

Estadística descriptiva

Sesión 2. Curso: 'Análisis estadístico aplicado con R Commander'

Paqui Corpas Burgos
correo: `corpas_fra@gva.es`

- 1 Estadística descriptiva
- 2 Análisis descriptivo univariable
- 3 Análisis descriptivo bivariable
- 4 Creación de informes con R Markdown
- 5 Ejercicio práctico

Estadística descriptiva

La **estadística descriptiva** comprende un conjunto de **técnicas numéricas y gráficas que permiten resumir la información recogida en una base de datos** y describir las características del grupo estudiado.

Es el **primer análisis estadístico en cualquier investigación** ya que permite conocer los datos y planificar análisis posteriores más complejos.

Las **principales técnicas descriptivas** son:

- Construcción de tablas de frecuencias.
- Elaboración de gráficos.
- Cálculo de medidas descriptivas de posición, dispersión y forma.

Tipos de variables

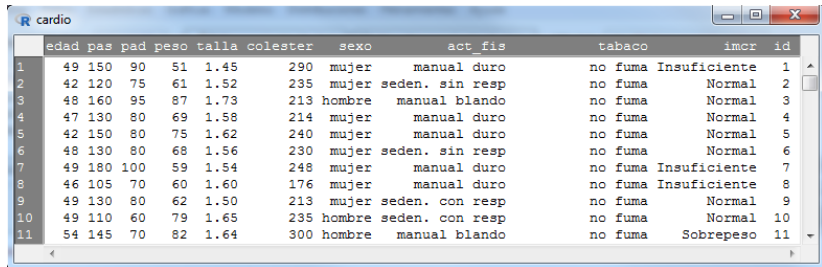
Para aplicar una técnica descriptiva es **necesario conocer previamente el tipo de variable** con la que se está trabajando.

Tipo	Clases	Ejemplo
Cualitativa	Nominal	Sexo, raza, color de ojos, ...
	Ordinal	Grado de contaminación, calificación, ...
Cuantitativa	Discreta	Nº de hermanos, Nº de materias, ...
	Continua	Peso, altura, ...

A continuación veremos las **principales técnicas** que se utilizan para describir las variables de una base de datos y cómo obtenerlas utilizando **R Commander**.

Ejemplo: Base de datos *cardio*

Una base de datos con 531 individuos (filas) y 11 variables (columnas).



	edad	pas	pad	peso	talla	colester	sexo	act_fis	tabaco	imcr	id
1	49	150	90	51	1.45	290	mujer	manual duro	no fuma	Insuficiente	1
2	42	120	75	61	1.52	235	mujer	seden. sin resp	no fuma	Normal	2
3	48	160	95	87	1.73	213	hombre	manual blando	no fuma	Normal	3
4	47	130	80	69	1.58	214	mujer	manual duro	no fuma	Normal	4
5	42	150	80	75	1.62	240	mujer	manual duro	no fuma	Normal	5
6	48	130	80	68	1.56	230	mujer	seden. sin resp	no fuma	Normal	6
7	49	180	100	59	1.54	248	mujer	manual duro	no fuma	Insuficiente	7
8	46	105	70	60	1.60	176	mujer	manual duro	no fuma	Insuficiente	8
9	49	130	80	62	1.50	213	mujer	seden. con resp	no fuma	Normal	9
10	49	110	60	79	1.65	235	hombre	seden. con resp	no fuma	Normal	10
11	54	145	70	82	1.64	300	hombre	manual blando	no fuma	Sobrepeso	11

- Variables cualitativas: sexo, act_fis, tabaco, imcr.
- Variables cuantitativas: edad, pas, pad, peso, talla, colester.

Análisis descriptivo univariable

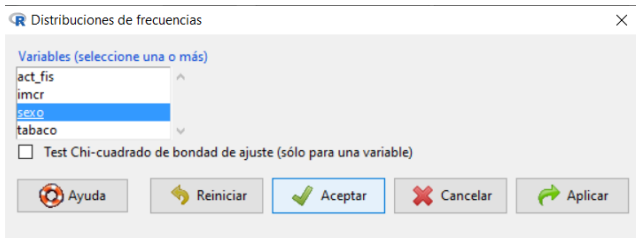
- **Tabla de frecuencias**

Una **variable cualitativa** se describe usualmente a través de una **tabla de frecuencias** mostrando el número de individuos que pertenecen a cada categoría de la variable y su porcentaje con respecto al total de individuos.

sexo	
hombre	255 (47.66 %)
mujer	280 (52.34 %)

Para obtener la tabla de frecuencias de una variable cualitativa en **R Commander** seleccionaremos:

Estadísticos > Resúmenes > Distribución de frecuencias



Indicamos la variable o variables de las que queremos obtener la tabla de frecuencias y pulsamos **Aceptar** (sólo aparecerán variables cualitativas en la selección).

