



# Módulo 4: Introducción al Machine Learning con R

## Parte II: Introducción a la Estadística Descriptiva con R

Autor: Raquel Dormido Canto

Actualizado Enero 2023

3. EXPLORANDO Y ANALIZANDO DATOS .....	3
3.1. Introducción.....	3
3.2. Estadística cuantitativa y cualitativa.....	3
3.2.1. Variables cuantitativas .....	4
3.2.2. Variables cualitativas.....	5
3.3. Tablas de frecuencia .....	5
3.3.1. Tablas de frecuencias de variables cualitativas.....	6
3.3.2. Tablas de frecuencias de variables cuantitativas .....	8
3.3.2.1. VARIABLES DISCRETAS .....	8
3.3.2.2. VARIABLES CONTINUAS .....	9
3.4. Medidas de una variable estadística .....	9
3.4.1. Medidas de posición.....	10
3.4.1.1. MEDIDAS DE TENDENCIA CENTRAL .....	10
3.4.1.2. MEDIDAS DE TENDENCIA NO CENTRAL .....	11
3.4.2. Medidas de dispersión.....	12
3.4.3. Medidas de forma .....	13
3.4.4. Funciones resumen.....	15
3.5. Descripción gráfica de datos en R.....	15
3.5.1. Gráficos básicos .....	15
3.5.1.1. FUNCIÓN PLOT: plot() .....	16
3.5.1.2. ALGUNOS PROCEDIMIENTOS DE BAJO NIVEL .....	18
3.5.1.3. PRIMERAS FUNCIONES GRÁFICAS INTERACTIVAS .....	19
3.5.2. Gráficos para variables cualitativas o cuantitativas discretas .....	21
3.5.2.1. GRÁFICOS DE BARRAS.....	21
3.5.2.2. GRÁFICOS DE SECTORES.....	23
3.5.2.3. POLÍGONOS DE FRECUENCIAS Y GRÁFICOS DE PUNTOS .....	25
3.5.3. Gráficos para variables cuantitativas continuas.....	27
3.5.3.1. HISTOGRAMAS .....	28
3.5.3.2. BOXPLOTS O GRÁFICO DE CAJA Y BIGOTES.....	30
3.5.3.3. DIAGRAMA DE TALLOS Y HOJAS.....	32
3.5.3.4. DIAGRAMAS DE DISPERSIÓN .....	33
3.6. Exportando gráficos en RStudio.....	33

## 3. EXPLORANDO Y ANALIZANDO DATOS

### 3.1. Introducción

Una vez que ya tenemos nuestros datos en R nos podemos preguntar ¿y ahora qué?. Deberíamos ser capaces de realizar un análisis de ellos. Los datos obtenidos cuando realizamos un experimento presentan variabilidad. Pensemos, por ejemplo, en cómo varía la cantidad de lluvia recogida en un día en una determinada zona, o el peso de un bebé al nacer, o la altura de una planta sometida a dos tipos de abonos distintos, etc. La estadística, es la disciplina que se ha desarrollado en respuesta a los experimentadores cuyos datos presentan variabilidad. Los conceptos y métodos estadísticos nos permiten describir la variabilidad, planificar la investigación teniéndola en cuenta y analizar los datos para extraer el máximo de información de ellos, así como determinar la fiabilidad de las conclusiones que podamos obtener a partir de esos datos. Y estos son los objetivos últimos de la estadística descriptiva, denominados normalmente como análisis descriptivo o exploratorio de datos. Un análisis exploratorio de datos engloba un conjunto de técnicas que permiten comprender de manera rápida la naturaleza de una colección de datos. Dicho análisis se basa principalmente en dos aspectos: 1) se calculan medidas que describen las características más importantes de los datos (estadísticas de resumen, tendencia central, dispersión, forma,...) y, 2) se realizan representaciones gráficas (técnicas de visualización de datos como gráficos de barras, sectores o histogramas, por ejemplo). Como vamos a ver, R permite implementar múltiples técnicas estadísticas que podemos utilizar para analizar los datos.

### 3.2. Estadística cuantitativa y cualitativa

Como hemos comentado en la introducción, la estadística descriptiva de datos analiza series de datos (por ejemplo, edad de una población, altura de los estudiantes, temperatura de los meses de verano, etc.) y trata de extraer conclusiones sobre el comportamiento de estas variables.

Por estadística cuantitativa entendemos aquellas propiedades que son medibles. Estadística cualitativa hace referencia a propiedades no medibles.

Vamos a definir algunos términos comunes en estadística básica: individuo, población y muestra.

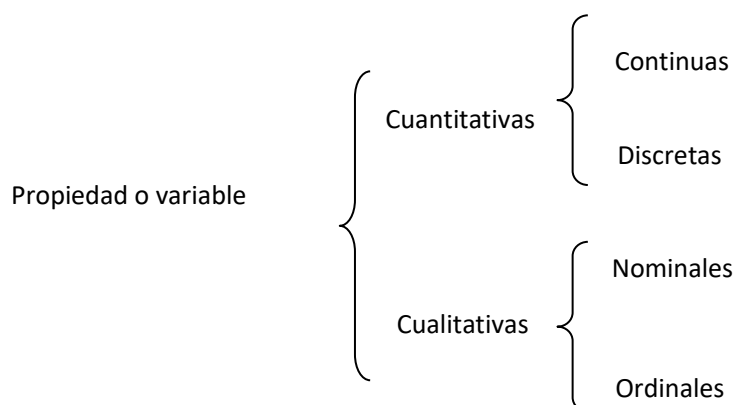
**Individuo:** cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si estudiamos el precio de una vivienda, cada vivienda es un individuo. Cada individuo u **observación** serían las filas de las tablas. Las columnas de las tablas son las características de cada individuo y se denominan variables.

Tanto la altura de los niños como el precio de cada vivienda (características de interés) serían variables.

**Población:** conjunto de datos de todos los individuos (personas, objetos, animales, etc.) que porten información sobre el fenómeno que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad. El número de individuos define el tamaño de la población.

**Muestra:** subconjunto que seleccionamos de la población. Es el conjunto de valores de la variable que observamos obtenidos de manera homogénea. Así, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

La manera de describir la muestra (nuestros datos) depende del tipo de atributo que tengan las variables. Los distintos tipos posibles son los siguientes:



### 3.2.1. Variables cuantitativas

Son aquellas que se pueden medir. Tienen un carácter intrínsecamente numérico. Estas variables estadísticas pueden ser **discretas** o **continuas**.

Las **variables discretas** sólo pueden tomar una cantidad finita de valores enteros, los valores posibles de estas variables son aislados.

**Ejemplos de variables estadísticas cuantitativas discretas:** El número de hermanos: pueden ser 1, 2, 3 ..., pero nunca podrá ser 3,45; el número de hijos; el número de empleados de una fábrica o el número de goles marcados por un equipo de fútbol en la liga.

Las **variables continuas** pueden tomar cualquier valor real (infinitos) dentro de un intervalo.

**Ejemplos de variables estadísticas cuantitativas continuas:** Velocidad de un vehículo (puede ser 20; 54,2; 100; ... km/h); temperaturas registradas en un observatorio cada hora; peso en Kg de los recién nacidos en un día en España.

### 3.2.2. Variables cualitativas

Las variables **cualitativas**, también llamadas **categóricas**, no se pueden medir numéricamente.

**Ejemplos de variables estadísticas cualitativas:** Color de los ojos; bondad de una persona; profesión de una persona.

Determinan modalidades. Por ejemplo, las modalidades de la variable profesión pueden ser: arquitecto, albañil, médico,... etc.

Podemos distinguir entre variables cualitativas nominales y ordinales. Las nominales no presentan orden entre sus valores, por ejemplo, el sexo. Las ordinales tienen valores ordenados, por ejemplo, el nivel de estudios (inicial, medio y avanzado).

## 3.3. Tablas de frecuencia

Vamos a definir en primer lugar lo que entendemos por frecuencia absoluta, frecuencia acumulada y frecuencia relativa. A continuación, veremos cómo calcular con R estos valores. Si tenemos todos los datos almacenados en una tabla en orden creciente o decreciente las distintas frecuencias se definen de esta manera:

**Frecuencia absoluta  $f_i$ .** La frecuencia absoluta  $f_i$  de un valor  $x_i$  es el número de veces que se repite dicho valor.

**Frecuencia absoluta acumulada  $F_i$ .** La frecuencia absoluta acumulada  $F_i$  de un valor  $x_i$  es la suma de las frecuencias absolutas de todos los valores anteriores a  $x_i$  más la frecuencia absoluta de  $x_i$ , esto es,  $F_i = f_1 + f_2 + \dots + f_i$

**Frecuencia relativa  $h_i$ .** La frecuencia relativa  $h_i$  de un valor  $x_i$  se define como el cociente entre la frecuencia absoluta de  $x_i$  y el número total de datos  $N$  que intervienen en la distribución, esto es,  $h_i = x_i/N$

**Frecuencia relativa acumulada  $H_i$ .** La frecuencia relativa acumulada  $H_i$  de un valor  $x_i$  es la suma de las frecuencias relativas de todos los valores anteriores a  $x_i$  más la frecuencia relativa de  $x_i$ , esto es,  $H_i = h_1 + h_2 + \dots + h_i$

Para crear tablas de frecuencia en R se emplea la función `table` o la función `prop.table`, dependiendo de si la tabla muestra las frecuencias absolutas o las frecuencias relativas. La sintaxis de estas órdenes es la siguiente:

```
> table(x) # para frecuencias absolutas  
> prop.table(tab) # para las frecuencias relativas
```

La principal diferencia entre estas dos funciones reside en el tipo de los argumentos que necesita cada una.

