","H","H","M"), levels = c("H","M"),
labels=c("Hombre","Mujer"))
```

- **Data.frame.** Ampliación de las matrices que admite variables de distintos tipos, pero con la misma longitud.

```
a <-  c(1,5,4)
b <-  c("hola", "adiós", "hasta luego")
c <-  c(TRUE, FALSE, FALSE)
dataframe <-  data.frame(a,  b,  c)

peso<-c(56,45,79,80,94,62)
talla<-c(150,155,178,180,210,145)
sexo<-c("mujer", "mujer", "hombre", "hombre", "hombre", "mujer")
sexo<-as.factor(sexo)
Aframe<-data.frame(peso, talla, sexo)
```

Acceso a los elementos de un data.frame: $, [,2], [2,]

Nombres de las variables: names(Aframe), colnames(Aframe)

- **Lista**. Es la estructura más flexible de R. Se entiende como un vector donde cada elemento a su vez puede ser cualquier otro objeto: números, vectores, matrices, data frames, funciones, listas… que no tienen por qué ser de la misma dimensión.

8. ESPACIO Y DIRECTORIO DE TRABAJO

Todos los objetos creados en una sesión de R se guardan en el conocido como espacio de trabajo (workspace).

Recuerde especificar siempre el directorio de trabajo en el que quiere trabajar al iniciar una sesión

a) ¿Qué objetos están en el espacio de trabajo?

ls()

b) Elimine alguno de los objetos del espacio de trabajo:

rm()

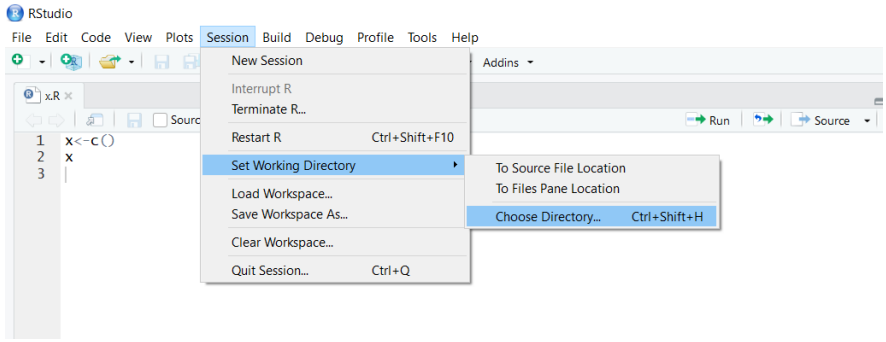
c) ¿Dónde está situado su directorio de trabajo?

getwd()

d) Cambie su directorio de trabajo y sitúelo en el escritorio.

setwd()

Menú Sesión -> set working directory -> Choose Directory



Al salir del programa, R preguntará si se desea guardar la imagen de su espacio de trabajo. En caso de guardarlo, se almacenarán todos los objetos creados y estarán disponibles en la siguiente sesión. En caso negativo, se perderán todos los objetos creados. En este caso, en una nueva sesión de R, tan solo estarán disponibles los objetos del espacio de trabajo creados en la última sesión guardada.

9. INSTALACIÓN DE PAQUETES

a) Utilice los comandos install.packages("") y library() para instalar el paquete foreign.

b) Utilice el menú de instalación de paquetes para instalar el paquete ggplot2.

9.1. Comandos de ayuda.

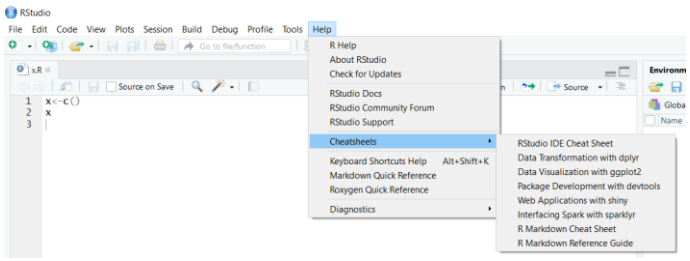
a) Revise la ayuda de la función $plot()$.

help(plot)
?plot

b) ¿Dónde encontrar manuales de R?

Menú Help

¡Las Hojas de Trucos son muy útiles (Cheat Sheets) !



10. IMPORTACIÓN DE BASES DE DATOS

A partir de datos almacenados en vectores es posible generar matrices de datos en R. Sin embargo, esto puede ser complicado para bases de datos con volúmenes de datos grandes. En estos casos, lo habitual es importar la base de datos desde otros formatos más manejables. Cuidado con la dirección del directorio de trabajo.

a) Importe la base de datos iris en formato csv:

- Copiando y pegando los datos del archivo. No siempre funciona.
  data.csv<-read.csv("clipboard", header=TRUE, sep=";")

- Eligiendo el archivo de una manera interactiva.
  data.csv <- read.csv(file.choose(), header= TRUE, sep=";")

- Directamente proporcionando el nombre del archivo.
  data.csv <- read.csv("iris.csv", header= TRUE, sep=";")

b) Importe la base de datos iris desde el archivo con extensión .txt

```
data.txt<-read.table("iris.txt", header=TRUE, sep="\t")
data.txt<-read.table(file.choose())
data.txt<-read.table("clipboard", header=TRUE)
```

c) Importe la base de datos iris desde el archivo con extensión .xlsx

Environment -> Import Dataset -> From Excel

También es posible utilizar bases de datos disponibles en R.

d) Cargue la base de datos: airquality del paquete datasets

```
data(airquality)
View(airquality)
```

11. EXPORTACIÓN DE BASES DE DATOS

a) Exporte la base de datos airquality en formato csv y en formato txt.

```
write.csv(airquality, file = "datos_prueba.csv")
write.table(airquality, file = "datos_prueba.txt", sep = "\t", dec=".")
```

b) Guarde toda la información creada en un archivo .RData

File => Save As => "Nombre.RData"

EJERCICIOS PROPUESTOS. **Use R para realizar los siguientes cálculos**

1) Construya un factor, denominado f, con los niveles “Sin estudios”, “Primaria”, “Secundaria”, “Universitarios” que tengan las frecuencias: 15, 35, 45, 40.
2) Considere los números: 8, 6.3, 10.1, 5.25, 4.02, 2.3, 20, 0.5
    a. Calcule su suma y su producto.
    b. Calcule la raíz cuadrada de los números.
    c. Obtenga los números mayores que su raíz cuadrada.
    d. ¿Cuántos valores son mayores que 5?
3) Genere un archivo de datos en Excel que contenga la información de las variables: edad, género, número de asignaturas aprobadas de una clase hipotética de 9 alumnos. Guárdelo en formato txt (delimitado por tabulaciones) e impórtelo a un data frame R.
4) Compruebe la clase del objeto importada, la dimensión y el nombre de las columnas.
5) Instale el paquete de R *Lock5Data.*
6) Busque en el menú de ayuda información sobre la base de datos *SleepStudy*.
7) Guarde en un data frame llamado *bebida* las variables “Gender”, “ClassYear”, “Happiness” y “AlcoholUse” del conjunto de datos *SleepStudy* del paquete *Lock5Data* de R

2. Ordenación de los datos: frecuencias y representaciones gráficas

OBJETIVOS
Resumir, ordenar y analizar conjuntos de datos
Ordenar los datos mediante tablas de frecuencias.
Representar gráficamente la información de una variable cualitativa.
Representar gráficamente la información de una variable cuantitativa.
Conocer y utilizar las funciones básicas de R de representación gráfica.

En esta práctica se emplean los datos del conjunto presidentes, disponible en el paquete *paqueteadp* (Cruz, González & González, 2022), que forma parte del material complementario del libro *AnalizaR Datos Políticos*. La base de datos fue propuesta por Reyes-Housholder (2019) y ampliada con información de los Indicadores de Desarrollo Mundial del Banco Mundial. Se recoge la información de las variables país, anio, trimestre (cuatrimestre), presidente, presidente_genero (Sexo del presidente), presidentes_neta (Aprobación presidencial neta (% de aprobación - % de desaprobación)), pib (Producto Interno Bruto del país), población, corrupcion (Corrupción del Poder Ejecutivo, según V-Dem. De 0 a 100 (un número mayor significa más corrupción)), desempleo (Tasa de desempleo) y crecimiento_pib, según los presidentes de los países y el año de gobernanza. La autora argumenta que las presidentas de América Latina sufren caídas más pronunciadas en la aprobación en comparación con sus homólogos masculinos ante escándalos de corrupción similares.

a) Importe la base de datos *presidentes*.

Import dataset -> From Excel

También podría cargar los datos directamente en R, al estar disponibles en R en la librería *paqueteadp* (bajo el nombre aprobación).

View(presidentes); class(presidentes)

presidentes<-as.data.frame(presidentes); dim(presidentes)

¿Tipo de objeto? data.frame - Dimensión de la base de datos: 1020x11

0. FUNCIONES GRÁFICAS BÁSICAS DE R

R dispone de algunas funciones básicas para la inspección visual y rápida de datos (véase Tabla 1). Existen otras librerías de R para generar gráficos más sofisticados o atractivos estéticamente, como el paquete ggplot2 con el que trabajaremos más adelante o incluso paquetes con funciones diseñadas para la representación de gráficos interactivos, como la librería plot_ly o la función ggplotly para hacer interactivos los gráficos de ggplot2.

En esta sección nos centraremos en conocer las funciones principales y básicas de R para la representación gráficas de variables cualitativas y cuantitativas. Además, trabajaremos con algunos (de los muchos) argumentos existentes para mejorar la apariencia de estos gráficos: ejes, colores, tamaño de letra, …

Tabla 1 Funciones gráficas básicas

Función	Significado
stem(data)	Diagrama de tallos y hojas
hist(data)	Histograma
boxplot(data)	Caja y bigotes
plot(data)	Gráfico de puntos / dispersión
pairs(data)	Gráfico de dispersión cruzado
barplot(data)	Diagrama de barras

0.1. Argumentos de las funciones gráficas

Los gráficos básicos de R pueden hacerse más atractivos a nivel estético a partir de la modificación de los parámetros gráficos de este tipo de funciones. Algunos de los argumentos más útiles se muestran en la tabla 2.

Tabla 2 Algunos de los argumentos más útiles para la modificación de parámetros gráficos

Función	Significado
main	Título del gráfico
Sub	Subtítulo del gráfico
type	Tipo de gráfico
xlab, ylab	Etiquetas de los ejes
las	Posición de las etiquetas de los ejes
xlim, ylim	Rango de valores de los ejes
col	Vector carácter que cambia los colores con los que se pinta
pch	Forma de los puntos
lty	Tipo de línea
lwd	Grosor de línea

Además, hay una serie de funciones que permiten añadir elementos sobre una gráfica ya creada (Tabla 3).

Tabla 3 Algunos de los posibles elementos a añadir sobre una gráfica ya creada

Función	Significado
points(x,y,…)	Añade nube de puntos
lines(x,y,…)	Añade una línea que une los puntos
abline()	Útil para gráficar línea vertical/horizontal
text(x,y,labels,…)	Añade etiquetas de texto en las coordenadas x,y
legend()	Añade una leyenda al gráfico

Personalización el entorno gráfico mediante cambios permanentes:

La función par()

Cada vez que se crea una gráfica nueva, se abre un dispositivo gráfico con una serie de parámetros predefinidos con unos valores iniciales por defecto. Hasta ahora hemos visto como modificar parámetros gráficos de manera **temporal**, incluyéndolos como **argumentos** en las llamadas a las funciones gráficas (como xlim, xlab, pch, …). Sin embargo, estos parámetros pueden ser modificados y fijados de forma **permanente** mediante la función **par**().

De entre todas las opciones que permite modificar la función par() (puede consultar su ayuda) se presentan a continuación dos parámetros útiles a la

hora de agregar varias gráficas en una misma ventana. Son los parámetros mfrow() y mfcol(). Veáse cómo funciona el primero de ellos mediante un ejemplo: par(mfrow=c(1,2)), permite al usuario colocar dos gráficos en una misma fila con dos columnas. Por otro lado, par(mfrow=c(2,1)) permitiría colocar dos gráficos en dos filas diferentes de una misma venta, sobre una sola columna. El parámetro mfcol() funciona de manera similar, pero rellenando la ventana con los gráficos por columnas.

1. ANÁLISIS DE UNA VARIABLE CUALITATIVA: TABLA DE FRECUENCIAS Y GRÁFICOS

Cualquier variable estadística toma distintos valores x_1; x_2; …, x_k, pero cada uno de estos valores puede aparecer repetido más de una vez, por lo que sería conveniente hacer una ordenación de los datos para resumir la información entregada por ellos. Cuando las variables son cualitativas, las categorías de agrupación vienen determinadas por la propia variable.

TABLAS DE FRECUENCIAS. Veamos en primer lugar como obtener las frecuencias absolutas de la variable cualitativa nominal *presidentes_genero*.

1.1. Cálculo de frecuencias absolutas y frecuencias absolutas acumuladas

a) Calcule las frecuencias absolutas y absolutas acumuladas de la variable presidentes_genero. ¿Cuál fue el sexo predominante en las presidencias del gobierno?