Resultados

The screenshot shows the R Commander application window. The menu bar includes: Archivo, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, Ayuda. The toolbar shows buttons for 'Conjunto de datos: cardio', 'Editar conjunto de datos', 'Visualizar conjunto de datos', and 'Modelo: <No hay modelo activo>'. The 'R Script' tab is active, displaying the following R code:

```
local({  
  .Table <- with(cardio, table(sexo))  
  cat("\ncounts:\n")  
  print(.Table)  
  cat("\npercentages:\n")  
  print(round(100 * .Table/sum(.Table), 2))  
})
```

The 'Salida' (Output) pane shows the execution results:

```
+ .Table <- with(cardio, table(sexo))  
+ cat("\ncounts:\n")  
+ print(.Table)  
+ cat("\npercentages:\n")  
+ print(round(100 * .Table/sum(.Table), 2))  
+ })  
  
counts:  
sexo  
hombre mujer  
255 280  
  
percentages:  
sexo  
hombre mujer  
47.66 52.34
```

Annotations with red circles and arrows point to the output tables:

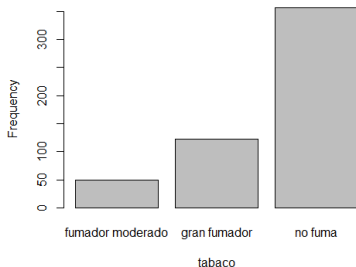
- Frecuencias absolutas** points to the 'counts:' table.
- Porcentajes** points to the 'percentages:' table.

The 'Mensajes' (Messages) pane at the bottom shows:

```
RGui with the single-document interface (SDI); see ?Commander.  
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.
```

Descripción de variables cualitativas

- **Gráficos: Diagrama de barras**

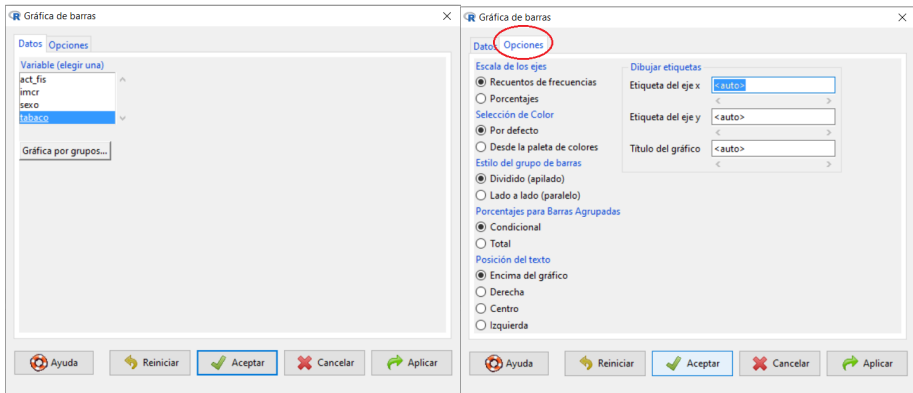


Para obtener un diagrama de barras de una variable cualitativa con **R Commander** seleccionaremos:

Gráficas > Gráfica de barras

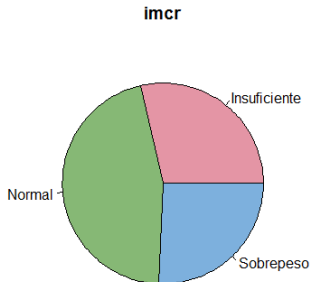
A continuación, indicamos la variable de la que queremos obtener el diagrama de barras y pulsamos **Aceptar**.

En la pestaña **Opciones** podemos personalizar la salida del gráfico: barras con las frecuencias absolutas o los porcentajes, modificar el color, el título del gráfico y las etiquetas en los ejes, etc.



Descripción de variables cualitativas

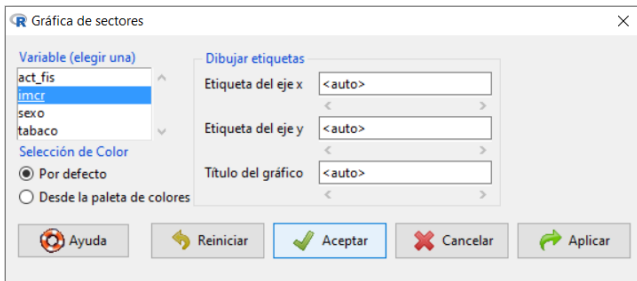
- Gráficos: Diagrama de sectores



Para obtener un diagrama de sectores de una variable cualitativa con **R Commander** seleccionaremos:

Gráficas > Gráfica de sectores

A continuación, seleccionamos la variable de la que queremos obtener el diagrama de sectores y pulsamos **Aceptar**.