`table` construye la tabla de frecuencias absolutas a partir de la variable  $x$  que recibe como argumento

`prop.table` recibe como argumento una tabla o una matriz `tab` que representa una tabla de frecuencias absolutas, y a partir de ella construye la tabla de frecuencias relativas asociada. Es decir, `prop.table` recibe como argumento el resultado que devuelve la función `table`.

### 3.3.1. Tablas de frecuencias de variables cualitativas

Para aprender a calcular las tablas de frecuencias con R vamos a utilizar un ejemplo. Vamos a utilizar el fichero **Datos personas-frecuencias.txt** que se encuentra en la documentación del curso.

En primer lugar, importamos el fichero (desde **Environment->Import Dataset**) y realizamos el `attach` correspondiente:

```
> Datos.personas.frecuencias <- read.delim("C:/Datos personas-frecu  
encias.txt")  
> View(Datos.personas.frecuencias)  
> attach(Datos.personas.frecuencias)
```

Se cargará en nuestro **Environment** la variable `Datos.personas.frecuencias` compuesta por 22 observaciones de 6 variables. Una visualización parcial de los datos se muestra en la Figura 3.3-1.

	Raza	Edad	Peso	Altura	Tension	Hermanos
1	Blanca	24	58.5	156	120	5
2	Negra	26	62.2	175	98	4
3	Blanca	62	61.4	165	76	5
4	Blanca	31	67.5	169	84	2
5	Asiatica	34	71.3	171	68	1
6	Negra	65	69.1	150	80	0
7	Negra	76	NA	190	95	6
8	Blanca	32	56.7	156	86	3
9	Blanca	12	43.4	130	76	7
10	Blanca	56	82.2	178	89	4
11	Asiatica	45	56.7	145	78	3
12	Negra	32	43.7	167	87	1
13	Negra	65	67.9	175	86	2
14	Asiatica	76	76.3	158	67	5

Showing 1 to 15 of 22 entries

Figura 3.3-1: Visualización parcial de Datos.personas.frecuencias.

Para obtener las **frecuencias absolutas** de la variable Raza utilizamos la función `table`, como sigue:

```
> tabla_raza<-table(Datos.personas.frecuencias$Raza)
> tabla_raza
```

```
Asiatica    Blanca    Negra
         5         9         8
```

Las **frecuencias relativas** de cada una de las razas para la muestra que tenemos la calculamos de la siguiente manera:

```
> prop.table(tabla_raza)

Asiatica    Blanca    Negra
0.2272727 0.4090909 0.3636364
```

Atendiendo a la definición que hemos dado anteriormente, y dado que el número de individuos es de 22, otra manera de calcular la frecuencia relativa sería la siguiente:

```
frel<-tabla_raza/22
> frel

Asiatica    Blanca    Negra
0.2272727 0.4090909 0.3636364
```

### 3.3.2. Tablas de frecuencias de variables cuantitativas

#### 3.3.2.1. VARIABLES DISCRETAS

En primer lugar, empezamos con las variables cuantitativas discretas. La variable `Hermanos` es una variable de este tipo.

Las **frecuencias absolutas** de la variable `Hermanos` se calcula utilizando la función `table`:

```
> tabla_hermanos<-table(Hermanos)
> tabla_hermanos
Hermanos
0 1 2 3 4 5 6 7
1 5 4 5 2 3 1 1
```

Las **frecuencias relativas** las podemos obtener utilizando `prop.table` o simplemente dividiendo la frecuencia absoluta por el número de individuos de la muestra, 22:

```
> prop.table(tabla_hermanos)
Hermanos
0      1      2      3      4      5
0.04545455 0.22727273 0.18181818 0.22727273 0.09090909 0.13636364
6      7
0.04545455 0.04545455
```

```
> frel<-tabla_hermanos/22
> frel
Hermanos
0      1      2      3      4      5
0.04545455 0.22727273 0.18181818 0.22727273 0.09090909 0.13636364
6      7
0.04545455 0.04545455
```

Además, ahora podemos obtener la **frecuencia absoluta y relativa acumuladas**. Para ello, podemos hacer lo siguiente:

```
> fabsacum<-as.table(cumsum(fabs))
> fabsacum
0 1 2 3 4 5 6 7
1 6 10 15 17 20 21 22

> frelacum<-as.table(cumsum(frel))
> frelacum
0      1      2      3      4      5
0.04545455 0.27272727 0.45454545 0.68181818 0.77272727 0.90909091
6      7
0.95454545 1.00000000
```



### 3.3.2.2. VARIABLES CONTINUAS

Para variables cuantitativas continuas las cosas se complican algo debido a que tenemos que determinar, en función del número de observaciones, si realizar un estudio de forma individual o cómo agrupar los valores de las variables en intervalos. La variable `Peso` es una variable de este tipo.

Supongamos que decidimos considerar un total de 7 intervalos. Las **frecuencias absolutas** y **relativas** de esta variable se pueden calcular, utilizando la función `table`, de la siguiente manera:

```
> fabs<-table(cut(Peso,breaks=7))
> fabs
```

(43.4,48.9]	(48.9,54.5]	(54.5,60]	(60,65.6]	(65.6,71.1]
4	0	5	4	4
(71.1,76.7]	(76.7,82.2]			
3	1			

```
frel<-table(cut(Peso,breaks=7))/22
> frel
```

(43.4,48.9]	(48.9,54.5]	(54.5,60]	(60,65.6]	(65.6,71.1]
0.18181818	0.00000000	0.22727273	0.18181818	0.18181818
(71.1,76.7]	(76.7,82.2]			
0.13636364	0.04545455			

Mientras que las **frecuencias acumuladas** son:

```
> fabsacum
```

(43.4,48.9]	(48.9,54.5]	(54.5,60]	(60,65.6]	(65.6,71.1]
4	4	9	13	17
(71.1,76.7]	(76.7,82.2]			
20	21			

```
> frelacum<-cumsum(frel)
> frelacum
```

(43.4,48.9]	(48.9,54.5]	(54.5,60]	(60,65.6]	(65.6,71.1]
0.1818182	0.1818182	0.4090909	0.5909091	0.7727273
(71.1,76.7]	(76.7,82.2]			
0.9090909	0.9545455			

### 3.4. Medidas de una variable estadística

En la sección anterior hemos aprendido a calcular las tablas de frecuencias de nuestros datos, lo que nos da una idea de la composición de la población que queremos estudiar. En esta sección vamos a resumir los datos recogidos en una tabla estadística en unos valores o medidas numéricas que representen el conjunto de datos. Estas medidas (de posición, dispersión y forma), que proporcionan

información sobre puntos importantes de la distribución, se revisan en el módulo del máster dedicado a estadística, por lo que no nos centraremos en sus definiciones matemáticas. Veremos cómo calcular los distintos valores utilizando R. En la sección 3.5 completaremos la información que nos han proporcionado las tablas de frecuencias y las medidas descritas en esta sección con representaciones gráficas. Dichas representaciones nos darán una representación visual de las variables que queramos estudiar.

En esta sección vamos a utilizar en los ejemplos el fichero **Datos personas-frecuencias.txt**.

### 3.4.1. Medidas de posición

#### 3.4.1.1. MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central se utilizan cuando el interés está en localizar el centro de la distribución. Estas medidas tratan de resumir los valores observados en un único valor asociado al localizado en el centro. Las medidas de tendencia central más habituales son: la **media**, la **mediana** y la **moda**.

En R la **media** y la **mediana** se calculan, respectivamente, con las funciones:

```
mean (x, na.rm = FALSE)
```

```
median (x, na.rm = FALSE)
```

donde  $x$  es el vector con los valores de la variable y `na.rm` un argumento lógico que indica si hay que eliminar los valores faltantes (NA) del conjunto de datos.

Cuando una función de R encuentra algún NA entre los valores de las observaciones que trata de analizar devuelve como resultado NA, indicando así que los cálculos no se han podido realizar. No obstante, asignando el valor TRUE al argumento `na.rm` se pueden eliminar los valores faltantes y obtener así un valor para la media o la mediana, basado en las observaciones restantes.

La **moda** de un atributo es el valor más frecuente observado. No existe función moda directamente en R, pero se puede calcular por ejemplo definiendo esta función:

```
Mode <- function(x) {  
  ux <- unique(x)  
  ux[which.max(tabulate(match(x, ux)))]  
}
```

Una vez hecho esto, podremos calcular la moda de un conjunto de datos  $x$  mediante `mode(x)`. En el caso de que existan varias modas (es decir, cuando estemos ante una distribución plurimodal),

esta función mostrará únicamente la menor de ellas (o la primera en orden alfabético, si se está analizando una variable cualitativa).

#### **EJEMPLO: Medidas de tendencia central**

```
# Importar fichero
> Datos.personas.frecuencias <- read.delim("C:/Datos personas-frecu
encias.txt")
> View(Datos.personas.frecuencias)

# Medias
> mean(Datos.personas.frecuencias$Altura)
[1] 165.7273
> mean(Datos.personas.frecuencias$Hermanos)
[1] 2.909091

# Medianas
> median(Datos.personas.frecuencias$Altura)
[1] 168
> median(Datos.personas.frecuencias$Hermanos)
[1] 3

# Modas
> source("Mode.R")
> Mode(Datos.personas.frecuencias$Altura)
[1] 156
> Mode(Datos.personas.frecuencias$Hermanos)
[1] 1
```

#### 3.4.1.2. MEDIDAS DE TENDENCIA NO CENTRAL

Entre las medidas de posición de tendencia no central, los **cuantiles** figuran entre las más utilizadas. El  $k$ -ésimo percentil de una variable numérica es un valor tal que  $k$  de las observaciones se encuentran debajo del percentil y el  $(100 - k)\%$  se encuentran sobre este valor. Los cuantiles son equivalentes a los percentiles expresados en fracciones en vez de en porcentajes.