```
frec.abs<-table(presidentes$presidente_genero)
frec.abs
frec.abs.acum<-cumsum(frec.abs)
```

	Frecuencias absolutas	***Frecuencias absolutas acumuladas (no sentido)***
Mujer	98	98
Hombre	922	1020

A continuación, para obtener las frecuencias relativas, podemos obtener estas directamente dividiendo por el tamaño muestral.

1.2. Cálculo de frecuencias relativas y frecuencias relativas acumuladas

a) Calcule las frecuencias relativas y porcentajes de la variable *presidente_genero*. ¿Cuál fue el porcentaje de la muestra de mujeres presidentas?

```
n<-dim(presidentes)[1]
frec.rel<-frec.abs/n
porc<-frec.rel*100
round(porc, digits=2)
porc.acum<-cumsum(round(porc, digits=2))
```

	Frecuencias relativas	***Frecuencias relativas acumuladas***	***Porcentaje (%)***
Hombre	0.096	0.096	9.61
Mujer	0.904	1.000	100.00

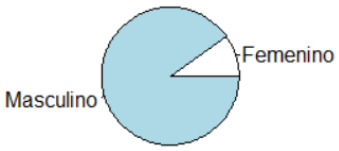
```
tabla.frec<-cbind(frec.abs, frec.abs.acum,
      round(frec.rel,3), round(porc,3), porc.acum)
colnames(tabla.frec)<-c("fi","Fi","hi", "%", "% acumulado")
tabla.frec

install.packages("epiDisplay")
library(epiDisplay)

tab1(presidentes$presidente_genero, bar.values = "percent")
```

REPRESENTACIONES GRÁFICAS. Una vez ordenados los datos mediante una tabla de frecuencias podemos resumir la información de una variable mediante una representación gráfica. Estas tienen el objetivo de transmitir la información de una variable de manera visual.

1.3. Gráfico de sectores.

a) Represente en un gráfico de sectores la información del género de los presidentes.

```
pie(table(presidentes$presidente_genero))
```

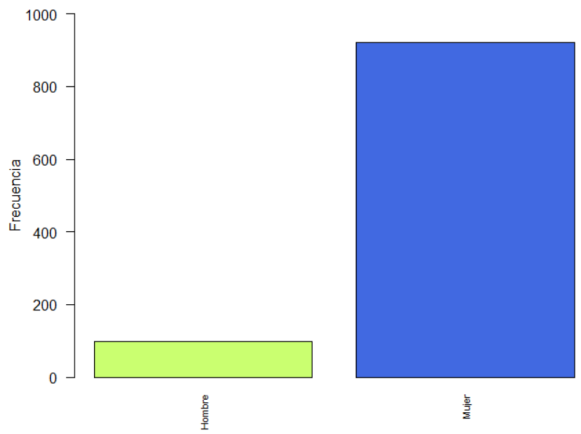



```
genero.data <- data.frame(genero = presidentes$presidente_genero)
install.packages("lessR")
library(lessR)
fa<-table(presidentes$presidente_genero)
PieChart(fa, hole = 0, values = "%", data = genero.data, main = "")
```

1.4. Diagrama de barras.

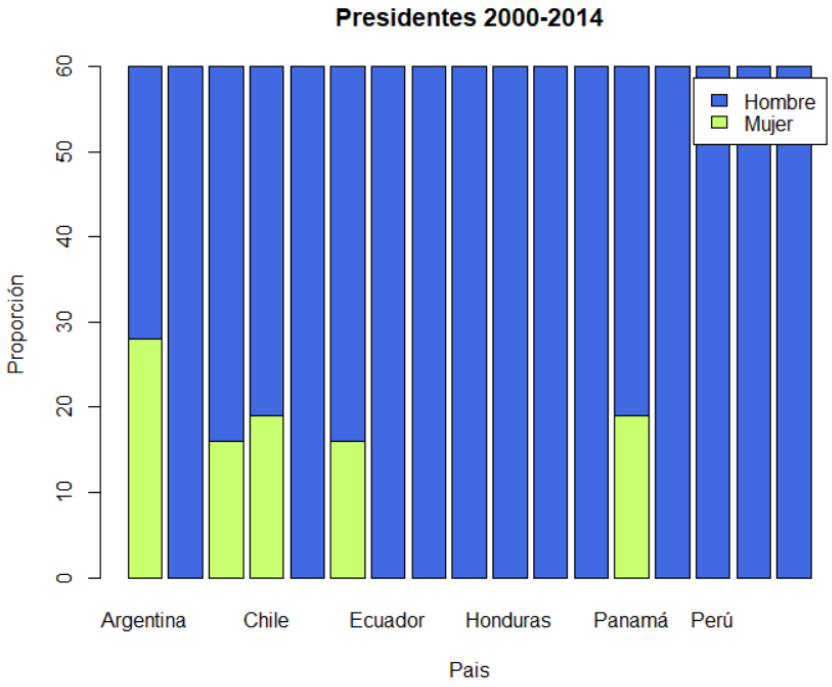
a) Represente en un diagrama de barras la información de las frecuencias absolutas y relativas de la variable *presidentes_genero*.

```
barplot(table(presidentes$presidente_genero),  main =" ",  xlab = "Género",
ylab = "Frecuencia", col = c("darkolivegreen1", "royalblue"))
```

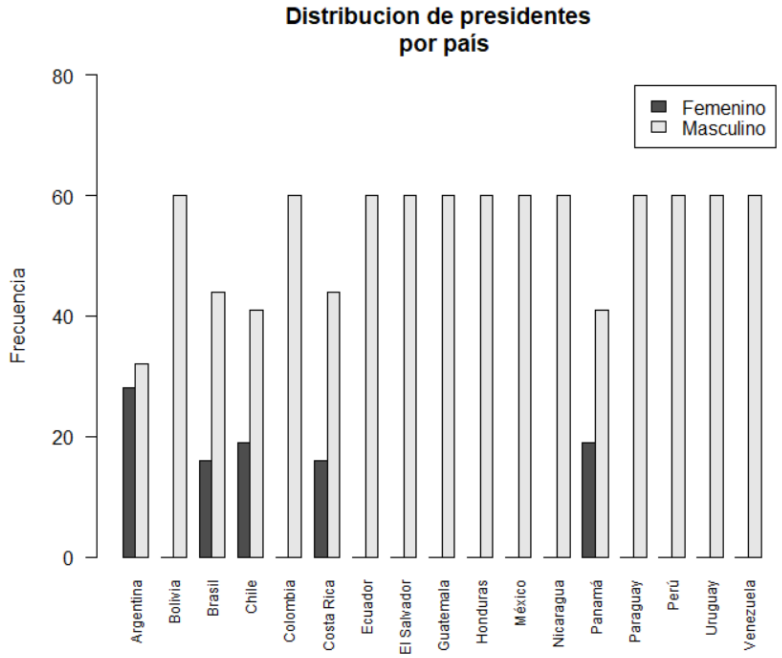


1.5. Diagrama de barras agrupadas/apiladas.

```
barplot(table(presidentes$presidente_genero,presidentes$pais),  main =
        "Presidentes 2000-2014",  xlab = "Pais", ylab = "Proporción",
col = c("darkolivegreen1", "royalblue"), legend.text=c("Mujer","Hombre"))
```




```
barplot(table(presidentes$presidente_genero,presidentes$pais), beside=T,
legend.text=
rownames(table(presidentes$presidente_genero,presidentes$pais)),
ylim=c(0,80), main="Distribucion de presidentes \n por país",
ylab="Frecuencia",cex.names =0.75,las=2)
```



2. ANÁLISIS DE UNA VARIABLE CUANTITATIVA: TABLAS DE FRECUENCIAS Y GRÁFICOS

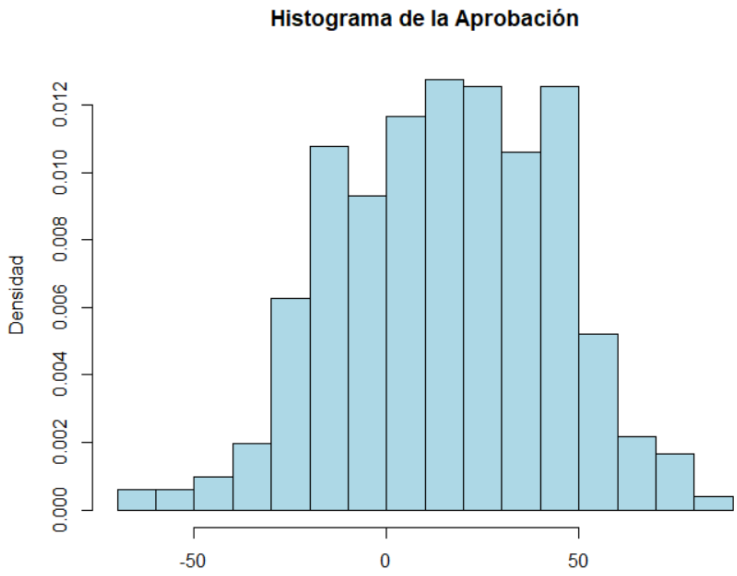
REPRESENTACIONES GRÁFICAS. En el caso de variables cuantitativas, los gráficos que más se suelen utilizar para explorar el comportamiento de la variable son el histograma, el diagrama de tallo y hojas y el diagrama de caja y bigotes. Este último se verá en futuras clases. También es de gran utilidad el diagrama de dispersión.

2.1. *Histograma.* Los principales argumentos de la función para graficar un histograma son:

hist(x, breaks = “Sturges”, right = TRUE, col = NULL, main = paste(“Histogram of” , xname))

a) Realice un histograma para la variable “presidentes”.

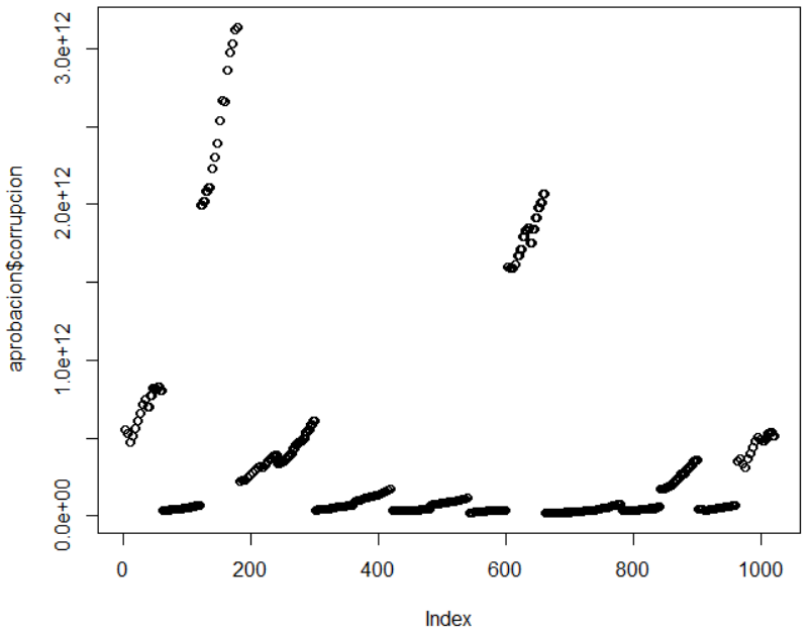
```
hist(presidentes$presidentes_neta, freq = FALSE, col = " lightblue",
main = "Histograma de la Aprobación", ylab = "Densidad", xlab = "",
breaks = 15)
```



2.2. *Diagrama de dispersión 2D.* Los gráficos de dispersión en R se obtienen a partir de la función plot:

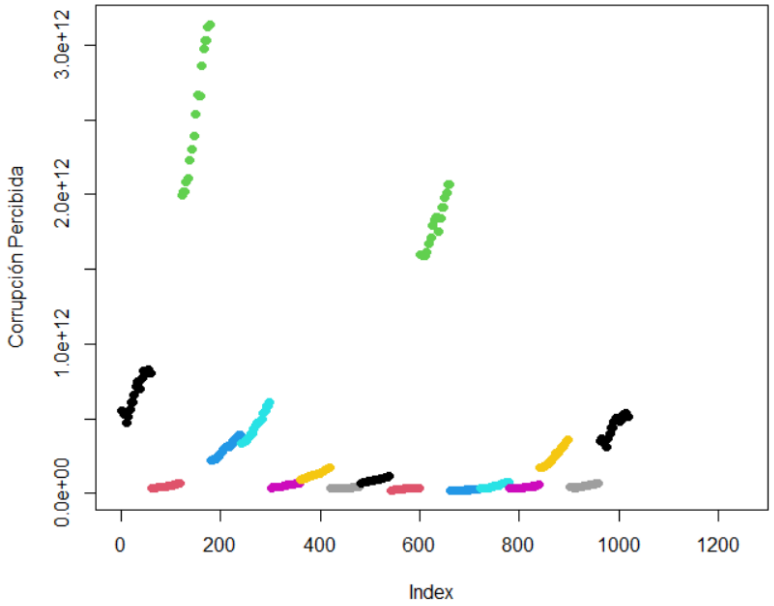
a) Gráfico de dispersión de la variable "corrupción"