Descripción de variables cuantitativas

Las principales medidas que se utilizan para describir una variable cuantitativa pueden clasificarse en:

- **Medidas de posición o localización:** indican el valor o valores entorno a los cuales se sitúan los datos.
 - **Tendencia central:** media, mediana y moda.
 - **Tendencia no central:** cuartiles, deciles y percentiles.
- **Medidas de dispersión:** indican cuánto se dispersan los datos.
 - **Absolutas:** desviación típica, varianza, rango y rango intercuartílico.
 - **Relativas:** coeficiente de variación.
- **Medidas de forma:** reflejan la forma en la que se distribuyen los datos (simetría, uni/multimodalidad).

A continuación veremos brevemente la interpretación de estas medidas. Las medidas de forma las analizaremos a nivel gráfico en este curso.

Medidas de posición o localización

Tendencia central

- **Media:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$. No es una medida robusta. Se ve influenciada por datos anormalmente altos o bajos (atípicos).
- **Mediana.** Si suponemos los datos de la muestra ordenados, es el valor por debajo del cual se encuentran el 50% de los casos. Es una medida robusta (se ve poco afectada por la presencia de valores atípicos).
- **Moda:** es el valor que más se repite (variables discretas).

Tendencia no central

- **Cuartiles.** Dividen la muestra en 4 partes iguales (Q_1, Q_2, Q_3).
 - Q_1 : valor por debajo del cual se encuentran el 25% de los casos.
 - Q_2 : valor por debajo del cual se encuentran el 50% de los casos (mediana).
 - Q_3 : valor por debajo del cual se encuentran el 75% de los casos.
- **Deciles.** Dividen la muestra en 10 partes iguales (D_1, \dots, D_9).
- **Percentiles.** Dividen la muestra en 100 partes iguales (P_1, \dots, P_{99}).

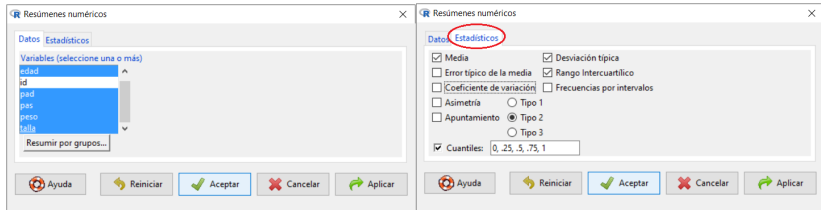
Medidas de dispersión

- **Absolutas:** dependen de las unidades en las que se miden las observaciones.
- **Desviación típica:** $s = \sqrt{\frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n}}$. Refleja cuánto se alejan los datos de la media. Está medida en las mismas unidades de los datos.
- **Varianza:** $s^2 = \frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n}$. Está medida en las unidades de los datos al cuadrado (no se puede comparar directamente con las medidas de posición).
- **Rango:** $\max(x_i) - \min(x_i)$. Se ve afectado por la presencia de datos atípicos.
- **Rango intercuartílico:** $Q_3 - Q_1$. Es menos sensible a datos atípicos.
- **Relativas:** no dependen de las unidades en las que se miden las observaciones.
- **Coefficiente de variación:** $CV = \frac{s}{\bar{x}}$. Permite comparar la variabilidad de variables medidas en distintas unidades.

Para obtener un resumen numérico de las variables cuantitativas del estudio con **R Commander** seleccionaremos:

Estadísticos > Resúmenes > Resúmenes numéricos

A continuación, seleccionamos las variables de las que queremos obtener un resumen numérico y pulsamos **Aceptar**. En la pestaña **Estadísticos** especificaremos las medidas que queremos calcular.



Resultados

The screenshot shows the R Commander application window. The menu bar includes: Fichero, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, Ayuda. The toolbar shows buttons for 'Conjunto de datos: cardio', 'Editar conjunto de datos', 'Visualizar conjunto de datos', and 'Modelo: [X] <No hay modelo activo>'. The 'R Script' tab is active, displaying the following R code:

```
library(colorspace, pos = 16)
with(cardio, pie(table(imcr), labels = levels(imcr), xlab = "", ylab = "", main = "imcr", col = rainbow_hcl(3)))
library(abind, pos = 17)
library(e1071, pos = 18)
numSummary(cardio[, c("colester", "edad", "pad", "pas", "peso", "talla"), drop = FALSE], statistics = c("mean",
  "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
```