Además, es muy común hablar de tres cuartiles específicos:

- El primer cuartil  $Q_1$  (cuartil inferior) es el percentil con  $k = 25$
- El segundo cuartil  $Q_2$  es con  $k = 50$  que coincide con la mediana
- El tercer cuartil  $Q_3$  (cuartil superior) es con  $k=75$

En R para obtener los cuantiles de una variable se emplea la función `quantile`:

```
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE)
```

donde `x` es el vector con los valores de la variable; `probs` es un argumento que indica los cuantiles que se van a calcular. Por defecto, se muestran los cuantiles: mínimo (valor 0), los tres cuantiles (25, 50 y 75) y el máximo (100). `na.rm` es un argumento lógico que indica si hay que eliminar los valores faltantes del conjunto de datos.

El mínimo y el máximo de un conjunto de datos, además de poderse calcular como los cuantiles 0 y 100 se pueden obtener utilizando las funciones de R `min` y `max`:

```
min(x, na.rm = FALSE)
max(x, na.rm = FALSE)
```

#### **EJEMPLO: Cuantiles**

```
# Cálculo de cuantiles por defecto de la variable Altura
> quantile(Datos.personas.frecuencias$Altura)
 0%    25%    50%    75%   100%
130.00 156.00 168.00 175.75 190.00

# Cálculo de cuantiles especificados de la variable Altura
> quantile(Datos.personas.frecuencias$Altura, probs=c(.1, .25, .5, .7, .9, .99))
 10%    25%    50%    70%    90%    99%
145.5 156.0 168.0 175.0 179.8 190.0
```

#### 3.4.2. Medidas de dispersión

Estas medidas nos dicen cómo de distintas o parecidas tienden a ser las observaciones respecto a un valor particular. Generalmente este valor se refiere a alguna medida de tendencia central.

Entre las medidas de dispersión más utilizadas se encuentran la **cuasi-varianza**, la **cuasi-desviación típica** y el **rango intercuartílico**, que en R se calculan a través de las funciones `var`, `sd` e `IQR`, respectivamente.

```
var(x, na.rm = FALSE)
sd(x, na.rm = FALSE)
IQR(x, na.rm = FALSE)
```

donde `x` es el vector con los valores de la variable que se está estudiando y `na.rm` indica si los valores faltantes han de ser eliminados antes del análisis.

Como se ha especificado, las funciones `var` y `sd` no calculan la varianza y la desviación típica de una variable, sino su cuasi-varianza y su cuasi-desviación típica. En caso de necesitar la varianza o la

desviación típica, basta con multiplicar el resultado de las funciones `var` y `sd` por  $(n - 1)/n$ , siendo  $n$  el número total de datos con el que se está trabajando.

A partir de las funciones anteriores se pueden calcular otras medidas, como el **coeficiente de variación de Pearson** o el **rango**. El coeficiente de variación se emplea para comparar la representatividad de la media entre distintas variables y se obtiene dividiendo la desviación típica de una variable entre su media. Por su parte, el **rango** es una medida de dispersión que se obtiene como la diferencia entre los valores máximo y mínimo.

#### **EJEMPLO: Medidas de dispersión**

```
# cuasi-varianza
> var(Datos.personas.frecuencias$Altura, na.rm = TRUE)
[1] 227.6364

# cuasi-desviación típica
> sd(Datos.personas.frecuencias$Altura, na.rm = TRUE)
[1] 15.08762

# rango intercuartílico
> IQR(Datos.personas.frecuencias$Altura, na.rm = TRUE)
[1] 19.75
```

#### 3.4.3. Medidas de forma

Estas medidas se centran en el estudio de la forma que presenta una distribución a través del análisis de la simetría y la curtosis o el apuntamiento de la distribución en cuestión.

Para determinar la simetría de una distribución se emplea la función `skewness`, contenida en el paquete `e1071`, por lo que tendremos que instalarlo (desde la ventana **Packages** de RStudio).

Una vez instalado y cargado el paquete `e1071`, ya podemos utilizar la función `skewness`. Su sintaxis es:

```
skewness(x, na.rm = FALSE)
```

donde `x` es el vector que incluye los valores de la variable y `na.rm`: es un argumento lógico que indica si hay que eliminar los valores faltantes del conjunto de datos.

Si el `skewness` o sesgo es cero la distribución es simétrica. Si su valor es positivo, la distribución es asimétrica positiva o a la derecha, tendrá una cola asimétrica hacia los valores positivos. Si es negativo, la distribución es asimétrica negativa o a la izquierda. La Figura 3.4-1 muestra los distintos casos.



Figura 3.4-1: Sesgo de una distribución.

De forma análoga, para estudiar la curtosis de un conjunto de datos emplearemos la función `kurtosis` que también está contenida en el paquete `e1071`.

```
kurtosis(x, na.rm = FALSE)
```

donde los parámetros `x` y `na.rm` se definen de forma similar al caso anterior.

Si el coeficiente es nulo, la distribución se denomina mesocúrtica. Si es positivo, la distribución es más puntiaguda y se denomina leptocúrtica (hay una mayor concentración de los datos en torno a la media). Si es negativo, se trata de una distribución platicúrtica y hay una menor concentración de datos en torno a la media. Sería una distribución más achatada que la normal (ver Figura 3.4-2).

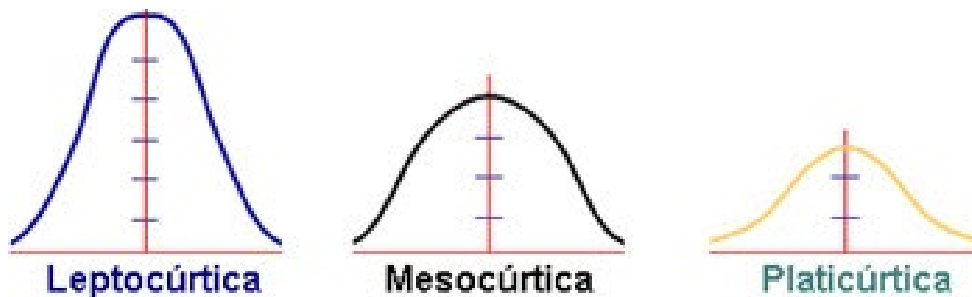


Figura 3.4-2: Curtosis de una distribución.

#### **EJEMPLO: Medidas de forma**

```
> install.packages("e1071")
> library(e1071)

# simetría de una distribución
> skewness(Datos.personas.frecuencias$Altura, na.rm = TRUE)
[1] -0.4437384

# coeficiente de curtosis
> kurtosis(Datos.personas.frecuencias$Altura, na.rm = TRUE)
[1] -0.4568731
```

### 3.4.4. Funciones resumen

Existen funciones en R que calculan a la vez algunas de las medidas que se han descrito hasta ahora. `summary` es un buen ejemplo de este tipo de funciones, ya que cuando se aplica a una variable cuantitativa devuelve el mínimo, el máximo, la media, la mediana y los cuartiles primero y tercero de la variable. Su sintaxis es la siguiente:

`summary(object)`

donde `object` es el objeto (la variable en nuestro caso) del cual queremos obtener el resumen.

Si las variables son categóricas `summary` nos devuelve la tabla de frecuencias.

#### **EJEMPLO: Funciones resumen**

```
> summary(Datos.personas.frecuencias$Altura)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
130.0   156.0   168.0   165.7   175.8   190.0
```

### 3.5. Descripción gráfica de datos en R

Las gráficas son la mejor forma de simplificar lo complejo. Un buen gráfico suele ser más accesible que una tabla. Sin embargo, es muy importante tener claro qué gráfico queremos hacer, puesto que R ofrece una gran variedad. Sus facilidades gráficas constituyen una de las componentes más importantes de este lenguaje. R incluye muchas y muy variadas funciones para hacer gráficas estadísticas, desde gráficos muy simples a figuras de gran calidad para incluir en artículos y libros. Además, permite exportar gráficas en distintos formatos (pdf, jpeg, gif, etc.). Para ver una demo de gráficos podemos utilizar el comando `demo(graphics)`. En esta sección vamos a ver algunas de las posibilidades atendiendo al tipo de datos.

#### 3.5.1. Gráficos básicos

Para realizar gráficos en R podemos utilizar el sistema tradicional (que es el que vamos a tratar en esta parte introductoria de gráficos) o bien algún paquete como por ejemplo `ggplot2`. Paquetes gráficos se estudiarán en otro módulo del máster.

Podemos dividir los comandos para efectuar las gráficas en tres grupos:







- 1) Funciones para crear gráficas de alto nivel, es decir, ya programadas y que admiten diferentes posibilidades.
- 2) Funciones de bajo nivel, que permiten un control más fino del dibujo y permiten crear gráficas a medida.
- 3) Funciones para el uso interactivo, para extraer información de una gráfica o una modificación mediante el ratón.