```
plot(presidentes$corrupcion) #Gráfico de dispersión de una variable
```



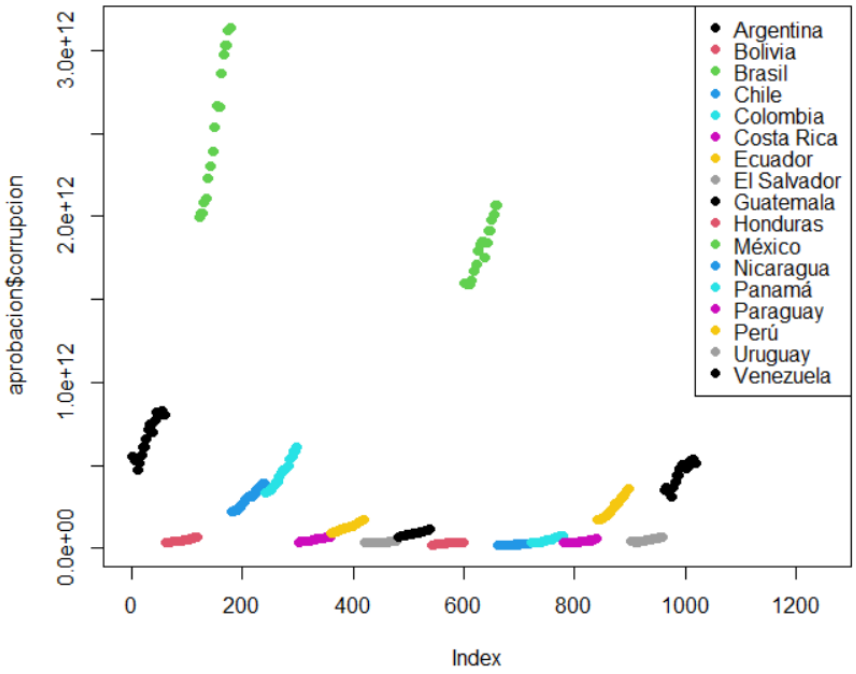
b) Mejore el gráfico anterior, modificando algunos argumentos de la función plot:

```
plot(presidentes$corrupcion, ylab="Corrupción Percibida")
#Etiquetas de ejes
plot(presidentes$corrupcion, xlim=c(0,1250)) #longitud del eje
plot(presidentes$corrupcion, col= as.factor(presidentes$pais))
#colores
plot(presidentes$corrupcion, pch=19) #forma
```



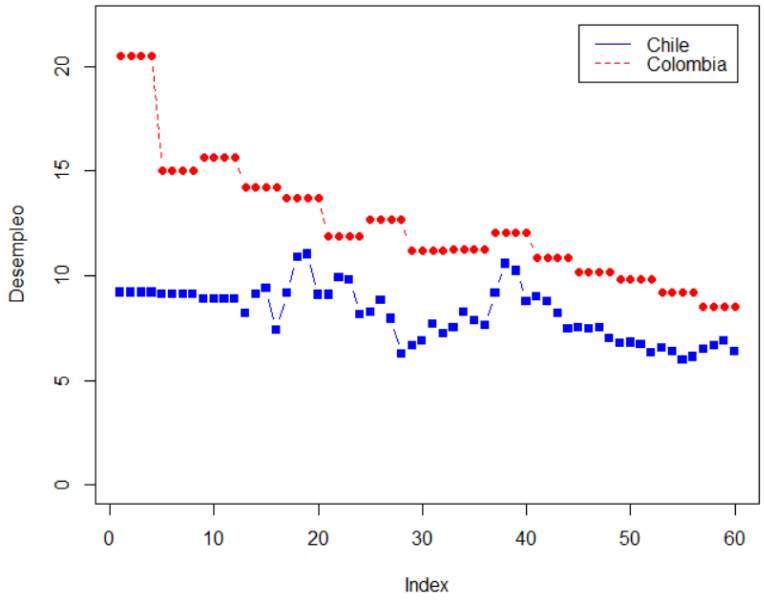
Añadir elementos externos:

```
plot(presidentes$corrupcion, col=as.factor(presidentes$pais),
pch=19,xlim=c(0,1250))
legend("topright",unique(presidentes$pais),col=unique(as.factor(pres
identes$pais)), pch=19) #legenda
```



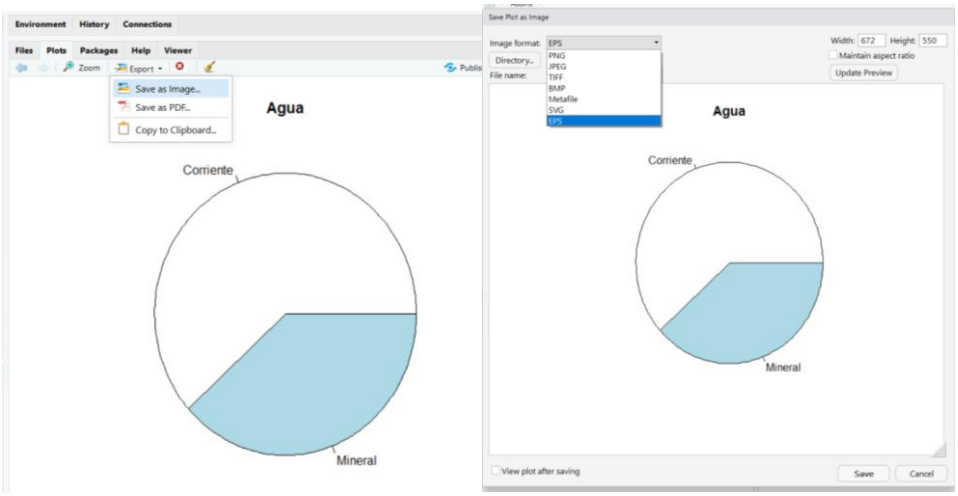
c) Gráfico con líneas:

```
plot(presidentes$desempleo[181:240], type="b", lty=1, col="blue",
pch=15,  ylim=c(0,22), ylab="Desempleo")
lines(presidentes$desempleo[241:300], type="b", lty=2, col="red",
pch=19)
legend(45,22,legend=c("Chile","Colombia"),col=c("blue","red"),
lty=1:2)
```



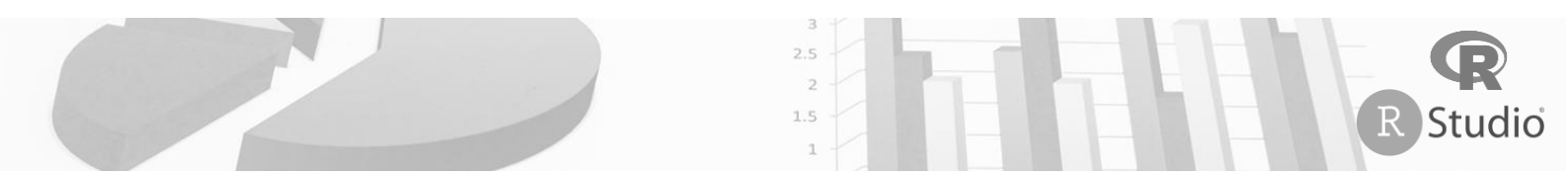
3. EXPORTACIÓN DE GRÁFICOS

Aunque existen también funciones para exportar gráficas, de manera sencilla estas pueden guardarse a partir del menú Exportar del panel de salidas gráficas, y guardar la imagen como un fichero de tipo pdf, png, tiff, jpg, eps…



REFERENCIAS

Cruz, A., González, G., & González, F. (2022). *AnalizaR Datos Políticos: Ciencia Política y R*. Santiago de Chile: Ediciones Universidad Alberto Hurtado. Disponible en https://github.com/arcruz0/paqueteadp



3. Medidas de resumen. Tendencia central, dispersión, posición y forma

OBJETIVOS
Análisis descriptivo de una base de datos: análisis univariante. **Cálculo las principales medidas de una variable estadística:** **1. Medidas de tendencia central y dispersión.** **2. Medidas de posición.** **3. Medidas de forma.** **Análisis exploratorio de los datos.** **Representar gráficamente el comportamiento de la muestra con respecto a una variable. El gráfico Box-Plot.**

En la práctica anterior se tomó contacto con la ordenación y representación de un conjunto de datos, mediante las tablas de frecuencias y las representaciones gráficas. En la práctica de hoy se completará dicho estudio mediante el análisis exploratorio de una base de datos, describiendo la información de las variables a través de las principales medidas de resumen o síntesis. Con ellas, se logrará obtener una idea acerca del comportamiento general de la muestra mediante las medidas de tendencia central (media, mediana, moda), dispersión (recorrido, desviación típica, varianza, rango intercuartílico) y posición (cuartiles, percentiles) así como de la forma de la distribución de los valores de una variable (a través de las medidas de simetría y curtosis). Toda la información se complementará mediante representaciones gráficas, algunas ya vistas en la clase anterior y otras novedosas como es el caso del Box-Plot.

MEDIDAS DE TENDENCIA CENTRAL	
Función	**Significado**
mean(x)	Media
median(x)	Mediana
MEDIDAS DE DISPERSIÓN	
Función	**Significado**
min(x)	Mínimo de un conjunto de datos
max(x)	Máximo de un conjunto de datos
range(x)	Rango / recorrido
var(x)	Varianza
sd(x)	Desviación típica
IQR(x)	Rango intercuartílico
std.error() Library: plotrix	Error estándar
MEDIDAS DE POSICIÓN	
Función	**Significado**
quantile(x)	Cuartiles de un conjunto de datos x
quantile(x, 15)	Percentil P_{15} de un conjunto de datos x
MEDIDAS DE FORMA	
Función	**Significado**
skewness() Library: e1071	Simetría de un conjunto de datos
kurtosis() Library: e1071	Curtosis de un conjunto de datos
FUNCIÓN RESUMEN	
summary()	

Los datos con los que vamos a trabajar se encuentran en el fichero "presidentes" ya trabajados en la práctica 2. Los datos proceden del conjunto presidentes del paquete *paqueteadp* (Cruz, González & González, 2022). Cargue o importe la base de datos desde R y realice las siguientes tareas:

1. Recodifique la variable "Aprobación" en una variable que llamará "aprobacionCOD", en la que se distribuirá la puntuación de aprobación del poder ejecutivo en los siguientes grupos: 1=[-100,-50); 2=[-50, 0); 3=[0,50), 4=[50,100).

```
aprobacionCOD<-cut(presidentes$ aprobacion_neta, breaks=c(-100,-50, 0, 50,100), labels=c("Desaprobación alta", "Desaprobación baja", "Aprobación baja", "Aprobación alta"))
```

2. Rellene la siguiente tabla y represente un diagrama de sectores para la variable aprobacionCOD:

Aprobación del poder ejecutivo	**Frecuencia absoluta**	**%**
Desaprobación alta	12	1.18
Desaprobación baja	299	29.31
Aprobación baja	613	60.10
Aprobación alta	96	9.41

```
table(aprobacionCOD)
round((table(aprobacionCOD)/1020)*100,2)
```

3. Realice un estudio descriptivo y exploratorio de las variables cuantitativas siguientes:

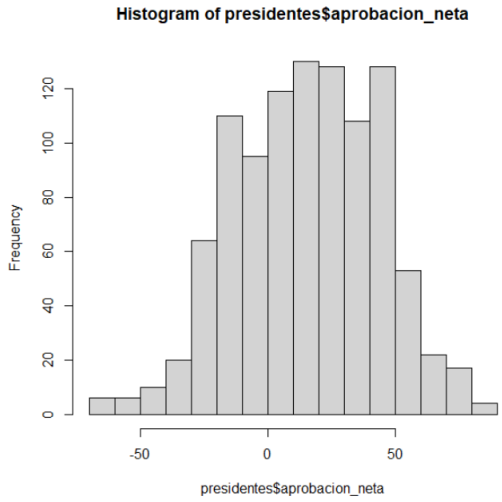
	Media	**Desviación típica**	**Mediana**	**Mínimo**	**Máximo**
PIB	43.35	26.9	39.98	2.13	94.42
Desempleo	7.043	3.61	6.41	1.30	20.52

```
summary(presidentes$pib)
summary(presidentes$desempleo)
```

```
sd(presidentes$pib)
sd(presidentes$desempleo)
library(psych)
describe(presidentes$desempleo)
```

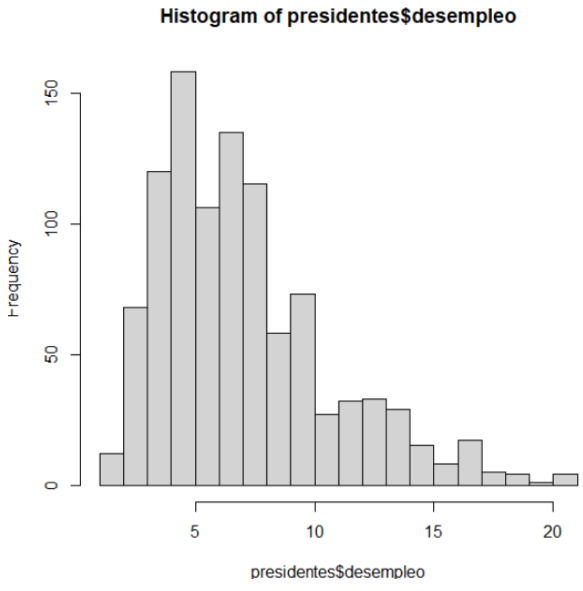
4. **Análisis gráfico de una variable cuantitativa**. Realice los gráficos adecuados para la variable "desempleo" para saber qué medida de tendencia central será más adecuada para describir los datos. ¿Qué se puede afirmar con respecto a la simetría de esta variable?