The 'Salida' (Output) pane shows the execution results:

```
> with(cardio, pie(table(imcr), labels = levels(imcr), xlab = "", ylab = "", main = "imcr", col = rainbow_hcl(3)))
> library(abind, pos = 17)
> library(e1071, pos = 18)

> numSummary(cardio[, c("colester", "edad", "pad", "pas", "peso", "talla"), drop = FALSE], statistics = c("mean",
+   "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
      mean      sd      IQR      0%      25%      50%      75%     100%    n NA
colester 223.504854 42.24430378 54.50 133.00 193.50 221.00 248.00 353.00 515 20
edad      49.467290  5.87975285 10.00  40.00  44.00  49.00  54.00  59.00  535  0
pad       84.054206 12.62639676 15.00  45.00  75.00  80.00  90.00 120.00  535  0
pas       137.138318 20.88356182 30.00  85.00 120.00 135.00 150.00 210.00  535  0
peso      71.971857 11.48197164 16.00  45.00  64.00  71.00  80.00 112.00  533  2
talla     1.616773  0.08770304  0.14   1.42   1.55   1.62   1.69   1.86  533  2
```

The 'Mensajes' (Messages) pane shows the following messages:

```
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.
```

Otra forma de obtener una descripción de todas las variables del estudio, tanto cualitativas como cuantitativas, sería seleccionando en **R Commander**:

Estadísticos > Resúmenes > Conjunto de datos activo

The screenshot shows the R Commander window with the 'Conjunto de datos: cardio' selected. The 'R Script' tab is active, displaying the following code:

```
with(cardio, Barplot(tabaco, xlab = "tabaco", ylab = "Frequency"))
library(reshape2)
library(abind, pos = 17)
library(e1071, pos = 18)
numSummary(cardio[, c("colester", "edad", "pad", "pas", "peso", "talla"), drop = FALSE], statistics = c("mean",
  "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
summary(cardio)
```

The 'Salida' (Output) pane shows the result of the `summary(cardio)` command:

```
> summary(cardio)
      edad      pas      pad      peso      talla      colester      sexo
Min.   :40.00  Min.   : 85.0  Min.   : 45.00  Min.   : 45.00  Min.   :1.420  Min.   :133.0  hombre:255
1st Qu.:44.00  1st Qu.:120.0  1st Qu.: 75.00  1st Qu.: 64.00  1st Qu.:1.550  1st Qu.:193.5  mujer :280
Median :49.00  Median :135.0  Median : 80.00  Median : 71.00  Median :1.620  Median :221.0
Mean   :49.47  Mean   :137.1  Mean   : 84.05  Mean   : 71.97  Mean   :1.617  Mean   :223.5
3rd Qu.:54.00  3rd Qu.:150.0  3rd Qu.: 90.00  3rd Qu.: 80.00  3rd Qu.:1.690  3rd Qu.:248.0
Max.   :59.00  Max.   :210.0  Max.   :120.00  Max.   :112.00  Max.   :1.860  Max.   :353.0
              NA's :2              NA's :2              NA's :20

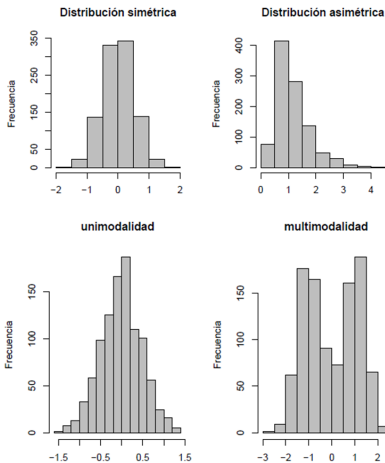
      act_fis      tabaco      imcr      id
manual blando :104  fumador moderado: 50  Insuficiente:152  Min.   : 1.0
manual duro  :210  gran fumador   :122  Normal       :244  1st Qu.:134.5
seden. con resp: 90  no fuma       :357  Sobre peso   :137  Median :268.0
seden. ain resp:131 NA's         : 6  NA's         : 2  Mean   :268.0
              3rd Qu.:401.5
              Max.   :535.0
```

The 'Mensajes' (Messages) pane at the bottom shows the following messages:

```
RGui with the single-document interface (SDI): see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.
```

Descripción de variables cuantitativas

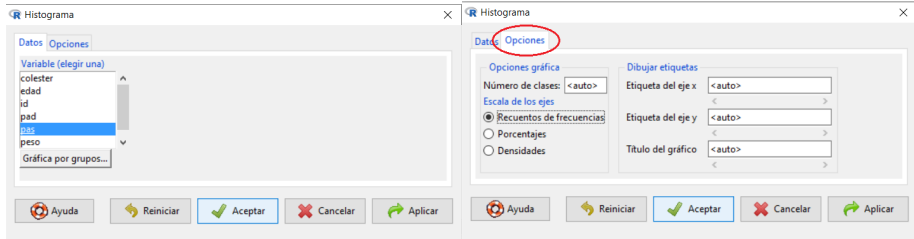
- **Gráficos: Histograma.** Da una idea clara de la forma de la distribución de los datos (simetría, uni/multimodalidad).