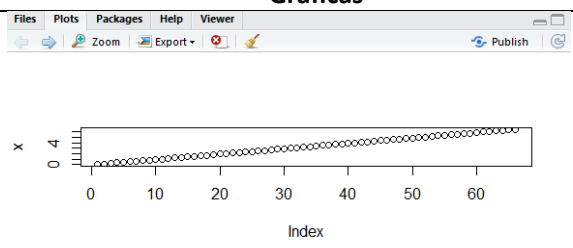
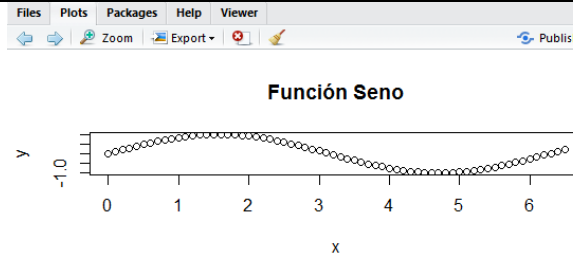
### 3.5.1.1. FUNCIÓN PLOT: plot()

Es el procedimiento gráfico de alto nivel más habitual para dibujar datos.

#### EJEMPLO: Dibujando con plot

Vamos a ver mediante un ejemplo distintas posibilidades de la función plot. En el fichero denominado GRAFICAS.R disponemos de todo el código que se utiliza en este ejemplo. Podemos ejecutarlo para ver cómo funciona. Las distintas gráficas que vayamos generando se pueden visualizar en la ventana **Plots** de RStudio. En este documento las mostramos en la Tabla 3.5-1.

Desde esta ventana podemos realizar diferentes acciones: acceder a gráficas anteriores , acceder a gráficas posteriores , realizar un zoom de la gráfica  **Zoom**, exportar la gráfica  **Export**, eliminar la gráfica actual  o eliminar todas las gráficas . Todas estas acciones las podemos realizar desde el menú **Plots** de la barra principal.

Código	Gráficas
<pre>x&lt;-(0:65)/10 y&lt;-sin(x) plot(x)</pre>	
<pre>plot(x,y) plot(x,y,main="Función Seno")</pre>	



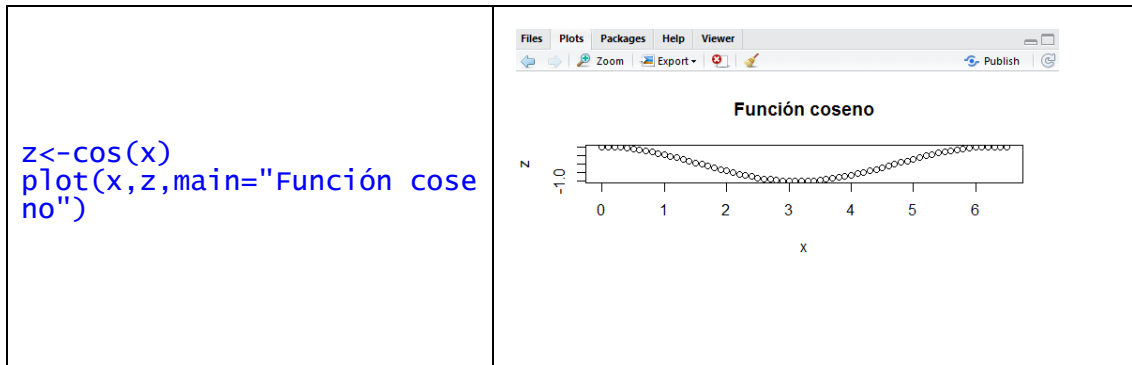


Tabla 3.5-1: Gráficas “Dibujando con `plot`” del fichero GRAFICAS.R.

### OPCIONES DE LA FUNCIÓN PLOT

Algunas de las opciones más útiles de la función `plot` son las que se muestran en la Tabla 3.5-2.

Función	Acción
<code>main</code>	Cambia el título del gráfico
<code>sub</code>	Cambia el subtítulo del gráfico
<code>type</code>	Tipo de gráfico (puntos, líneas, etc.)
<code>xlab, ylab</code>	Cambia las etiquetas de los ejes
<code>xlim, ylim</code>	Cambia el rango de valores de los ejes
<code>lty</code>	Cambia el tipo de línea
<code>lwd</code>	Cambia el grosor de línea
<code>col</code>	Color con el que dibuja

Tabla 3.5-2: Algunas opciones de `plot`.

### EJEMPLO: Algunas opciones de `plot`

El código de las distintas gráficas de este ejemplo lo tenemos en el fichero denominado GRAFICAS.R y en la Tabla 3.5-3.

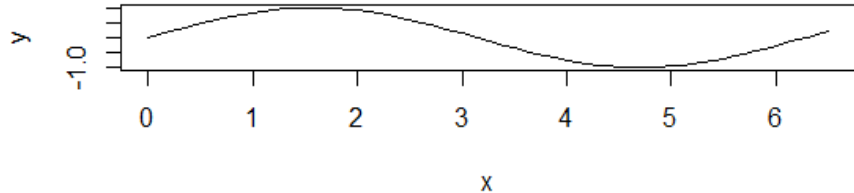
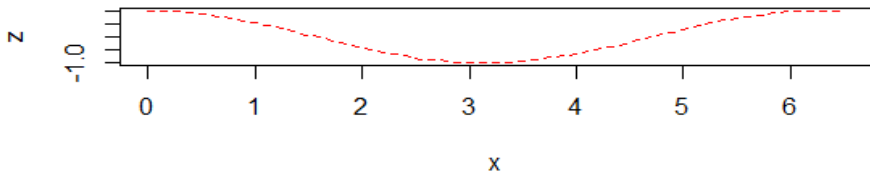
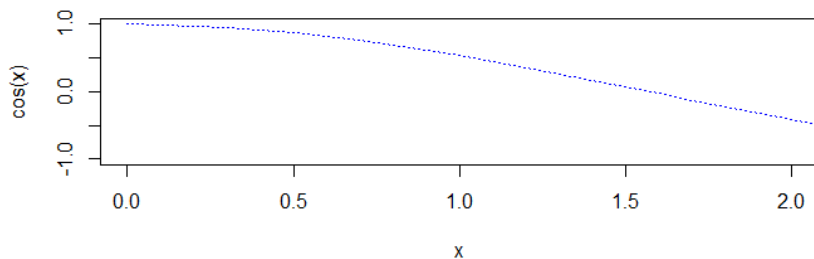
<code>plot(x,y,main="Seno",type='l')</code>

<code>plot(x,z,main="Coseno",lty=2, col="red", type='l')</code>

<code>plot(x,z,main="Coseno",lty=3, col="blue",type='l',xlim=c(0,2),ylab="cos(x)")</code>


Tabla 3.5-3: Gráficas “Algunas opciones de `plot`” del fichero GRAFICAS.R.

### 3.5.1.2. ALGUNOS PROCEDIMIENTOS DE BAJO NIVEL

Hay una serie de funciones que permiten dibujar sobre una gráfica ya creada. Las más habituales son las que mostramos en la Tabla 3.5-4.

Función	Acción
<code>points (x,y,...)</code>	Dibuja una nube de puntos
<code>lines (x,y,...)</code>	Dibuja una línea que une todos los puntos
<code>polygons (x,y,...)</code>	Dibuja un polígono cerrado
<code>Text (x,y,labels,...)</code>	Escribe texto en unas coordenadas

Tabla 3.5-4: Algunos procedimientos de bajo nivel.

### **EJEMPLO: Algunos procedimientos de bajo nivel**

El código de la gráfica de la Figura 3.5-1 generada en este ejemplo lo tenemos en el fichero denominado GRAFICAS.R. Es el siguiente:

```
plot(x,y,main="Funciones seno y coseno",type="l")
points(x,y)
lines(x,z,col="blue",lty=2) #col=4 es equivalente
text(x=c(0.5,0.5),y=c(0,1),labels=c("sin(x)","cos(x)"),col=c("black","blue"))
```

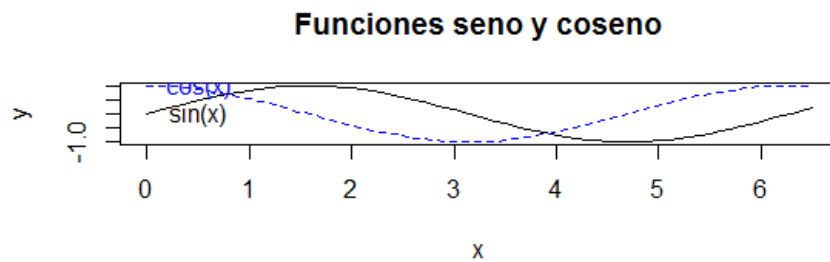


Figura 3.5-1: Gráficas “Algunos procedimientos de bajo nivel” del fichero GRAFICAS.R.

### 3.5.1.3. PRIMERAS FUNCIONES GRÁFICAS INTERACTIVAS

En R hay una serie de funciones que permiten completar los gráficos de manera interactiva por parte del usuario. Por ejemplo:

<code>identify(x,y,etiquetas)</code>	Identifica los puntos con el ratón y escribe la correspondiente etiqueta
<code>locator()</code>	Devuelve las coordenadas de los puntos

Tabla 3.5-5: Algunas funciones para completar gráficos de forma interactiva.