```
hist(presidentes$aprobacion_neta)
```



¿Y qué ocurriría con la variable desempleo?

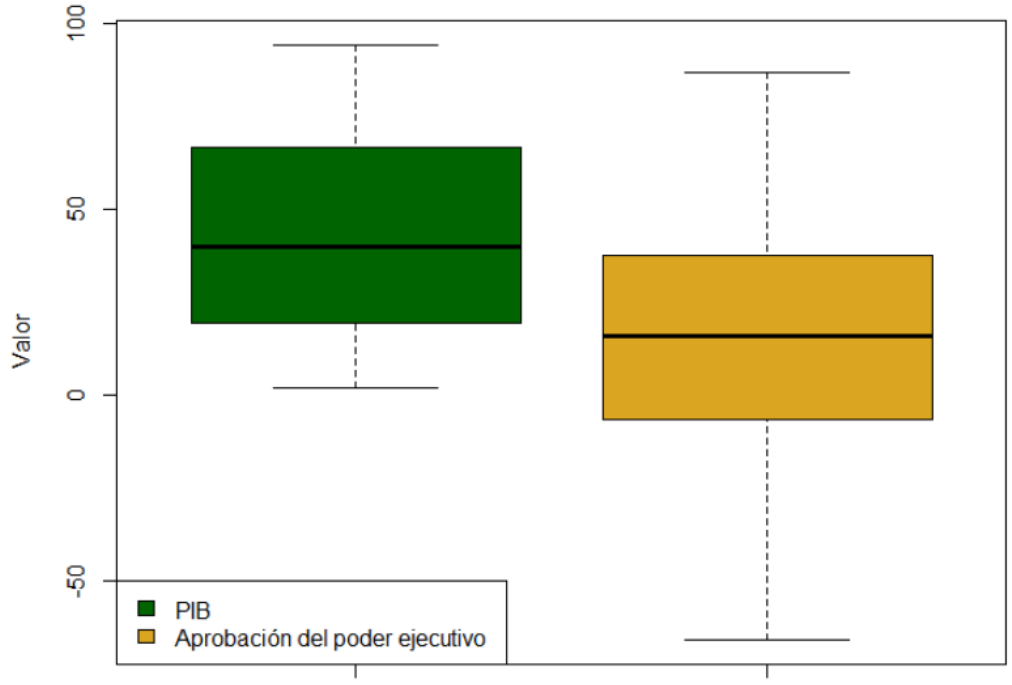
```
hist(presidentes$desempleo)
```



5. Complete la siguiente tabla. Además, realice dos boxplot: uno para la variable "PIB" y otro para "Aprobación" y compare los resultados con los de la pregunta anterior.

	Mediana	**RI**	**Q_1**	**Q_2**	**Q_3**	**P_{80}**
Desempleo	6.41	4.52	4.38	6.41	8.90	9.4

```
median(presidentes$desempleo)
IQR(presidentes$desempleo)
quantile(presidentes$desempleo)
quantile(presidentes$desempleo, 0.8)
```




```
boxplot(presidentes$pib,presidentes$aprobacion_neta,
        ylab="Valor", col=c("darkgreen","goldenrod"))
legend("bottomleft",c("PIB","Aprobación del poder ejecutivo"),
fill = c("darkgreen","goldenrod"))
```

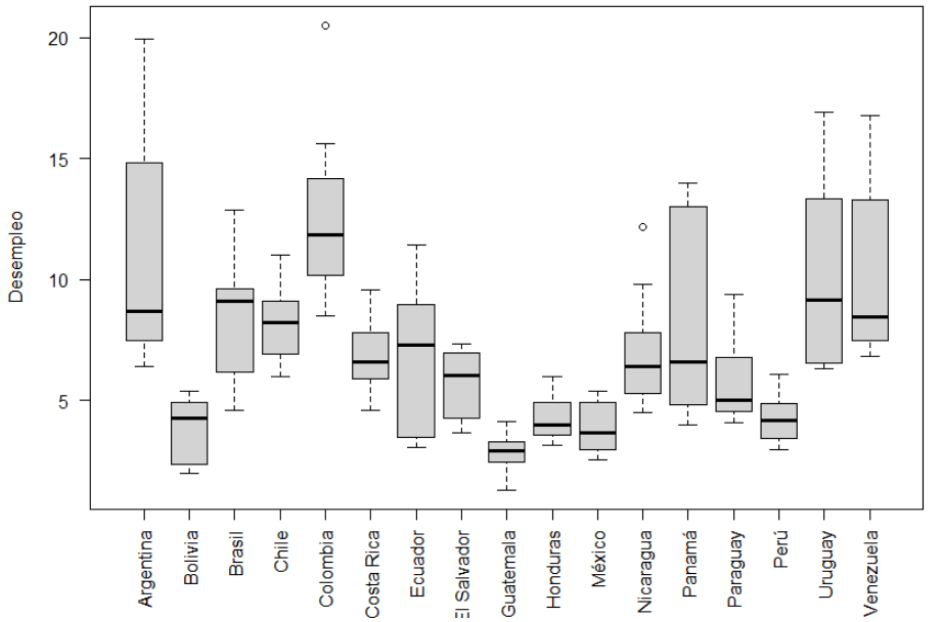
6. **Seleccionar casos**. Realice un estudio descriptivo de la variable "desempleo" para los datos de Brasil y Guatemala. ¿Qué observa para ambos países?

```
library(dplyr)
brasil<-filter(presidentes, pais=="Brasil")
dim(brasil)

guatemala<-filter(presidentes, pais=="Guatemala")
dim(guatemala)
```

	Mediana	RI	Q_1	Q_2	Q_3	P_{80}
Desempleo global	6.41	4.52	4.38	6.41	8.90	9.4
Brasil	9.1	3.32	6.25	9.10	9.57	10
Guatemala	2.91	0.86	2.46	2.91	3.32	3.34

7. **Segmentar casos.** Realice un boxplot de la variable "Desempleo", pero diferenciando los datos por país (segmentar por pais). ¿En qué país la tasa de desempleo es mayor? Complete la siguiente tabla

```
boxplot(presidentes$desempleo~presidentes$pais,las=2, xlab="",
ylab="Desempleo")
```

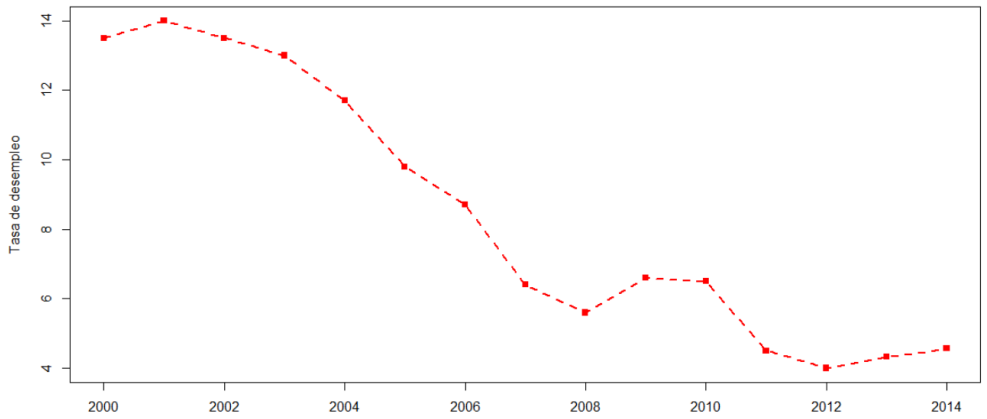


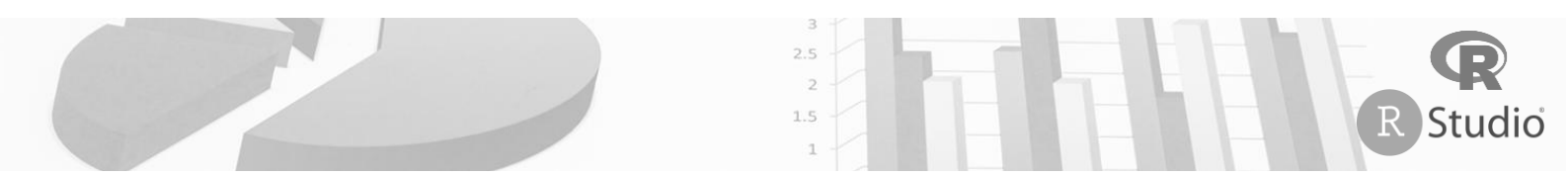
```
library(dplyr)
presidentes%>%
 group_by(pais)%>%
 summarise(mean(desempleo), sd(desempleo))
```

Desempleo	Media	Desv. Estándar
Bolivia	3.7	1.3
Colombia	12.5	2.98
Nicaragua	6.78	2.02

8. Guarde los resultados de la tasa de desempleo por países calculados en el apartado anterior mediante la función group_by en un objeto de nombre *evolución*. Posteriormente, analice cómo ha sido la evolución de la tasa de desempleo media en Panamá mediante un gráfico de líneas desde el año 2000 hasta el 2014. ¿En qué país año fue mayor la tasa de desempleo?

```
evolucion<-presidentes%>%
group_by(pais, anio)%>%
summarise(mean(desempleo), sd(desempleo))
evolucion2<-filter(evolucion, pais=="Panamá")

plot(evolucion2$anio, evolucion2$`mean(desempleo)`, xlab="", ylab="Tasa de desempleo", type="o", pch=15, col="red", lwd=2, lty=2)
```





REFERENCIAS

Cruz, A., González, G., & González, F. (2022). *AnalizaR Datos Políticos: Ciencia Política y R* [Paquete de software y material docente]. Ediciones Universidad Alberto Hurtado. Disponible en https://github.com/arcruz0/paqueteadp

4. Análisis de regresión lineal (simple y múltiple) y regresión no lineal

OBJETIVOS
Estudio de la correlación entre variables gráfica y analíticamente: diagrama de dispersión y coeficiente de correlación de Pearson. **Ajuste del correspondiente modelo de regresión lineal simple y múltiple y estimación de los parámetros.** **Análisis de la bondad de ajuste. Poder explicativo vs poder predictivo.** **Análisis de los residuos.** **Ajuste de los datos a otro tipo de modelos: logarítmico, potencial, exponencial y parabólico.**

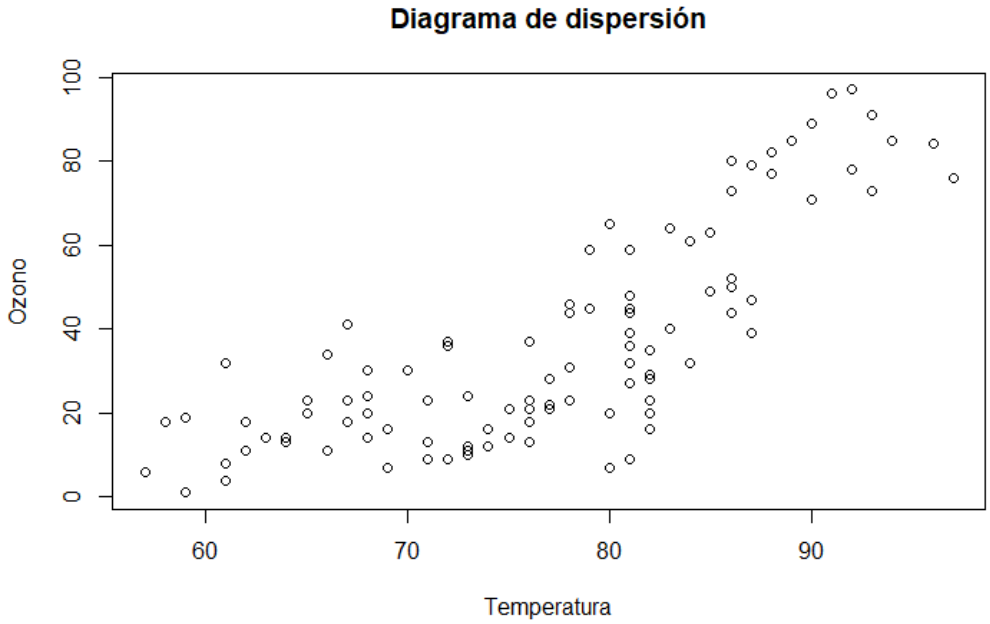
PARTE 1 – MODELO LINEAL

El objetivo del análisis de regresión es la determinación de un modelo que permita explicar el comportamiento de una variable (conocida como variable dependiente o explicada) a partir de los valores de otra u otras dadas (las llamadas variables independientes o explicativas). Esto es, se tratará de encontrar la función que mejor explique la relación entre las variables dadas.

Para la práctica de hoy trabajaremos analizando el conjunto de datos: "*airquality*", de la librería datasets de R. La base de datos almacena información acerca de una serie de mediciones diarias de la calidad del aire en Nueva York, en el año 1973. En concreto, se recoge información del día y mes en que se realiza la medición, temperatura, velocidad del viento y concentración de ozono en el aire. Según los especialistas, los valores de la temperatura pueden afectar a la concentración de ozono en el aire. Por ello, el objetivo es analizar la hipótesis de los investigadores y examinar la posible relación existente entre la temperatura y los niveles de ozono.

Cargue la base de datos *airquality*, disponible en Studium, y realice las siguientes tareas:

1. Represente un diagrama de dispersión para inspeccionar la posible relación existente entre las variables *Temperatura* y *Ozono.*



plot(airq$Temp, airq$Ozone, main="Diagrama de dispersión", xlab="Temperatura", ylab="Ozono")

¿Cuál es la variable independiente? Temperatura

¿De qué tipo parece ser la relación entre ambas variables? Lineal directa ¿por qué?

En la disposición de la nube de puntos del diagrama de dispersión se observa una relación directa entre la temperatura y el ozono. A mayor temperatura, mayor ozono.

2. Calcule e interprete la covarianza y el coeficiente de correlación de Pearson entre las variables Temperatura y Ozono.