Para obtener un histograma de las variables cuantitativas del estudio con **R Commander** seleccionaremos:

Gráficas > Histograma

A continuación, seleccionamos la variable para la que queremos construir un histograma y pulsamos **Aceptar**. En la pestaña **Opciones** podemos personalizar la salida del histograma.

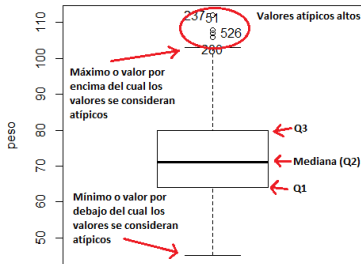


Descripción de variables cuantitativas

- **Gráficos: Diagrama de cajas.** Es útil para detectar valores atípicos en los datos.

Los datos atípicos son aquellos que quedan fuera del intervalo:

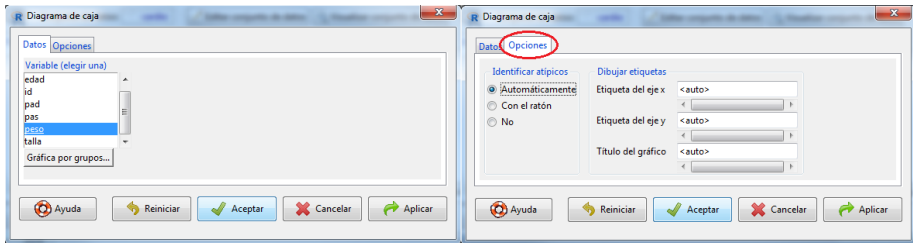
$$[Q_1 - 1.5RIC, Q_3 + 1.5RIC]$$



Para obtener un diagrama de cajas de las variables cuantitativas del estudio con **R Commander** seleccionaremos:

Gráficas > Diagrama de caja

A continuación, seleccionamos la variable para la que queremos construir el diagrama de cajas y pulsamos **Aceptar**. En la pestaña **Opciones** podemos modificar el título del gráfico y las etiquetas de los ejes.



- Descripción de variables cualitativas:
 - Tabla de frecuencias
 - Diagrama de barras
 - Diagrama de sectores
- Descripción de variables cuantitativas:
 - Resumen numérico (mínimo, cuartiles, media, máximo, desviación típica)
 - Histograma
 - Diagrama de cajas

Análisis descriptivo bivariable

Uno de los principales objetivos de cualquier investigación es **estudiar la relación entre dos variables**, observando cómo cambian los valores de una de ellas cuando se modifican los de la otra.

El análisis bivariable engloba varias técnicas estadísticas que permiten describir la relación entre dos variables. **El uso particular de cada técnica dependerá del carácter cualitativo o cuantitativo de las variables.**

Dos variables cualitativas

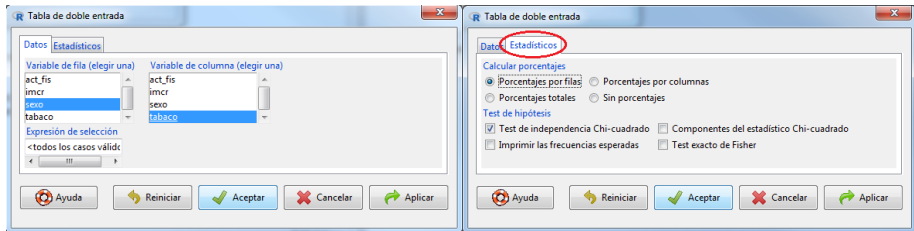
- Tablas de contingencia (de doble entrada).

		tabaco	
sexo	fumador moderado	gran fumador	no fuma
hombre	25 (10 %)	115 (46.2 %)	109 (43.8 %)
mujer	25 (8.9 %)	7 (2.5 %)	248 (88.6 %)

Para obtener una tabla de contingencia de doble entrada con **R Commander** seleccionaremos:

Estadísticos > Tablas de contingencia > Tabla de doble entrada

A continuación indicamos la variable cuyas categorías aparecerán como filas en la tabla y la variable cuyas categorías aparecerán como columnas en la tabla. En la pestaña **Estadísticos** podemos indicar si, además de las frecuencias absolutas, queremos el cálculo de los porcentajes por filas, por columnas o totales. Pulsamos **Aceptar**.



Resultados

R Commander

Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda

Conjunto de datos: **cardio** Editar conjunto de datos Visualizar conjunto de datos Modelo: **<No hay modelo activo>**

R Script R Markdown

```
.table <- xtabs(~sexo + tabaco, data = cardio)
cat("\nFrequency table:\n")
print(.table)
cat("\nRow percentages:\n")
print(rowPercents(.table))
.Test <- chisq.test(.table, correct = FALSE)
print(.Test)
})
```

Salida

Frequency table:

sexo	tabaco	fumador moderado	gran fumador	no fuma
hombre		25	115	109
mujer		25	7	248

Row percentages:

sexo	tabaco	fumador moderado	gran fumador	no fuma	Total	Count
hombre		10.0	46.2	43.8	100	249
mujer		8.9	2.5	88.6	100	280

Pearson's Chi-squared test

data: .table
X-squared = 148.42, df = 2, p-value < 2.2e-16

Mensajes

RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.