### **EJEMPLO: Algunas funciones gráficas interactivas**

El código de la gráfica de la Figura 3.5-2 generada en este ejemplo lo tenemos en el fichero denominado GRAFICAS.R. Es interesante ejecutar el código para que nos demos cuenta de cómo R escribe las correspondientes etiquetas una vez que se ha seleccionado su posición con el ratón. El código es el siguiente:

```
plot(x,y,main="Funciones seno y coseno",type="l")  
lines(x,z,col=2,lty=2)  
legend(locator(1),legend=c("sin(x)","cos(x)"),lty=c(1,2),col=c(1,2)  
)
```

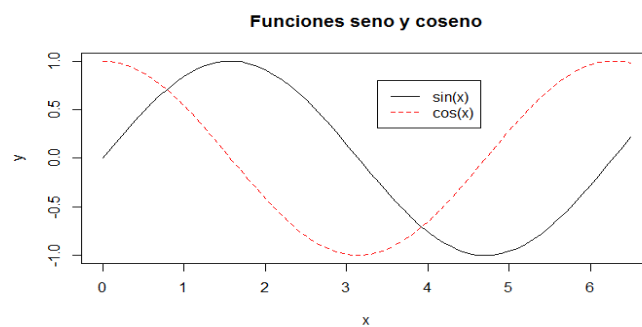


Figura 3.5-2: Gráficas "Funciones gráficas interactivas" del fichero GRAFICAS.R.

El código de la gráfica de la Figura 3.5-3 lo tenemos también en el fichero GRAFICAS.R. Con él aprendemos cómo utilizar la función `identify`. El código es el siguiente:

```
x<-1:10; y<-sample(1:10)  
nombres<-paste("punto", x,"", y, sep="")  
#nombres<-paste("punto",x)  
plot(x,y);identify(x,y,labels=nombres)
```

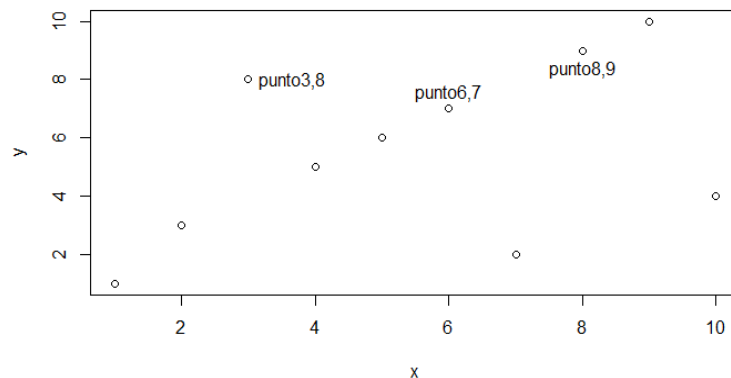


Figura 3.5-3: Gráficas "Funciones gráficas interactivas" del fichero GRAFICAS.R.

### 3.5.2. Gráficos para variables cualitativas o cuantitativas discretas

Los gráficos más habituales para representar variables cualitativas o cuantitativas discretas son los siguientes:

- Gráficos de barras: `barplot()`
- Gráficos de sectores: `pie()`
- Polígonos de frecuencias y gráficos de puntos: `dotchart()`

En esta sección vamos a trabajar las distintas representaciones utilizando los datos del fichero **Datos personas-frecuencias.txt** (ver sección 3.3.1). En el fichero `GRAFICOS VARIABLES CUALIT CUANTDISCR.R` podemos encontrar todo el código de los distintos ejemplos de creación gráficas para variables cualitativas y cuantitativas discretas. A continuación, vamos a ver cómo realizar estos resúmenes gráficos de las variables cualitativas o de las frecuencias de las variables cuantitativas discretas a partir de las frecuencias.

#### 3.5.2.1. GRÁFICOS DE BARRAS

La sintaxis de la función `barplot` con sus argumentos más importantes es:

```
barplot(x, horiz = FALSE, col = NULL, main = NULL, sub = NULL, xlab = NULL, ylab = NULL)
```

donde

x	Representa el vector con las frecuencias de las observaciones. Puede ser una tabla de frecuencia (de las obtenidas con <code>table</code> o <code>prop.table</code> )
horiz	Es un argumento lógico que indica si las barras del gráfico de barras se dibujan de forma vertical ( <code>horiz = FALSE</code> , que es la opción por defecto) u horizontal ( <code>horiz = TRUE</code> )
col	Es el vector que indica los colores de las barras
main y sub	Son las cadenas de caracteres que especifican el título y el subtítulo del gráfico
xlab e ylab	Son las cadenas de caracteres que especifican los nombres de los ejes X e Y

#### EJEMPLO: Creando un gráfico de barras

Utilizaremos en primer lugar la variable cualitativa Raza del fichero **Datos personas-frecuencias.txt** (ver sección 3.3.1). Los distintos gráficos de barras se realizan de la siguiente forma:

```
#Lectura del fichero de datos: importar fichero
Datos.personas.frecuencias <- read.delim("C:/Datos personas-frecuen
cias.txt")
View(Datos.personas.frecuencias)
attach(Datos.personas.frecuencias)

#Frecuencia absoluta de razas
tabla_raza<-table(Datos.personas.frecuencias$Raza)
tabla_raza

#Diagrama de barras de la variable Raza
barplot(tabla_raza,ylab="Frecuencias absolutas",main="Diagrama de b
arras de Razas")
```

El gráfico de barras generado se muestra en la Figura 3.5-4.

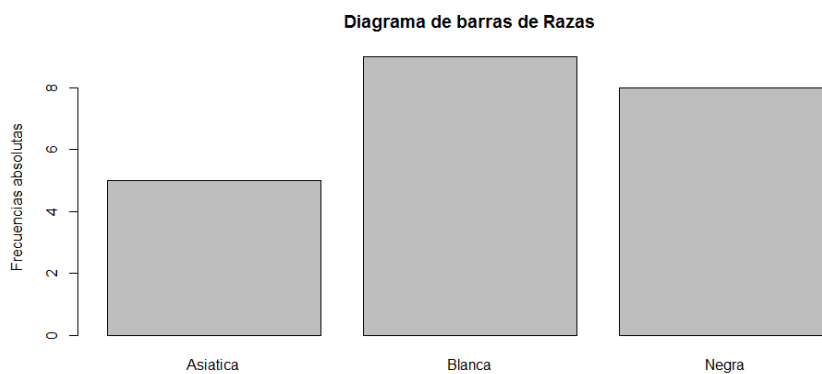


Figura 3.5-4: Gráfico de barras de las frecuencias absolutas de la variable cualitativa Raza.

Ver que también podemos hacer un diagrama de barras de la variable Raza utilizando las frecuencias relativas (ver Figura 3.5-5) y el código siguiente:

```
#Frecuencia relativa de razas
frel<-prop.table(tabla_raza)

#Diagrama de barras de frecuencias relativa de la variable Raza
barplot(frel,ylab="Frecuencias relativas",main="Diagrama de barras
de Equipo")
```

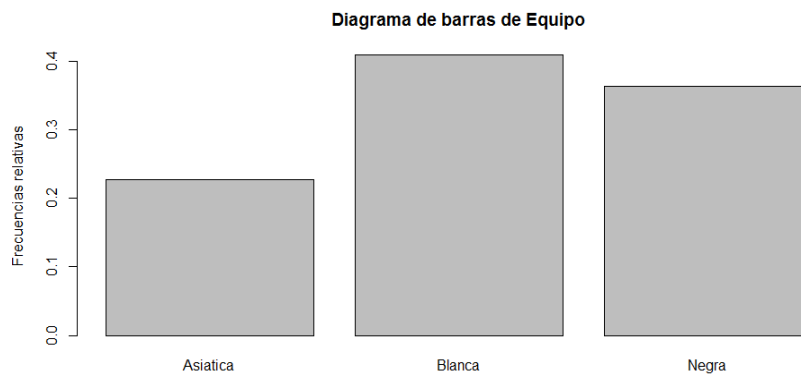


Figura 3.5-5: Gráfico de barras de las frecuencias relativas de la variable cualitativa Raza.

De manera similar al caso de variables cualitativas podemos realizar diagramas de barras de variables cuantitativas discretas (ver Figura 3.5-6). Por ejemplo, para hacer un diagrama de barras de la variable Hermanos del fichero **Datos personas-frecuencias.txt** (ver sección 3.3.1) basta con ejecutar el código:

```
tabla_hermanos<-table(Hermanos)
frel<-prop.table(tabla_hermanos)
fabsacum<-as.table(cumsum(tabla_hermanos))
frelacum<-as.table(cumsum(frel))

#Diagrama de barras de frecuencias absolutas de la variable cuantitativa discreta Hermanos
barplot(tabla_hermanos,ylab="Frecuencias absolutas",main="Diagrama de barras de Hermanos")

#Diagrama de barras de frecuencias absolutas acumuladas de la variable cuantitativa discreta Hermanos
barplot(fabsacum,ylab="Frecuencias absolutas acumuladas",main="Diagrama de barras de Hermanos")
```

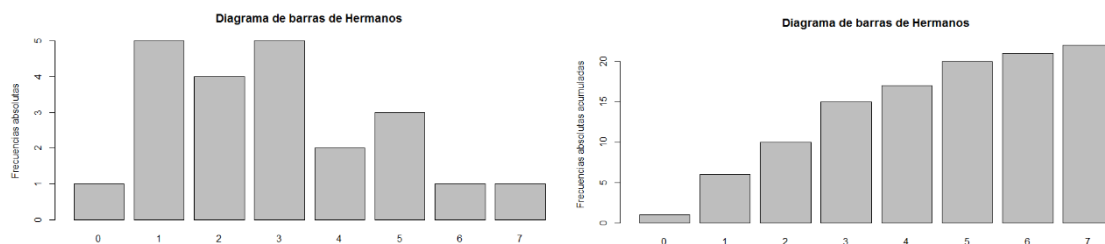


Figura 3.5-6: Gráficos de barras de las frecuencias absolutas y frecuencias absolutas acumuladas de la variable cuantitativa discreta Hermanos.