```
cov(airq$Temp, airq$Ozone, use = "pairwise.complete.obs")
cor(airq$Temp, airq$Ozone, use = "pairwise.complete.obs")
```

¿Existe relación lineal entre las variables estudiadas? Si ¿Qué coeficiente utiliza para responder a esta pregunta? El coeficiente de correlación de Pearson

¿Cuánto vale la covarianza entre las dos variables estudiadas? ¿Y el coeficiente de correlación?

$$S_{xy} = 189{,}79 > 0$$

$$r = 0{,}78 > 0$$

El valor del coeficiente de correlación de Pearson entre Temperatura y Ozono es de 0,78. Por tanto, nos indica que la relación entre ambas es directa (su signo es positivo) y fuerte (su valor absoluto está próximo a 1).

3. Ajuste un modelo de regresión lineal e interprete cada uno de los coeficientes del modelo.

```
recta <- lm(Ozone~Temp, airq)
```

$$\boldsymbol{Y = a + bX + e}$$

$$\boldsymbol{\hat{Y} = -125{,}52 + 2{,}098X}$$

$$\boldsymbol{\widehat{Ozono} = -125{,}52 + 2{,}098 \cdot Temp}$$

El valor estimado de a es -125,52 y nos indica el valor estimado de ozono para una temperatura de 0 grados. En este caso no tiene sentido esta interpretación, ya que el rango de temperaturas usado en este ejemplo está entre 57 y 97 grados, por lo que el valor 0 queda fuera del rango de temperatura. El valor estimado de b, el coeficiente de regresión, es 2,098 e indica el incremento promedio de ozono por cada incremento en una unidad de la temperatura (incremento por su signo positivo). Esto es, se espera que los valores de concentración de ozono aumenten 2,098 cuando la temperatura aumente un grado.

4. Estudie la bondad del ajuste para el modelo obtenido.

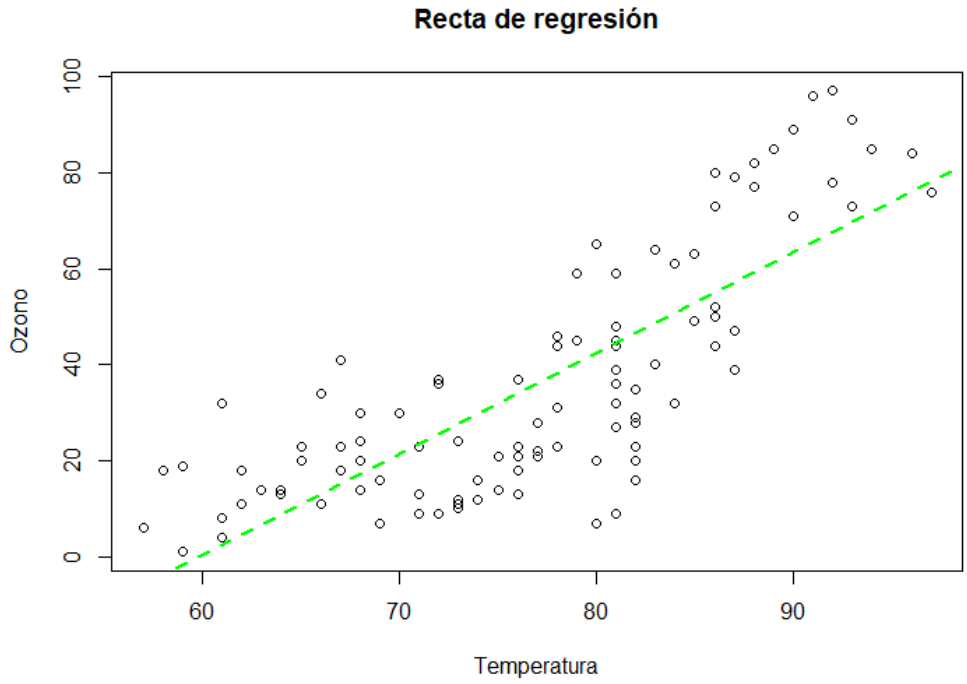
¿Considera que el modelo permite explicar adecuadamente la variabilidad del ozono a través de su relación con la temperatura? ¿Por qué? ¿Qué coeficiente, o medidas utiliza para responder?

summary(recta)

$R^2 = 0{,}606 = 60{,}6\%$

El coeficiente de determinación R^2 nos ayuda a evaluar el poder explicativo. Vale 0,606, lo que indica que usando el modelo de regresión lineal, la temperatura explica el 60,6% de la variabilidad total de los valores de ozono. Se corresponde con el cuadrado del coeficiente de correlación. También se dispone del coeficiente de determinación ajustado, R^2_{adj}=0,602; cuyo valor es similar y se interpreta en los mismos términos.

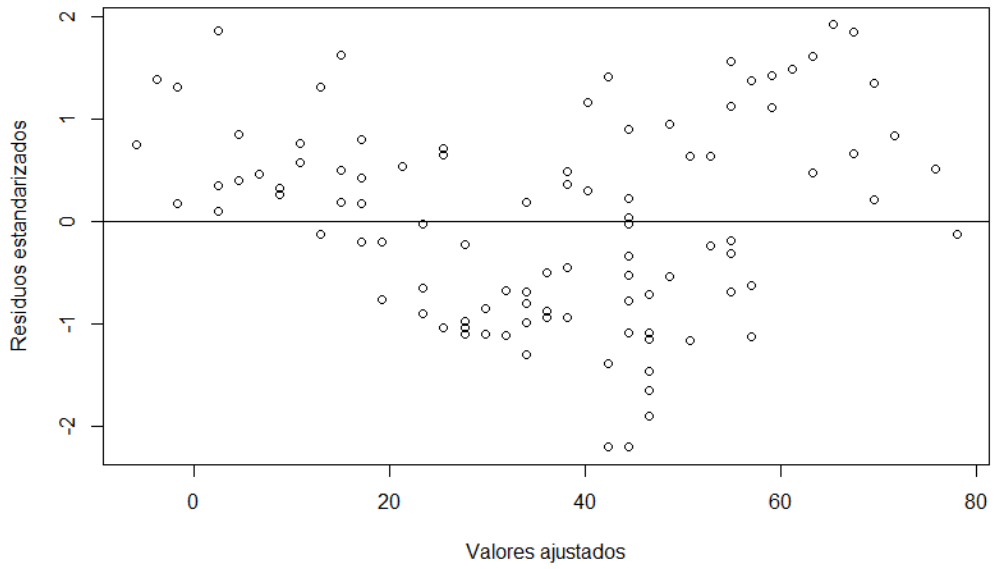
5. Represente sobre el gráfico de dispersión la línea recta del modelo que acaba de ajustar.



plot(airq$Temp, airq$Ozone, main="Recta de regresión",
xlab="Temperatura", ylab="Ozono")
abline(recta, col = "green", lty=2, lwd=2)

6. Analice el poder predictivo del modelo. ¿Cuál será el valor esperado de concentración de ozono con una temperatura de 75 grados? ¿Y si fuese de 250? ¿Son igualmente fiables dichos pronósticos?

```
plot(fitted.values(modelo),rstandard(modelo), xlab="Valores ajustados", ylab="Residuos estandarizados")  # gráfico 2D de los valores ajustados vs. los residuos estandarizados

abline(h=0) # dibuja la recta en cero
```



¿Cuál será el valor esperado de concentración de ozono con una temperatura de 75 grados?

$\hat{Y} = -125{,}52 + 2{,}098 \cdot 75 = 31{,}83$ ppb

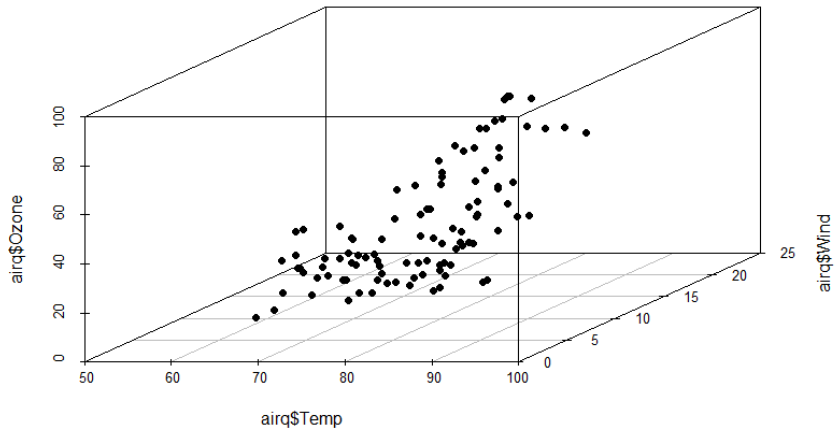
¿Y si fuese de 250?

$\hat{Y} = -125{,}52 + 2{,}098 \cdot 250 = 398{,}98$ ppb

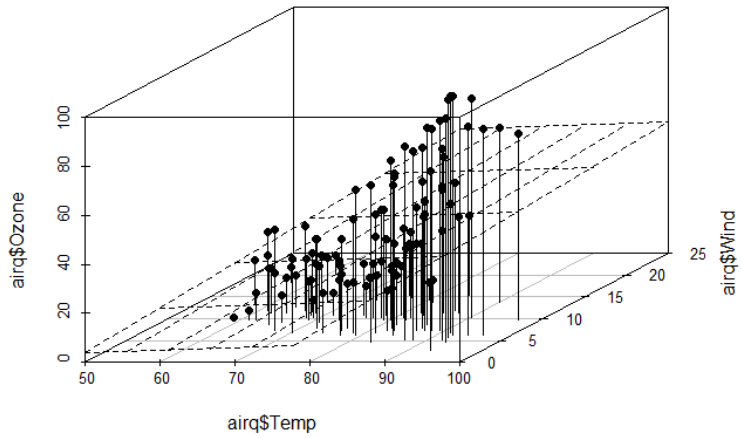
¿Serían fiables dichos pronósticos? ¿E igualmente fiables?

7. **Regresión lineal múltiple.** Ajuste un modelo lineal que permita predecir el nivel de ozono en función de la temperatura y el viento. Represente gráficamente los datos, estime un modelo en el que se consideren estas dos variables regresoras, interprete los coeficientes y analice la bondad de ajuste. ¿Qué puede decir acerca del poder predictivo del modelo?

Gráfico de dispersión 3d:




```
library(scatterplot3d)

scatterplot3d(airq$Temp, airq$Wind, airq$Ozone, pch = 16)
```



Ajuste del modelo:

```
regr_multiple<- lm(Ozone~airq$Temp+airq$Wind, data = airq)

summary(regr_multiple)
```

$$Y = a + bX_1 + cX_2 + e$$

con X_1 la temperatura y X_2 el viento. Además, b y c son los llamados coeficientes de regresión.

$$\widehat{Ozono} = -86,87 + 1,82 \cdot Temp - 1,67 \cdot Viento$$

El valor estimado de a es -86,97 e indica el valor estimado de ozono para una temperatura de 0 grados y una velocidad de viento nula (0 mph). En este caso, no tiene sentido la interpretación ya que el 0 no está dentro del rango de valores de la temperatura ni de viento (aunque sería suficiente con que no estuviera en el rango de una de ellas para no poder utilizar dicha interpretación). El valor estimado de b es 1,82, que indica que se espera que los valores de Ozono aumenten una concentración de 1,82 cuando la temperatura aumente un grado y el viento se mantenga constante. El valor estimado de c es -1,67 e indica que se espera que los valores de concentración de ozono disminuyan 1,67 cuando la velocidad del viento aumente 1 mph, manteniéndose fija la temperatura.

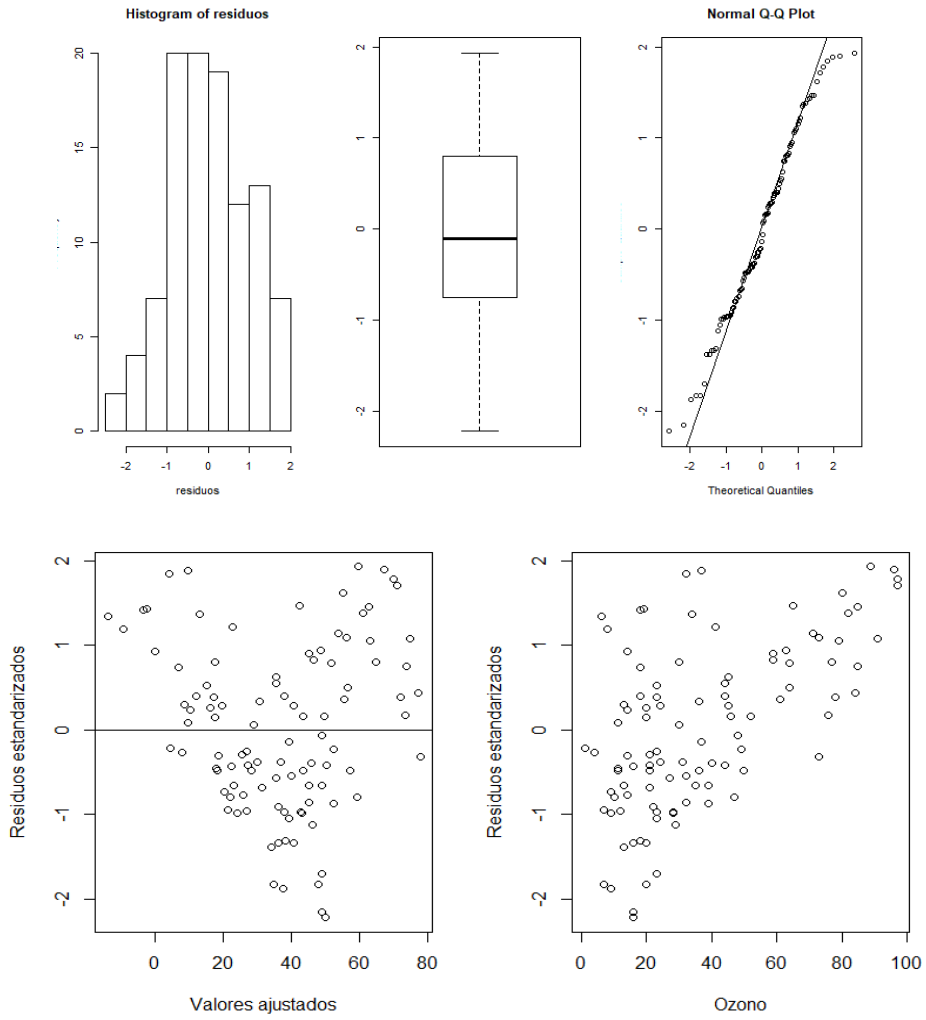
Bondad de ajuste

El coeficiente de determinación R^2_{adj}=0,638 indica que la temperatura y el viento, utilizando el modelo de regresión lineal múltiple, explican el 64% de la variación total de los valores del ozono. Se ha mejorado mínimamente con respecto al modelo de regresión lineal simple, lo que era de esperar por utilizar una nueva variable independiente.

Plano de ajuste

```
s3d <- scatterplot3d(airq$Temp, airq$Wind, airq$Ozone,
type="h",pch=16, main="Temp + Viento + Ozono")
s3d$plane3d(regr_multiple)
```

Poder predictivo



PARTE 2 – MODELOS NO LINEALES

A menudo se supone que la relación que guardan la variable dependiente y las independientes es lineal. En estos casos, se utilizan los modelos de regresión lineal. Sin embargo, aunque las relaciones lineales aparecen de forma frecuente, también es posible considerar otro tipo de relación entre las variables, que se modelizan mediante otros modelos de regresión, como pueden ser el modelo de regresión cuadrático o parabólico, o los modelos de regresión logarítmicos y exponenciales.

El objetivo de la práctica de hoy es explicar la relación entre temperatura y ozono mediante modelos no lineales. Se pretende encontrar el modelo lineal o no lineal que mejor ajuste a los datos. ¿Mejora la bondad de ajuste de los modelos no lineales con respecto al modelo lineal? Para responder a esta pregunta realice las siguientes tareas.

En primer lugar, ajuste los datos a un modelo lineal y valore el poder explicativo del modelo. Represente en un diagrama de dispersión la relación entre temperatura y ozono, añadiendo la recta de ajuste a los datos.

A continuación:

1) Ajuste los datos a un modelo logarítmico y valore el poder explicativo del modelo. Represente en un diagrama de dispersión la relación entre temperatura y ozono, añadiendo el modelo ajustado a los datos.
2) Ajuste los datos a un modelo exponencial y valore el poder explicativo del modelo. Represente en un diagrama de dispersión la relación entre temperatura y ozono, añadiendo el modelo ajustado a los datos.
3) Ajuste los datos a un modelo potencial y valore el poder explicativo del modelo. Represente en un diagrama de dispersión la relación entre temperatura y ozono, añadiendo el modelo ajustado a los datos.
4) Ajuste los datos a un modelo parabólico y valore el poder explicativo del modelo. Represente en un diagrama de dispersión la relación entre temperatura y ozono, añadiendo el modelo ajustado a los datos.

Y complete la siguiente tabla:

Modelo	**Ecuación**	**Modelo linealizado**	**Bondad de ajuste (R^2)**
LINEAL	$Y=A+BX$	-	60,6%
LOGARÍTMICO	$Y=A+B\ln X$	$X'=\ln X$ - $Y=Y$	56,7%
EXPONENCIAL	$Y=AB^x$	$X'=X$ - $Y'=\ln Y$	55,6%
POTENCIAL	$Y=AX^b$	$X'=\ln X$ - $Y'=\ln Y$	54,4%
PARABÓLICO	$Y=A+BX^2+CX$	-	71.8%

#MODELOS NO LINEALES

```
modelo_log <- lm(airq$Ozone~log(airq$Temp))
summary(modelo_log)
plot(log(airq$Temp), airq$Ozone)
abline(modelo_log, col="green", lty=2, lwd=2)

modelo_exp <- lm(log(airq$Ozone)~airq$Temp)
summary(modelo_exp)
plot(airq$Temp, log(airq$Ozone))
abline(modelo_exp, col="green", lty=2, lwd=2)

modelo_pot <- lm(log(airq$Ozone)~log(airq$Temp))
summary(modelo_pot)
plot(log(airq$Temp), log(airq$Ozone))
abline(modelo_pot, col="green", lty=2, lwd=2)

modelo_cuad <- lm(airq$Ozone~airq$Temp+(airq$Temp)^2)
summary(modelo_cuad)
plot(airq$Temp, airq$Ozone)
curve(355.4675-10.67573*x+0.0835*x^2, 55, 99, add=TRUE)
```

6. Probabilidad, distribuciones y simulación

OBJETIVOS
Conocer las principales funciones de cálculo de probabilidades.
Calcular probabilidades de las distribuciones Binomial y Normal.
Evaluación de funciones de densidad.
Generación de números de distribuciones discretas y continuas más frecuentes.
Gráficos de distribuciones.

R dispone de varias funciones asociadas a distintos tipos de distribuciones de probabilidad. Existen muchas distribuciones de probabilidad, pero en esta práctica se trabajará sobre las distribuciones de probabilidad discretas de Poisson y binomial y sobre la distribución normal para ilustrar distribuciones continuas.

Función	Distribución
norm()	Normal
t()	t de Student
unif()	Uniforme
f()	F de Snedecor
exp()	Exponencial
binom()	Binomial
chisq()	Ji-Cuadrado
gamma()	Gamma
pois()	Poisson
beta()	Beta

Todas las funciones de distribuciones en R están generadas para que se comporten de una manera común. Cada una de las distribuciones de probabilidad más conocidas tienen un conjunto de cuatro funciones que nos permiten:

Random: rxxx(n,) => Devuelve una simulación aleatoria de longitud n

Density: dxxx(x,) => Devuelve la función de densidad o de probabilidad

Probability: pxxx(q,) => Devuelve la función de distribución (acumulada)

Quantile: qxxx(p,) => Devuelve los cuantiles de la distribución asociados a una probabilidad q

Realice las siguientes tareas:

BINOMIAL

1. Genere 100 observaciones de una distribución binomial con n = 7 y p = 0.3.

   rbinom(100, 7, 0.3)

2. Calcule la P(4≤X≤7), es decir, la probabilidad de observar 4, 5, 6 o 7 aciertos para una v.a. X~B(10,0,5).

   sum(dbinom(x = 4:7, size = 10, prob = 0.5))

   O:

   pbinom(q = 7, size = 10, prob = 0.5) - pbinom(q = 3, size = 10, prob = 0.5)

3. *Fuente: Anaya.* El 42% de los habitantes de un municipio está a favor de la gestión del alcalde y el resto son contrarios a este. Se toma una muestra de 64 individuos,

   a. Defina la variable aleatoria que cuenta el número de votos que recibe el alcalde e identifique la distribución de probabilidad que sigue la variable aleatoria.

   X = "Nº de habitantes que le votan"

   Esta variable aleatoria tiene una distribución Binomial de parámetros

   $$X \sim B(64, 0.42)$$

   b. ¿Cuál es la probabilidad de que todos voten a favor del alcalde?

      dbinom(64,64, 0.42)

   c. ¿Cuál es la probabilidad de que lo voten como mucho 50 personas? ¿Y la de que lo voten al menos 5 habitantes?

      pbinom (5,64, 0.42)

      $P[X \geq 5] = 1 - P[X \leq 4] = 1 - F(4)$

      1 – pbinom (4,64, 0.42)

   d. ¿Cuál es la probabilidad de que lo voten entre 25 y 35 personas?

      pbinom (c(35, 24),64, 0.42) y se restan

NORMAL

1. Genere 800 números aleatorios procedentes de una distribución normal N(3,2.1) mediante la función rnorm.

```
x<-rnorm(800, mean = 3, sd = 2.1)
hist(x, probability=TRUE)
xx <- seq(from=min(x), to=max(x), length.out=100)
lines(xx, dnorm(xx, mean=3, sd=2.1))
```

2. Implemente la función de densidad y de distribución de N(0,1) como una función de R y calcule la $P(Z \leq 1.7)$:

```
#Funcion de densidad normal
fd.normal<-function(x)
    {
          (1/sqrt(2*pi))*exp(-(x^2/2))
    }
```

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

$$Normal\ estándar: f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (-\infty < x < \infty)$$

```
quantiles <- c(-1.96, 0, 1.96)

fd.normal(quantiles)
#> [1] 0.05844094 0.39894228 0.05844094

dnorm(quantiles, mean = 0, sd = 1)

fd.normal(quantiles) == dnorm(quantiles, mean = 0, sd = 1)

curve(fd.normal(x),xlim=c(-5,5))   #Pintar
```

Función de distribución

```
integrate(fd.normal,
        lower = -Inf,
        upper = 1.7)
es lo mismo que:  pnorm(1.7)

o que:

integrate(dnorm,
        lower = -Inf,
        upper = 1.7)
```

3. Grafique la función de densidad de una Normal Estándar y la función de distribución acumulada.