Frecuencias absolutas

Porcentajes por filas

Dos variables cualitativas

• Tablas de contingencia (de entradas múltiples).

Hombre

	imcr		
act_fis	Insuficiente	Normal	Sobrepeso
manual blando	24 (26.7 %)	51 (56.7 %)	15 (16.7 %)
manual duro	16 (34.8 %)	23 (50 %)	7 (15.2 %)
seden. con resp	26 (32.9 %)	42 (53.2 %)	11 (13.9 %)
seden. sin resp	10 (25.6 %)	20 (51.3 %)	9 (23.1 %)

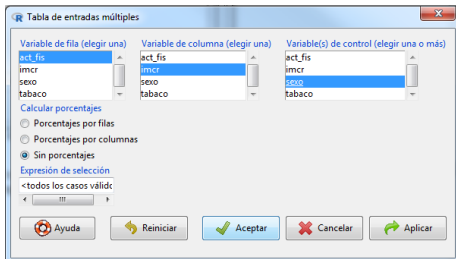
Mujer

	imcr		
act_fis	Insuficiente	Normal	Sobrepeso
manual blando	3 (21.4 %)	7 (50 %)	4 (28.6 %)
manual duro	48 (29.3 %)	61 (37.2 %)	55 (33.5 %)
seden. con resp	2 (20 %)	3 (30 %)	5 (50 %)
seden. sin resp	23 (25.3 %)	37 (40.7 %)	31 (34.1 %)

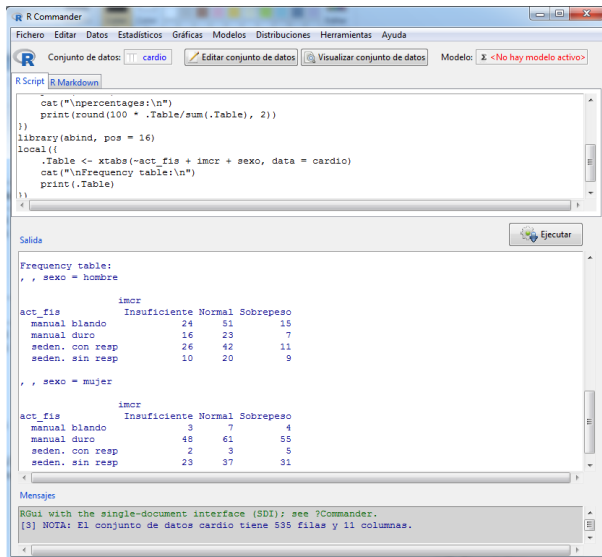
Para obtener una tabla de contingencia de entradas múltiples con **R Commander** seleccionaremos:

Estadísticos > Tablas de contingencia > Tabla de entradas múltiples

A continuación indicamos la variable cuyas categorías aparecerán como filas en la tabla, la variable cuyas categorías aparecerán como columnas en la tabla y la variable de control. Es posible especificar si, además de las frecuencias absolutas, se desean obtener los porcentajes por filas o por columnas. Pulsamos **Aceptar**.



Resultados



The screenshot shows the R Commander application window. The menu bar includes: Archivo, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, Ayuda. The toolbar shows buttons for 'Conjunto de datos: cardio', 'Editar conjunto de datos', 'Visualizar conjunto de datos', and 'Modelo: No hay modelo activo'. The 'R Script' tab is active, displaying the following R code:

```
cat("\npercentages:\n")
print(round(100 * .Table/sum(.Table), 2))
})
library(abind, pos = 16)
local({
  .Table <- xtabs(~act_fis + imcr + sexo, data = cardio)
  cat("\nFrequency table:\n")
  print(.Table)
})
```

The 'Salida' (Output) pane shows the execution results:

Frequency table:
, , sexo = hombre

act_fis	imcr		
	Insuficiente	Normal	Sobrepeso
manual blando	24	51	15
manual duro	16	23	7
seden. con resp	26	42	11
seden. sin resp	10	20	9

, , sexo = mujer

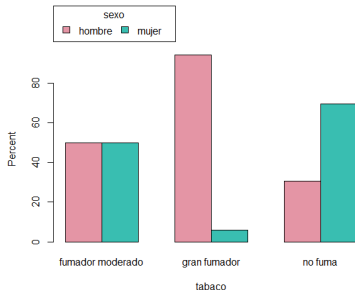
act_fis	imcr		
	Insuficiente	Normal	Sobrepeso
manual blando	3	7	4
manual duro	48	61	55
seden. con resp	2	3	5
seden. sin resp	23	37	31

The 'Mensajes' (Messages) pane shows the following messages:

RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.