### 3.5.2.2. GRÁFICOS DE SECTORES

Estos gráficos representan la frecuencia de los elementos en un círculo. Cada elemento tiene una participación proporcional a su frecuencia relativa.

La sintaxis de la función `pie` con sus argumentos más importantes es:

```
pie(x, labels = names(x), clockwise = FALSE, init.angle = if(clockwise) 90 else 0, col = NULL, main = NULL)
```

donde

<code>x</code>	Es el vector con las frecuencias de las observaciones. Puede ser una tabla de frecuencia (de las obtenidas con <code>table</code> o <code>prop.table</code> )
<code>labels</code>	Es el vector de cadenas de caracteres que indican los nombres de cada una de las categorías que aparecen en el gráfico de sectores
<code>clockwise</code>	Es el argumento lógico que indica si los sectores se dibujan en sentido horario ( <code>clockwise = TRUE</code> ) o en sentido antihorario ( <code>clockwise = FALSE</code> , que es la opción por defecto).
<code>init.angle</code>	Es el valor numérico que indica el ángulo (en grados) en el que se sitúa el primer sector. Por defecto, el primer sector empieza a dibujarse a los 90 grados (- a las 12 en punto -, cuando <code>clockwise</code> es igual a <code>TRUE</code> ) o a los 0 grados (- a las 3 en punto -, cuando <code>clockwise</code> es igual a <code>FALSE</code> )
<code>col</code>	Es el vector que indica los colores de los sectores del gráfico
<code>main</code>	Es la cadena de caracteres que especifica el título del gráfico

### **EJEMPLO: Creando un gráfico por sectores**

Para crear el gráfico por sectores usaremos la variable cualitativa Raza (ver Figura 3.5-7) de nuestro fichero **Datos personas-frecuencias.txt** utilizando la función `pie`:

```
#Gráfico por sectores de la variable cualitativa Raza
pie(tabla_raza,col=rainbow(6),main=c("Grafico por sectores de Razas"))
```



Figura 3.5-7: Gráfico por sectores de las frecuencias relativas de la variable cualitativa Raza.

De manera similar podemos realizar gráficos por sectores de variables cuantitativas discretas. Por ejemplo, para hacer un gráfico por sectores de la variable Hermanos (ver Figura 3.5-8) del fichero **Datos personas-frecuencias.txt** (ver sección 3.3.1) basta con ejecutar el código:



```
#Gráfico por sectores de la variable cuantitativa discreta Hermanos  
pie(tabla_hermanos,col=rainbow(6),main=c("Grafico por sectores de H  
ermanos"))
```

Grafico por sectores de Hermanos



Figura 3.5-8: Gráfico por sectores de las frecuencias absolutas de la variable cuantitativa discreta Hermanos.

### 3.5.2.3. POLÍGONOS DE FRECUENCIAS Y GRÁFICOS DE PUNTOS

El polígono de frecuencia es un gráfico que se crea a partir de un histograma de frecuencia. Se realiza uniendo los puntos de mayor altura de las columnas del histograma. En R podemos utilizar el comando `plot` para hacer un polígono de frecuencias absolutas (ver Figura 3.5-9) de la manera siguiente:

```
#Polígono de frecuencias absolutas de Raza  
plot(tabla_raza,type="l",main=c("Poligono de frecuencias absolutas  
de Razas"),ylab= "Frecuencias absolutas")
```

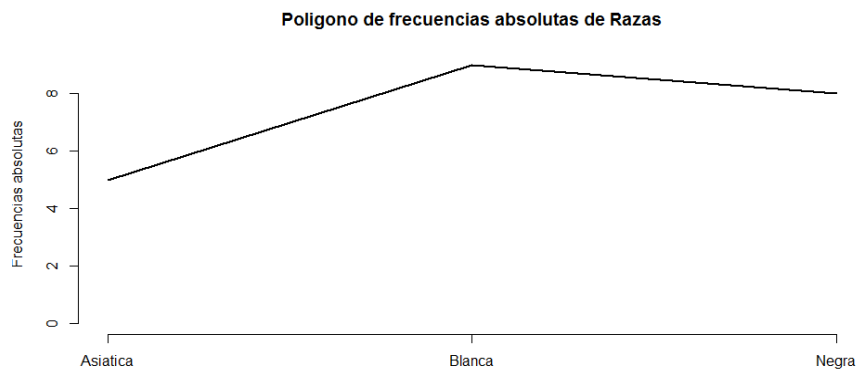


Figura 3.5-9: Polígono de frecuencias absolutas de la variable cualitativa Raza.

Análogamente, el polígono de frecuencias relativas de la variable Raza (ver Figura 3.5-10) se genera con el comando:

```
#Polígono de frecuencias relativas de Raza  
plot(frel,type="l",main=c("Poligono de frecuencias relativas de Raz  
as"),ylab="Frecuencias relativas")
```

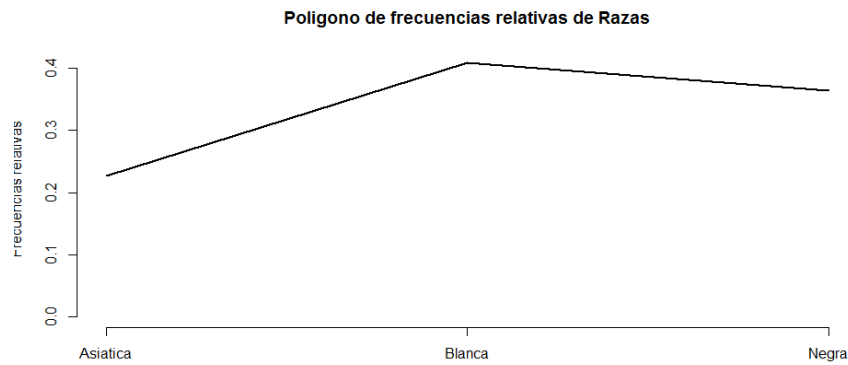


Figura 3.5-10: Polígono de frecuencias relativas de la variable cualitativa Raza.

Una alternativa al polígono de frecuencias es el gráfico de puntos `dotchart()`, que es como un polígono de frecuencias salvo que no se conectan las frecuencias con líneas (ver Figura 3.5-11). El comando para generar esta figura es el siguiente:

```
#Gráfico de puntos  
dotchart(frel, labels=c("Asiatica", "Blanca", "Negra"), main="Gráfico de puntos por Raza")
```

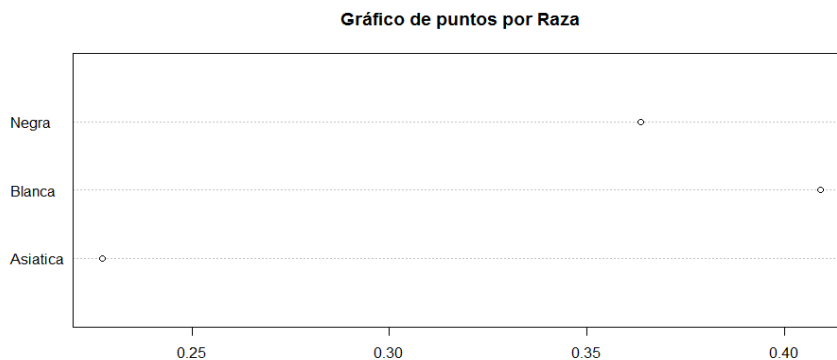


Figura 3.5-11: Gráfico de puntos de la variable cualitativa Raza.

Para realizar polígonos de frecuencias para la variable cuantitativa discreta Hermanos (ver Figura 3.5-12) basta con ejecutar los comandos siguientes:

```
#Polígono de frecuencias absolutas de la variable cuantitativa discreta Hermanos  
plot(tabla_hermanos, type="l", main="Polígono de frecuencias absolutas de Hermanos", ylab="Frecuencias absolutas")
```

```
#Polígono de frecuencias absolutas acumuladas de la variable cuantitativa discreta Hermanos  
plot(fabsacum, type="l", main="Polígono de frecuencias absolutas acumuladas de Hermanos", ylab="Frecuencias absolutas acumuladas")
```

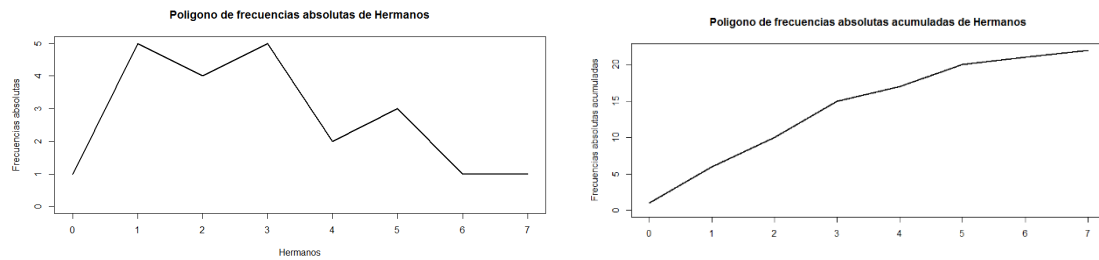


Figura 3.5-12: Polígonos de frecuencias absolutas y frecuencias absolutas acumuladas de la variable cuantitativa discreta Hermanos.