```
curve(dnorm(x), xlim = c(-5, 5),
ylab = "Densidad", main = "Funcion de Densidad – Normal Estandar")

curve(pnorm(x),
xlim = c(-5, 5),
ylab = "Probabilidad",
main = "Funcion Distribucion – Normal Estandar")
```

4. En una distribución $N(0,1)$, calcular:

a) $P(Z \leq 0.72) = 0.7642$

```
pnorm(0.72,mean = 0, sd = 1)
```

b) $P(Z > 1.24) = 0.1075$

```
pnorm(1.24,mean = 0, sd = 1, lower.tail = F)
1-pnorm(1.24,mean = 0, sd = 1, lower.tail = T)
```

c) $P(0.5 < Z \leq 1.76) = 0.2693$

```
pnorm(1.76)-pnorm(0.5)
```

d) Calcular a para que $P(Z \leq a) = 0.975$:

$$a = 1.96$$

```
qnorm(0.975,mean = 0, sd = 1, lower.tail = TRUE)
qnorm(0.975)
```

5. Para $X \sim N(5,2)$, calcular x tal que $P(X \leq x) = 0.45$

$$x = 4.74$$

```
qnorm(0.45, mean=5, sd=2)
```

6. Suponga que la variable "puntuación de un examen de acceso a la universidad" se ajusta a una distribución normal de media 6.8 y desviación típica 1.13. ¿Cuál es la probabilidad de que los estudiantes obtengan una puntuación de más de 8.7 en el examen?

Aplicamos la función pnorm de la distribución normal con media 6.8 y desviación estándar 1.13. Dado que buscamos el porcentaje de alumnos con una puntuación superior a 8.7, nos interesa la cola superior de la distribución normal.

```
pnorm(8.7, mean=6.8, sd=1.13, lower.tail=FALSE)
```

El porcentaje de estudiantes que obtienen una puntuación de 8.7 o más en el examen de acceso a la universidad es del 4.6%.

7. Halle el 70-ésimo percentil de la distribución normal de parámetros $\mu = 2$, $\sigma = 0.1$.

```
qnorm(0.70, mean = 2, sd = 0.1)
```

8. Se sabe que el nivel de colesterol (en mg/100 ml) en una población se distribuye, aproximadamente, según un modelo de normal con media 205 mg/100 ml y desviación típica 36 mg/100 ml. Se selecciona al azar un individuo de dicha población y se pretende calcular la probabilidad de que:

Se define: X = "Concentración en plomo de un individuo".

$X \rightarrow N(205, 36)$

a. Su nivel de colesterol sea inferior a 190 mg/100 ml. Sol: 0.33

```
pnorm(190, mean = 205, sd = 36)
```

b. Un nivel de colesterol superior a 270 mg/100 ml se considera extremadamente alto, ¿cuál es la probabilidad de que un individuo seleccionado al hacer de esta población esté incluido en esta categoría? Sol: 0.035

```
1 – pnorm(270, mean = 205, sd = 36)
```

c. Su nivel de colesterol se encuentre entre 180 y 220 mg/100 ml. Sol: 0.418

```
pnorm(c(220,180), mean = 205, sd = 36)
pnorm(220, mean = 205, sd = 36)- pnorm(180, mean = 205, sd = 36)
```

d. Determine el nivel de colesterol mínimo del 40% de los individuos con más colesterol. Sol: 214

```
qnorm(0.60, mean = 205, sd = 36)
```

e. Determine la mediana de la distribución. Sol: 205

```
qnorm(0.50, mean = 205, sd = 36)
```

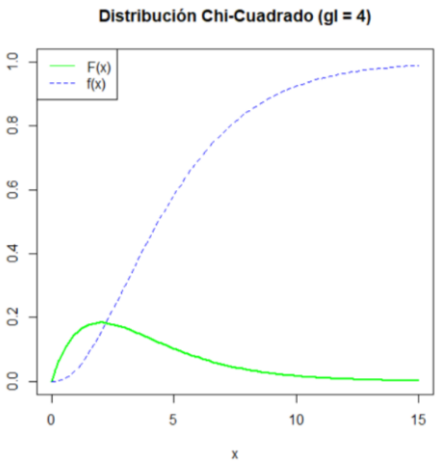
f. Genere una muestra de tamaño 50 para esta distribución normal.

```
rnorm(50, mean = 205, sd = 36)
```

OTRAS DISTRIBUCIONES: χ^2

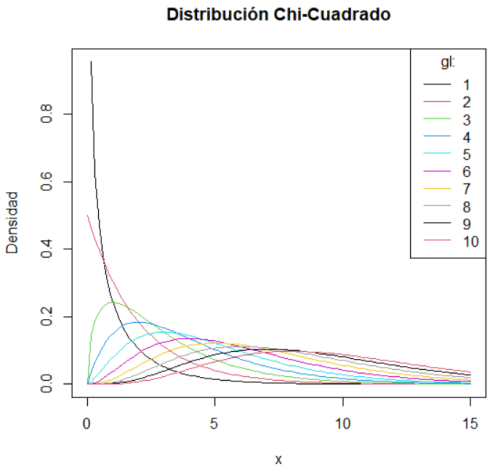
9. Grafique la función de densidad y de distribución de una v.a. que sigue una distribución Chi-Cuadrado, de 5 grados de libertad.

```
# Funcion de densidad
curve(dchisq(x, df = 4),
      xlim = c(0, 15),
      ylim = c(0, 1),
      col = "green",
      ylab = "",
      lwd=2,
      main = "Distribución Chi-Cuadrado (gl = 4)")
# Añadimos la curva de la función de distribución
curve(pchisq(x, df = 4),
      xlim = c(0, 15),
      add = TRUE,
      col = "blue",
      lwd=1.5, lty=2)
# Leyenda
legend("topleft",
       c("F(x)", "f(x)"),
       col = c("green", "blue"),
       lty = c(1, 2))
```



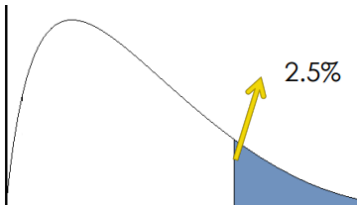
Distintos grados de Libertad:

```
# gl=1
curve(dchisq(x, df = 1),
      xlim = c(0, 15),
      xlab = "x",
      ylab = "Densidad",
      main = "Distribución Chi-Cuadrado")

# gl=2,...,10 usando el bucle for
for (gl in 2:10) {
  curve(dchisq(x, df = gl),
        xlim = c(0, 15),
        add = T,
        col = gl)
}
# añadir leyenda
legend("topright",
       as.character(1:10),
       col = 1:10,
       lty = 1,
       title = "gl:")
```

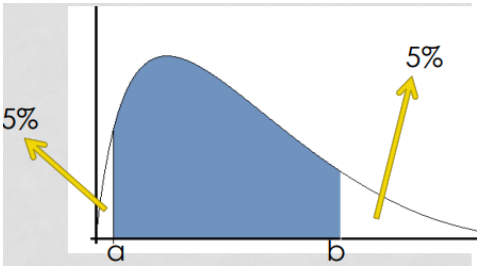


10. Sea X una variable aleatoria que sigue una distribución Chi-Cuadrado con 9 grados de libertad. Calcular: a) ¿Cuál es el valor de la variable que deja a su derecha una probabilidad de un 2.5%? Sol: 19.02

```
qchisq(0.025,df=9, lower.tail=FALSE)

qchisq(0.975,df=9, lower.tail=T)
```



11. Sea X una variable aleatoria que sigue una distribución Chi-Cuadrado con 6 grados de libertad. Calcular: a) ¿Cuáles son los valores "a" y "b" de la variable que deja en sus extremos una probabilidad de un 10%? Sol: 1.635, 12.59

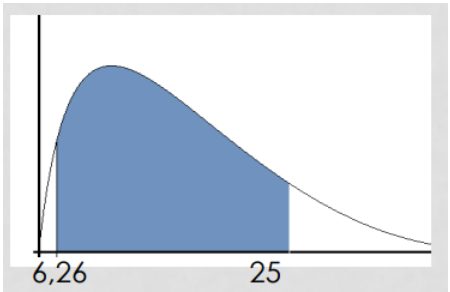



```
qchisq(0.05,df=6)

qchisq(0.95,df=6)
```

12. Calcula el percentil 95 de la distribución Chi-Cuadrado con 9 grados de libertad.

```
qchisq(.95, df=9)        # 9 degrees of freedom
```

13. Sea X una variable aleatoria que sigue una distribución Chi-Cuadrado con 15 grados de libertad, calcula la probabilidad que equivaldría a la siguiente área en azul: Sol: 0.975-0.05=0.97




```
pchisq(25,df=15)-pchisq(6.26,df=15)
```

OTRAS DISTRIBUCIONES: T-Student

14. Grafique la función de densidad de una distribución T-Student, de diferentes grados de libertad.

```
# Normal estandar
curve(dnorm(x),
      xlim = c(-4, 4),
      xlab = "x",
      lty = 2,
      lwd=2,
      ylab = "Densidad",
      main = "FD Normal")


#T-Student gl=2,3,4
for (gl in 2:4) {
  curve(dt(x, df = gl),
        xlim = c(-4, 4),
        add = T,
        lwd=2,
        col = gl)
}
# gl=25
curve(dt(x, df = 25),
      xlim = c(-4, 4),
      col = "orange",
      lwd=2,
      add = T)
# leyenda:
legend("topright",
       c("N(0,1)", "M=2","M=3","M=4", "M=25"),
       col = c(1:4,"orange"),
       lty = c(2, 1, 1, 1, 1),
       lwd=2)
```

15. Sea X una variable aleatoria que sigue una distribución t de Student con n=15 grados de libertad. Calcular el valor "t" tal que P(X>t)=0.025. Sol: -2.1315

```
qt(0.025, df=15,lower.tail = F)
```

16. Encuentra los percentiles 5 y 95 de la distribución t de Student con 4 grados de libertad. Sol: -2.132 y 2.132

```
qt(c(.05, .95), df=4)
```

17. Sea X una variable aleatoria que sigue una distribución T-Student con 10 grados de libertad, calcula: P(X > 2.764) Sol: 0.01

```
pt(2.764, df=10, lower.tail=F)
```

OTRAS DISTRIBUCIONES: F de Snedecor

18. Grafique la función de densidad de una distribución F de Snedecor de grados de libertad 8 y 12.

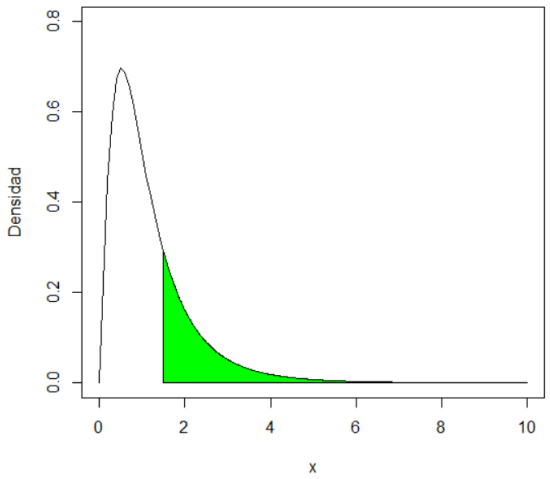
```
# F_{8, 12}
curve(df(x,8 ,12),
      ylim = c(0, 1),
      xlim = c(0, 15),
      ylab = "Densidad",
      main = "Función de Densidad")
```

19. Sea X una variable aleatoria que sigue una distribución F-Snedecor con n=2 y m=5 grados de libertad, calcular el valor de la variable que deja a su derecha una probabilidad del 5%. Sol: $F_{2,5}=5.79$

```
qf(0.05,df1=2,df2=5, lower.tail = F)
```

20. Calcule la P(Y>1.5), donde Y es una v.a. que sigue la distribución F con 5 y 12 grados de libertad respectivamente y represente gráficamente dicha probabilidad. Sol: 0.2611

```
pf(1.5, df1 = 5, df2 = 12, lower.tail = F)

# Definir coordenadas:
coord.x <- c(1.5, seq(1.5, 10, 0.01), 10)
coord.y <- c(0, df(seq(1.5, 10, 0.01), 5, 12), 0)

curve(df(x ,5 ,12),
      ylim = c(0, 0.8),
      xlim = c(0, 10),
      ylab = "Densidad")

polygon(coord.x, coord.y, col = "green")
```



21. Encuentre el percentil 83 de la distribución F con (3,9) grados de libertad. Sol: 2.1037

```
qf(0.83, df1=3, df2=9)
```

7. Intervalos de confianza

OBJETIVOS
Cálculo del intervalo de confianza para la media de una población con varianza conocida.
Cálculo del intervalo de confianza para la media de una población con varianza desconocida.
Calcular e interpretar intervalos de confianza para la diferencia de medias en dos poblaciones normales independientes con varianzas desconocidas y suponiéndolas iguales.
Calcular e interpretar intervalos de confianza para la diferencia de medias en dos poblaciones normales independientes con varianzas desconocidas y suponiéndolas diferentes.
Graficar intervalos de confianza.

R no dispone de funciones implementadas para el cálculo de intervalos de confianza, pero si funciones diseñadas para otro tipo de estudios que los incluyen. Además, la obtención del intervalo de confianza manualmente es muy sencilla.

Cargue la base de datos “Datos flores” y realice las siguientes tareas:

1. Estime el intervalo de confianza para la EVA media poblacional (suponiendo que sigue la distribución Normal en la población) a un nivel de confianza del 95%:

```
media <- mean(flores$EVA)
cuasi= sd(flores$EVA)
gl<-length(flores$EVA)-1
cuantil<-qt((1 - alpha/2), gl, lower.tail = T)
cuantil
media - cuantil * cuasi / sqrt(n)
media + cuantil * cuasi / sqrt(n)
t.test(flores$EVA) #AI 95%
```

1.0. ¿Cuál es la estimación puntual del valor medio de EVA?

media <- mean(flores$EVA)

1.1. ¿Qué puede concluirse en relación con la media de toda la población?

2. Calcule el intervalo de confianza para la EVA media poblacional ahora con un nivel de confianza del 99%. Antes de obtener este intervalo, ¿será más o menos amplio que el calculado antes con una confianza del 95%? ¿y si se considerase una confianza del 90%?

t.test(flor_A, flor_B, conf.level = 0.99, var.equal = TRUE) **#2. Al 99%:**

t.test(flor_A, flor_B, conf.level = 0.90, var.equal = TRUE) **#Al 90%:**

3. Calcule los intervalos de confianza para la media poblacional de “Agua en ml” según el color final de las flores (blanco, amarillo) a un nivel de confianza del 90%. Compare los resultados obtenidos en cada grupo.

flor_A<-flores$Agua.ml[which(flores$Color=="Blanco")]

flor_B<-flores$Agua.ml[which(flores$Color=="Amarillo")]

t.test(flor_A, conf.level = 0.90, var.equal = TRUE)

t.test(flor_B, conf.level = 0.90, var.equal = TRUE)

3.0. Sin calcular los intervalos de confianza, ¿la conclusión sería la misma al 80% de confianza? ¿Y al 95%?

t.test(flor_A, conf.level = 0.80, var.equal = TRUE)

t.test(flor_B, conf.level = 0.80, var.equal = TRUE)

3.1. Si queremos ver qué ocurre en relación con los dos colores: ¿qué puede afirmarse a la vista de los resultados obtenidos hallando los intervalos al 95%?

No hay diferencias significativas entre ambos grupos

4. Grafique los intervalos correspondientes al análisis anterior.

library(gplots)

plotmeans(Agua.ml ~ Color, data = flores, p=0.9, ci.label=T, connect=F)

plotmeans(Agua.ml ~ Color, data = flores, frame=F, p=0.8, ci.label=T, connect=F)

5. Calcule los intervalos de confianza al 90% para la puntuación media poblacional de la escala visual analógica EVA para las flores según si recibieron Aspirina o Ibuprofeno. Interprete todos los intervalos.

```
flor_A<-flores$EVA[which(flores$Tratamiento=="Aspirina")]
flor_B<-flores$EVA[which(flores$Tratamiento=="Ibuprofeno")]
t.test(flor_A, conf.level = 0.90, var.equal = TRUE)
t.test(flor_B, conf.level = 0.90, var.equal = TRUE)
```

5.0. ¿Aseguraría la existencia de diferencias en el estado a nivel poblacional entre las flores tratadas con "aspirina" o "ibuprofeno"?

```
flores1<-
flores[c(which(flores$Tratamiento=="Aspirina"),which(flores$Tratamiento=="Ibu
profeno")),]
library(gplots)
plotmeans(EVA ~ Tratamiento, data = flores1,p=0.99, ci.label=T, connect=F)
```

6. Calcule los intervalos de confianza al 90% para la puntuación media de la escala visual analógica EVA para las flores según todos los tipos de tratamiento que recibieron. Interprete todos los intervalos y explique porque tienen diferentes amplitudes.

```
aspirina<-flores$EVA[which(flores$Tratamiento=="Aspirina")]
ibuprofeno<-flores$EVA[which(flores$Tratamiento=="Ibuprofeno")]
paracetamol<-flores$EVA[which(flores$Tratamiento=" Paracetamol")]
control<-flores$EVA[which(flores$Tratamiento=" Control")]

t.test(aspirina, conf.level = 0.90, var.equal = TRUE)
t.test(ibuprofeno, conf.level = 0.90, var.equal = TRUE)
t.test(paracetamol, conf.level = 0.90, var.equal = TRUE)
t.test(control, conf.level = 0.90, var.equal = TRUE)
```

¿Considera que todos los grupos tienen la misma media poblacional?

7. Represente los intervalos de confianza del apartado anterior en un gráfico.

```
plotmeans(EVA ~ Tratamiento, data = flores,p=0.9, ci.label=T, connect=F)
```

REPASO - AUTOEVALUACIÓN

Trabajo con los datos que encontrará en el fichero "Pulmon.xlsx". La base de datos contiene información de 149 sujetos con cáncer de pulmón avanzado anonimizado. Las variables registradas en la base de datos son:

- Tiempo: Tiempo de supervivencia en días
- Estado: estado del paciente (1=vivo, 2=muerto)
- Edad: Edad en años
- Sexo: (1=Hombre, 2=Mujer)
- RendimientoECOG: Puntuación de rendimiento ECOG calificada por el médico (0= asintomático, 1= sintomático pero completamente ambulatorio, 2= en cama <50% del día, 3= en cama > 50% del día pero sin estar en cama, 4 = en cama)
- Karnofsky.medico: Puntuación de rendimiento de Karnofsky (malo=0-bueno=100) calificada por el médico
- Karnofsky.paciente: Puntuación de rendimiento de Karnofsky calificada por el paciente
- Calorias: Calorías consumidas en las comidas
- Calorias2: Calorías consumidas en las comidas con dieta
- PerdidaPeso: Pérdida de peso en los últimos seis meses

Las puntuaciones de rendimiento califican lo bien que el paciente puede realizar las actividades diarias habituales.

Los datos proceden del conjunto de datos *lung* (o *cancer*) incluido en el paquete *survival* de R (Therneau, 2024). Con fines didácticos, hemos agregado dos nuevas variables que no corresponden a la base de datos original: talla (m), peso (kg)

Realice las siguientes tareas:

1. Recodificar la variable "edad" de diez en diez años en una variable que llamaremos "edadREC", en la que se distribuirá la edad de los pacientes en los siguientes grupos: 1=30-45; 2=46-60 ; 3=61-75 ; 4= >75 .
2. Obtener la distribución de frecuencias y representar un diagrama de sectores para las variables "edadREC" y "estado".
3. Obtener la variable Talla expresada en centímetros, que llamaremos"TALLA_CM".
4. Calcule una nueva variable de Índice de Masa Corporal, que llamaremos "IMC", y que se calcula a partir de:

$$IMC = \frac{peso}{talla^2}$$

5. Realizar un gráfico apropiado para las variables "calorías", "Calorias2" y "Edad" para estudiar su asimetría. ¿Qué se podría afirmar respecto de la normalidad de estas distribuciones?

   Realizar el correspondiente a la talla empezando en 1,35 metros y con una amplitud de 5 centímetros de cada intervalo.

CONTINÚE ANALIZANDO SOLO LOS PACIENTES QUE HAN FALLECIDO

6. Realizar un estudio descriptivo y exploratorio de las variables siguientes:

	Media	Mínimo	Máximo	Desv.Tip.	Error estándar	CV	Q1	Q2	Q3
Tiempo Supervivencia									
Karnofsky.medico									
Karnofsky.paciente									
Calorías									
Calorías2									
Imc									

Interprete los resultados obtenidos. ¿Qué variables muestran una mayor dispersión? De entre las variables "Calorías" y "Calorías2", ¿cuál es la que está menos dispersa?

A PARTIR DE ESTE PUNTO VAMOS A TRABAJAR CON TODA LA MUESTRA COMPLETA

7. Realice un Box-plot de "calorías", "Calorias2" y compare los resultados con los del apartado 4. Comente los resultados.

8. ¿Cuál es el tiempo medio de supervivencia de los pacientes vivos y de los fallecidos? ¿Observa alguna diferencia?

9. Realice el análisis que considere más adecuado para saber si existen diferencias en la puntuación de rendimiento de Karnofsky calificada por el paciente según su RendimientoECOG. Represente gráficamente la puntuación de rendimiento de Karnofsky según los distintos grupos de rendimiento ECOG.

CONTINÚE ANALIZANDO SOLO LOS PACIENTES QUE HAN FALLECIDO

10. Realice los pasos necesarios para ajustar un modelo de regresión lineal en el que se explique la relación entre las variables Tiempo de Supervivencia y Calorías. Compare los resultados entre sí.

11. A la vista de los resultados obtenidos en el apartado anterior, ¿qué cambios en el modelo propone para estudiar el efecto de las calorías en el tiempo de supervivencia de los pacientes fallecidos? Realice dichas modificaciones, interprete y compare los resultados obtenidos (utilice el modelo parabólico o modelos no lineales)

AUTOEVALUACIÓN

A partir de los resultados obtenidos en la práctica y los gráficos generados en R, responda las siguientes afirmaciones marcando *Verdadero* o *Falso*.

Pregunta 1. Observando el diagrama de sectores, la proporción de pacientes entre los 46 y 60 años es similar a la de pacientes entre 61 y 75 años.
☐ Verdadero ☐ Falso

Pregunta 2. De acuerdo con la tabla de frecuencias, el grupo de edad mayoritario es el de pacientes entre 61 y 75 años.
☐ Verdadero ☐ Falso

Pregunta 3. El 33,8 % de los pacientes muestreados tiene 60 años o menos.
☐ Verdadero ☐ Falso

Pregunta 4. Observando el histograma de la variable *Calorías*, se observa que hay algún paciente que consume más de 2000 calorías.
☐ Verdadero ☐ Falso

Pregunta 5. Hay 102 pacientes en la muestra que han fallecido.
☐ Verdadero ☐ Falso

Pregunta 6. El índice de Karnofsky (médico) medio de los pacientes supera los 80 puntos, con una dispersión en la muestra de 12,87 puntos.
☐ Verdadero ☐ Falso

Pregunta 7. Si comparamos las dispersiones de *Calorías* y *Calorías2* en el grupo de pacientes fallecidos, se observa que *Calorías2* es menos dispersa.
☐ Verdadero ☐ Falso

Pregunta 8. El tiempo medio de supervivencia es mayor para los pacientes vivos que para los fallecidos.
☐ Verdadero ☐ Falso

Pregunta 9. Los pacientes fallecidos fueron mayoritariamente de sexo femenino, mientras que los vivos fueron mayoritariamente de sexo masculino.
☐ Verdadero ☐ Falso

Pregunta 10. 19,14 kg/m^2 es el IMC que supera el 75 % de los pacientes fallecidos de la muestra.
☐ Verdadero ☐ Falso

Pregunta 11. A la vista del boxplot de las *Calorías*, el rango intercuartílico está comprendido aproximadamente entre 1000 y 1200 calorías.
☐ Verdadero ☐ Falso

Pregunta 12. La relación entre el tiempo de supervivencia y las calorías consumidas en el grupo de pacientes fallecidos es directa.
☐ Verdadero ☐ Falso

Pregunta 13. A la vista del modelo de regresión lineal, la bondad de ajuste es baja, con un coeficiente de determinación del 20,2%.
☐ Verdadero ☐ Falso

Pregunta 14. Analizando el modelo lineal, se observa que por cada caloría que aumenta un paciente, su tiempo de supervivencia disminuye aproximadamente 0,3 días.
☐ Verdadero ☐ Falso

Pregunta 15. Analizando el gráfico de residuos, se observa que el poder predictivo del modelo no es bueno.
☐ Verdadero ☐ Falso


SOLUCIONES

1	F	**6**	V	**11**	F
2	V	**7**	V	**12**	V
3	F	**8**	V	**13**	F
4	V	**9**	V	**14**	F
5	V	**10**	F	**15**	V

BIBLIOGRAFÍA Y RECURSOS RECOMENDADOS

Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

Hillebrand, J., & Nierhoff, M. H. (2015). *Mastering RStudio–Develop, Communicate, and Collaborate with R*. Packt Publishing Ltd.

Kronthaler, F., & Zöllner, S. (2021). *Data Analysis with RStudio*. Springer: Berlin/Heidelberg, Germany.

Lander, J. P. (2014). *R for everyone: Advanced analytics and graphics*. Pearson Education.

Verzani, J. (2018). *Using R for introductory statistics*. Chapman and Hall/CRC.

Coursera: Cursos de RStudio

Owen, W.J. (2006). The R Guide. http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf

Paradis, E. (2003). R para Principiantes: http://cran.r-project.org/doc/contrib/rdebuts_es.pdf

Terry M. Therneau, Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York. ISBN 0-387-98784-3.

Therneau T (2024). *A Package for Survival Analysis in R*. R package version 3.8-3, <https://CRAN.R-project.org/package=survival>.

Short, T. (2004). R Reference Card: http://cran.r-project.org/doc/contrib/Short-refcard.pdf


Nota:
Este material se ha utilizado como apoyo en la asignatura *Estadística* del Grado en Matemáticas y la Doble Titulación de Grado en Física y en Matemáticas y de la Universidad de Salamanca.