Los gráficos de puntos para la variable Hermanos (ver Figura 3.5-13) los podemos obtener mediante:

#Gráfico de puntos de las frecuencias absolutas de la variable cuantitativa discreta Hermanos  
`dotchart(tabla_hermanos,main="Graficos de puntos de Hermanos")`

#Gráfico de puntos de las frecuencias absolutas acumuladas de la variable cuantitativa discreta Hermanos  
`dotchart(fabsacum,main="Graficos de puntos de Hermanos")`

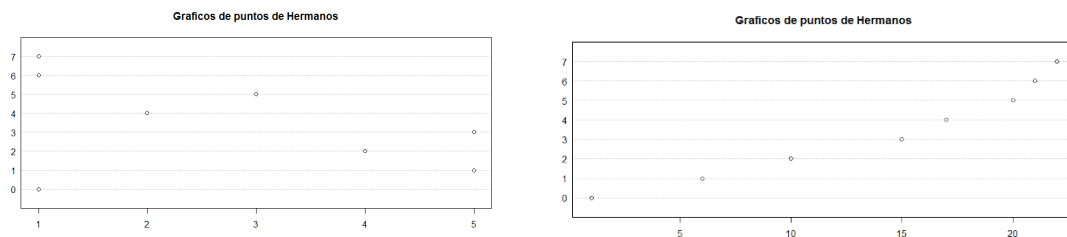


Figura 3.5-13: Gráficos de puntos de frecuencias absolutas (izqda.) y frecuencias absolutas acumuladas (drcha.) de la variable cuantitativa discreta Hermanos.

### 3.5.3. Gráficos para variables cuantitativas continuas

Los gráficos más habituales para representar variables cualitativas o cuantitativas discretas son:

- Histograma: `hist()`
- Boxplot o gráfico de caja y bigotes: `boxplot()`
- Diagrama de tallos y hojas: `stem()`
- Diagramas de dispersión

En esta sección vamos a trabajar las distintas representaciones utilizando los datos del fichero **Datos personas-frecuencias.txt** (ver sección 3.3.1). En el fichero GRAFICOS VARIABLES CUANTITATIVAS

CONTINUAS.R podemos encontrar todo el código de los distintos ejemplos desarrollados en esta sección.

### 3.5.3.1. HISTOGRAMAS

Los histogramas muestran la distribución de los valores de una variable. Un histograma consiste en coger todos los datos de la tabla y separarlos en grupos (contenedores) que cubran todos los valores posibles de la variable que estamos estudiando, por ejemplo, de 0 a 10, de 10 a 20, de 20 a 30... después contamos cuántos registros hay dentro de cada grupo y lo representamos mediante gráfico de barras. Esto es, un histograma consiste en crear gráficos de barra por cada contenedor en los que hemos dividido los valores de los datos que queremos representar. La altura de cada barra indica el número de elementos o frecuencia del contenedor. La forma del histograma depende del número de contenedores.

La sintaxis de la función `hist` con sus argumentos más importantes es:

```
hist(x, breaks = "Sturges", right = TRUE, col = NULL, main = paste(
"Histograma de" , xname))
```

donde

x	Es el vector de valores de la variable a partir de los cuales se dibujará el gráfico
breaks	Indica la forma en la que se calcularán los intervalos en el histograma. Las opciones disponibles para este parámetro son "Sturges" (que es la opción por defecto) "Scott" y "FD" (Freedman-Diaconis)
right	Es el argumento lógico que indica si los intervalos son cerrados por la izquierda y abiertos por la derecha (en cuyo caso, <code>right = TRUE</code> , que es la opción por defecto) o viceversa ( <code>right = FALSE</code> )
col	Es el vector que indica los colores del histograma
main	Es la cadena de caracteres que especifica el título del gráfico

### EJEMPLO: Creando un histograma

Vamos a crear un histograma para la variable cuantitativa continua `Peso` del **fichero Datos personas-frecuencias.txt**

En la Figura 3.5-14 se muestra un histograma con la distribución de personas según su peso.

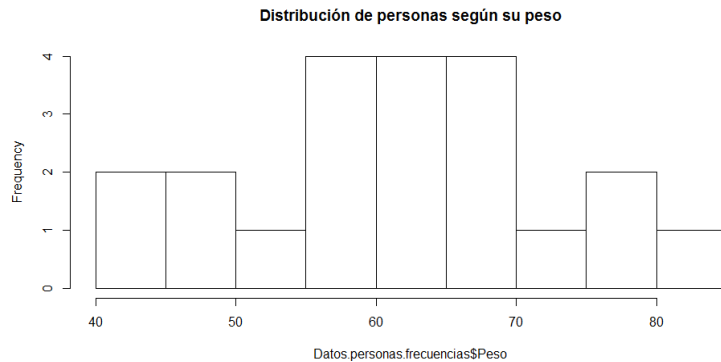


Figura 3.5-14: Histograma de la variable `Peso`. Número de contenedores por defecto.

El comando utilizado en R para generar este histograma es:

```
#Histograma de la variable Peso
hist(Datos.personas.frecuencias$Peso,main="Distribución de personas
según su peso")
```

R selecciona por defecto selecciona el número de contenedores siguiendo un método interno llamado método de Sturges. Para poder utilizar el número de clases que a nosotros más nos interese, tenemos que crear un vector con los puntos de corte de las clases. Las siguientes líneas de código muestran cómo hacer esto para el caso de 12 contenedores. En la Figura 3.5-15 se muestra el resultado de la ejecución de estas líneas.

```
#Creación de histograma de la variable Peso eligiendo el número de
contenedores
#Definiendo número de contenedores del histograma
n.clases=12

#elimino NA de la variable Peso
Peso<-na.omit(Peso)
puntos=min(Peso)+(0:n.clases)*(max(Peso)-min(Peso))/n.clases
hist(Peso,breaks=puntos,col="yellow",main="Distribución de personas
según su peso")
```

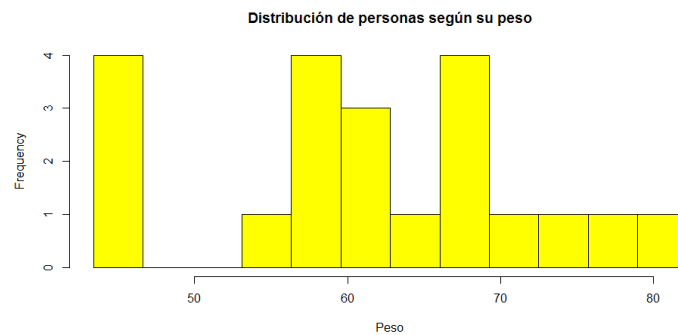


Figura 3.5-15: Histograma de la variable Peso. Número de contenedores seleccionados.

### 3.5.3.2. BOXPLOTS O GRÁFICO DE CAJA Y BIGOTES

Los boxplots o diagramas de caja y bigotes se construyen a partir de cinco medidas: el mínimo valor, el primer cuartil  $Q_1$ , la mediana, el tercer cuartil  $Q_3$  y el máximo valor. Se construye un rectángulo como el de la Figura 3.5-16, en el que el alto de la caja está definido por  $Q_1$  y  $Q_3$ . La altura es el rango intercuartil RIC ( $Q_3 - Q_1$ ). A mayor altura de la caja mayor variabilidad de los datos. Del centro de la caja salen dos segmentos, uno hasta el mínimo y otro hasta el máximo. Estos segmentos representan los datos que están por fuera del rango inter-cuartílico (brazos de largo  $Q_1 - 1,5 \cdot \text{RIC}$  para la recta inferior y  $Q_3 + 1,5 \cdot \text{RIC}$  para la recta superior). Dentro de la caja se dibuja una línea que indica la mediana de los datos. Los valores por fuera del largo de los brazos son considerados atípicos.

El boxplot nos da información sobre la simetría de la distribución de los datos. Si la mediana no está en el centro del rectángulo, la distribución no es simétrica. Estos gráficos son útiles para detectar la presencia de valores atípicos o outliers.

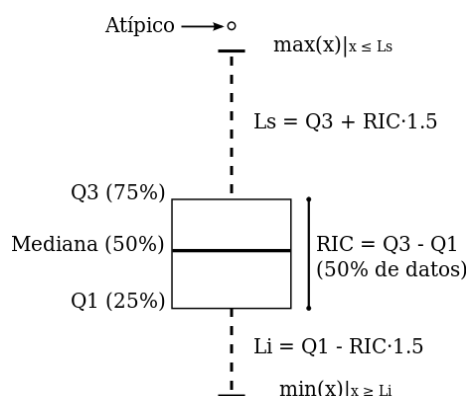


Figura 3.5-16: Parámetros de caja en boxplot.

La sintaxis de la función `boxplot` con sus argumentos más importantes es:

`boxplot(x, range=1.5, col=NULL, main=NULL)`

donde

x	Es el vector con las frecuencias de las observaciones. Puede ser una tabla de frecuencia (de las obtenidas con <code>table</code> o <code>prop.table</code> )
range	Es el valor numérico que determina la extensión de los bigotes de la caja. Para un valor positivo de <code>range</code> , los bigotes se extienden hasta el último dato que no supere 1.5 veces la longitud de la caja. Para un valor de 0, los bigotes se extienden hasta el dato más lejano
col	Es el vector que indica los colores de las barras o los sectores del gráfico
main	Es la cadena de caracteres que especifican el título del gráfico

#### **EJEMPLO:** Creando un gráfico de caja y bigotes

En este ejemplo vamos a utilizar los datos contenidos en el fichero “**Datos personas-boxplot.txt**” para representar el diagrama de caja y bigotes de la variable creatinina contenida en él.

El gráfico que se obtiene se muestra en la Figura 3.5-17 y el código para obtener este gráfico es el siguiente:

```
#Carga del fichero
Datos.personas.boxplot <- read.delim("D:/Datos personas-boxplot.txt")
> view(Datos.personas.boxplot)

#Representación boxplot
> boxplot(Datos.personas.boxplot$creatinina, xlab="Pesos", main = "
Cajas y bigotes para la variable peso")
```

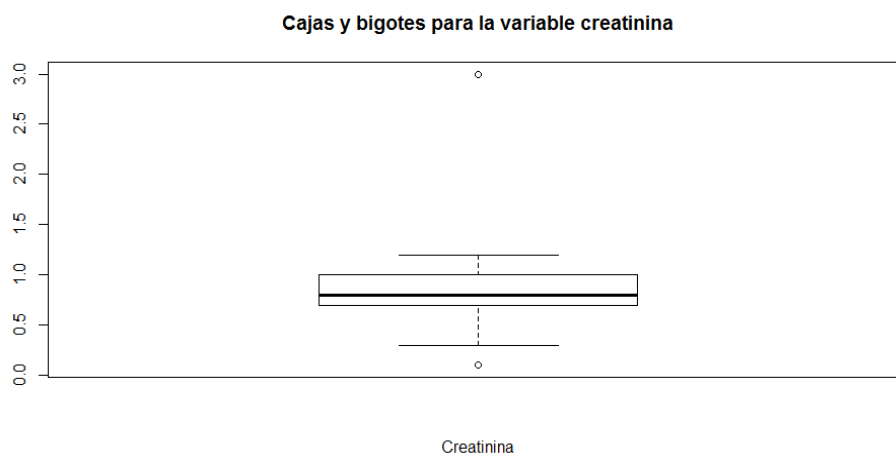


Figura 3.5-17: Boxplot de la variable creatinina.

En la Figura 3.5-17 los círculos representan los *outliers*. La Figura 3.5-18 muestra un histograma de la variable creatinina para poder comparar ambas representaciones.

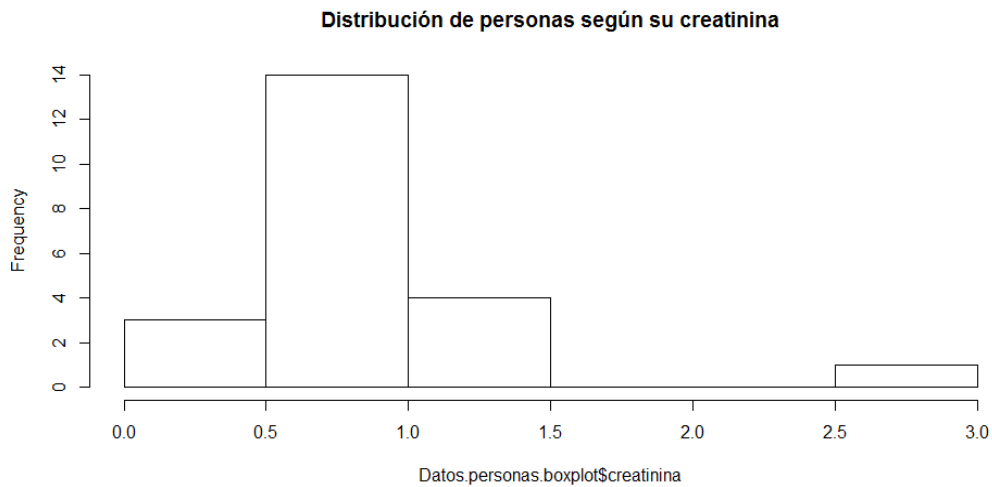


Figura 3.5-18: Histograma de la variable creatinina.

### 3.5.3.3. DIAGRAMA DE TALLOS Y HOJAS

Este gráfico permite presentar la distribución de una variable cuantitativa. Para construirlo basta separar en cada dato el último dígito de la izquierda (que constituye la **hoja**) y el resto (que representa el **tallo**).

La sintaxis de la función `stem` con sus argumentos más importantes es:

`stem(x)`

donde `x` es el vector de valores de la variable a partir de los cuales se dibujará el gráfico.

**EJEMPLO:** Creando un diagrama de tallos y hojas

```
> stem(Datos.personas.boxplot$Peso)
```

```
The decimal point is 1 digit(s) to the right of the |
```

```
4 | 3466
5 | 57779
6 | 11247889
7 | 166
8 | 2
```



#### 3.5.3.4. DIAGRAMAS DE DISPERSIÓN

Los diagramas de dispersión o scatter plots usan coordenadas cartesianas para mostrar los valores de dos variables de la misma longitud. Los valores de los atributos determinan la posición de los elementos.

En R podemos utilizar el comando plot para realizar un scatterplot de dos variables numéricas, como vamos a ver en el siguiente ejemplo.

##### **EJEMPLO: Diagrama de dispersión**

Vamos a realizar un diagrama de dispersión de la variable Tensión frente a la variable Peso y representarlo en distintos colores para cada una de las razas. Esta representación se muestra en la Figura 3.5-19. Para ello tenemos que escribir y ejecutar el código siguiente:

```
> plot(Datos.personas.boxplot$Tension,Datos.personas.boxplot$Peso,
col=Datos.personas.boxplot$Raza)
> legend('topright', levels(Datos.personas.boxplot$Raza), lty=1, col=1:3,bty='n')
```

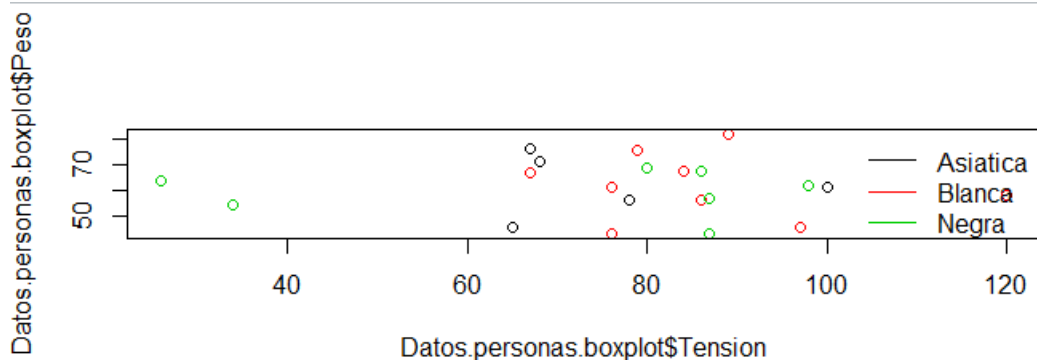
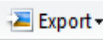


Figura 3.5-19: Ejemplo de diagrama de dispersión.

#### 3.6. Exportando gráficos en RStudio

Lo más sencillo para exportar gráficos utilizando Rstudio es hacerlo desde la pestaña Plot. En la barra del menú de esta pestaña encontramos el botón  Export. Haciendo clic en él podremos acceder a distintas opciones (ver Figura 3.6-1).

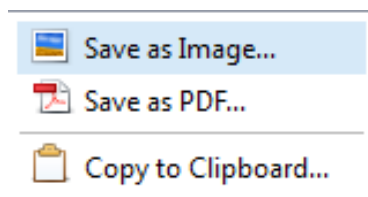


Figura 3.6-1: Opciones de Exportación de gráficos con RStudio.

Si seleccionamos **“Save as Image”** se accede a la pantalla que se muestra en la Figura 3.6-2 en la que tenemos la opción de guardar la gráfica en diferentes formatos. Además, desde esta pantalla se puede cambiar el tamaño de la imagen.



Figura 3.6-2: Formatos de exportación de gráficos.

Si seleccionamos **“Save as PDF...”** en la Figura 3.6-1 se accede a la pantalla de la Figura 3.6-3 para poder guardar la gráfica como .pdf.

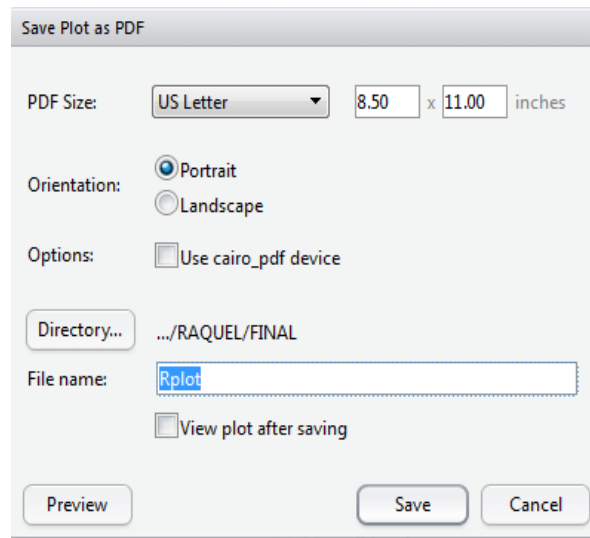


Figura 3.6-3: Guardar gráfico en formato pdf.

Si seleccionamos **“Copy to Clipboard”** en la Figura 3.6-1 RStudio copia la gráfica generada para poder pegarla en otra aplicación (Paint, MSWord,...).