Dos variables cualitativas

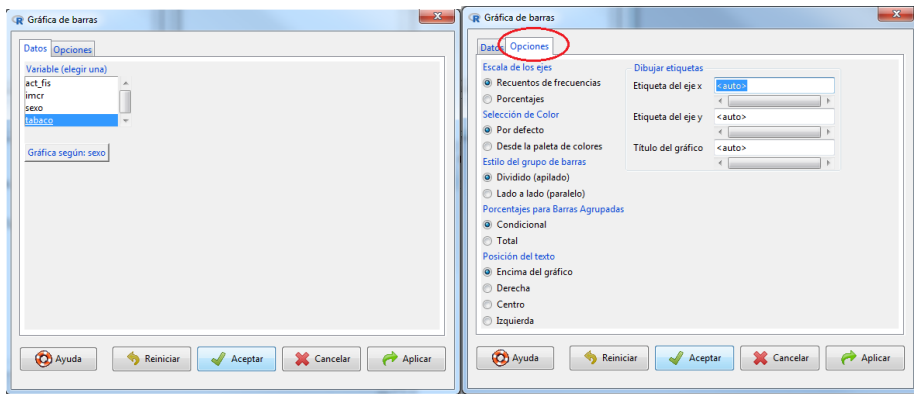
- Gráficos: Diagramas de barras por grupos



Para obtener un diagrama de barras de una variable cualitativa según las categorías de otra variable cualitativa seleccionaremos en **R Commander**:

Gráficas > Gráfica de barras > Gráfica por grupos...

A continuación, seleccionamos la variable para la que queremos obtener el diagrama de barras y la variable “grupo”. En la pestaña **Opciones** podemos personalizar la salida del gráfico (escala de los ejes: frecuencias absolutas/porcentajes, color de las barras, título y etiquetas de los ejes, etc.). Pulsamos **Aceptar**.



- **Resumen numérico por grupos**

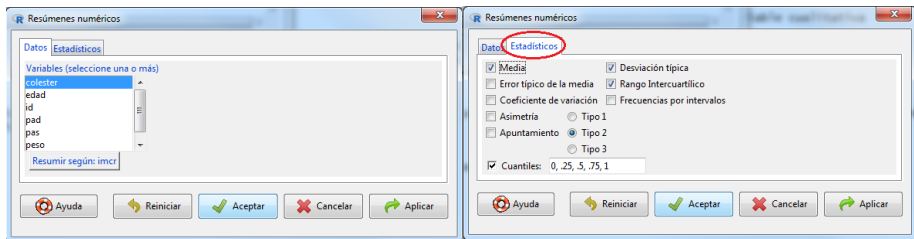
Nivel de colesterol (media y desviación típica) según IMC.

IMC	Media	Desv.típica
Insuficiente	216.28	42.51
Normal	228.04	42.44
Sobrepeso	224.31	40.54

Para obtener un resumen numérico de una variable cuantitativa según las categorías de otra variable cualitativa seleccionaremos en **R Commander**:

Estadísticos > Resúmenes > Resúmenes numéricos > Resumir según:

A continuación, seleccionamos la variable de la que queremos obtener el resumen numérico y la variable “grupo”. Pulsamos **Aceptar**. En la pestaña **Estadísticos** podemos especificar las medidas que deseamos obtener.



Resultados

The screenshot shows the R Commander window with the following components:

- Menu Bar:** Archivo, Editar, Datos, Estadísticos, Gráficas, Modelos, Distribuciones, Herramientas, Ayuda.
- Toolbar:** Includes buttons for 'Conjunto de datos: cardio', 'Editar conjunto de datos', 'Visualizar conjunto de datos', and 'Modelo: <No hay modelo activo>'. There are also tabs for 'R Script' and 'R Markdown'.
- Script Editor:** Contains the following R code:

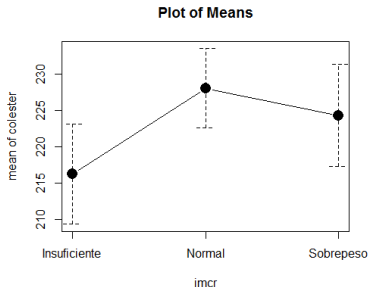
```
load("H:/FISABIO/Curso de Estadística con RCommander/datos/cardio.RData")
library(abind, pos = 16)
library(e1071, pos = 17)
```
- Salida (Output) Panel:** Displays the execution results, including a summary of the 'coleston' variable from the 'cardio' dataset. The output is as follows:

```
> load("H:/FISABIO/Curso de Estadística con RCommander/datos/cardio.RData")
> library(abind, pos = 16)
> library(e1071, pos = 17)

> numSummary(cardio[, "coleston", drop = FALSE], groups = cardio$imcr, statistics = c("mean", "sd",
+ "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
      mean      sd  0%   25%  50%   75% 100% coleston:n coleston:NA
Insuficiente 216.2838 42.51259 133 185.5 215 240.0 341      148      4
Normal       228.0383 42.44578 140 196.5 226 251.5 353      235      9
Sobrepeso    224.3077 40.54280 143 196.0 220 249.5 335      130      7
```
- Mensajes (Messages) Panel:** Shows the following messages:

```
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.
```

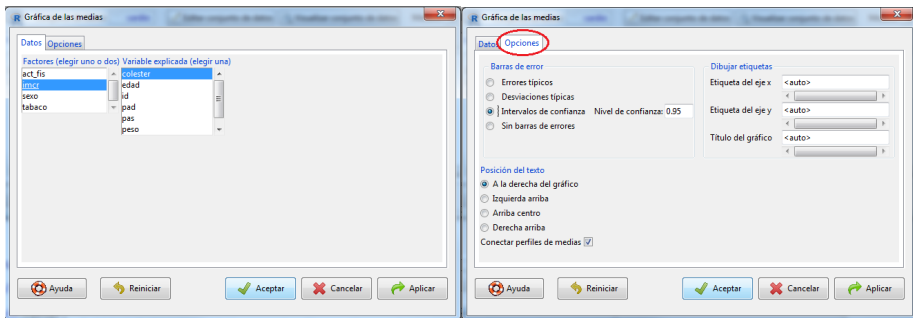
- Gráficos: Gráfico de medias



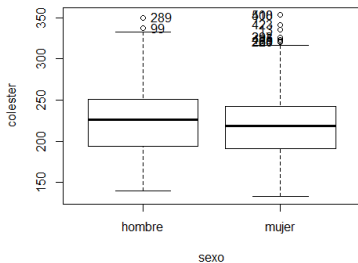
Para obtener el gráfico de medias de una variable cuantitativa del estudio según las categorías de otra variable cualitativa seleccionaremos en **R Commander**:

Gráficas > Gráfica de las medias

A continuación, indicamos como variable explicada la variable cuantitativa y como factor la variable cualitativa. En la pestaña **Opciones** podemos especificar la salida del gráfico (además de la media, mostrar intervalos de confianza, personalizar título del gráfico y etiquetas de los ejes, etc.). Pulsamos **Aceptar**.



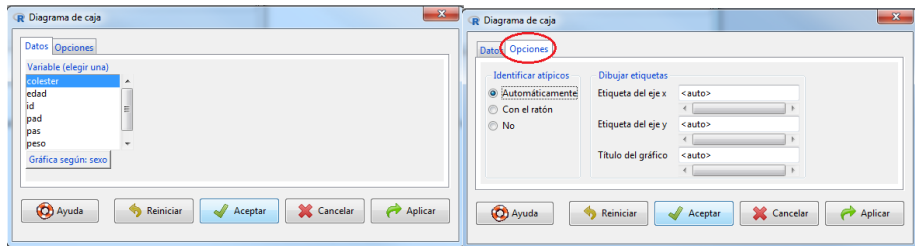
- Gráficos: Diagrama de cajas por grupos



Para obtener el diagrama de cajas de una variable cuantitativa del estudio según las categorías de otra variable cualitativa seleccionaremos en **R Commander**:

Gráficas > Diagrama de caja > Gráfica según:

A continuación, seleccionamos la variable para la que queremos obtener el diagrama de cajas y la variable “grupo”. En la pestaña **Opciones** podemos personalizar el título del gráfico y las etiquetas de los ejes. Pulsamos **Aceptar**.

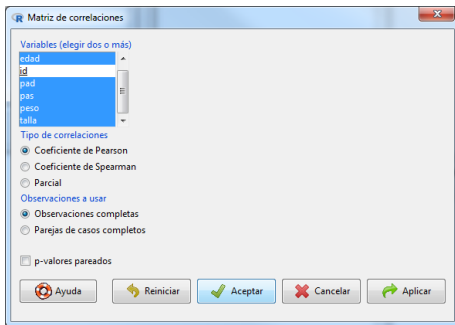


- **Coeficiente de correlación lineal de Pearson (r):** mide la fuerza de la relación lineal entre dos variables cuantitativas continuas. Puede tomar valores entre -1 y 1.
 - Si $r \approx 1$, existe relación lineal positiva entre las variables.
 - Si $r \approx 0$, no existe relación lineal entre las variables.
 - Si $r \approx -1$, existe relación lineal negativa entre las variables.

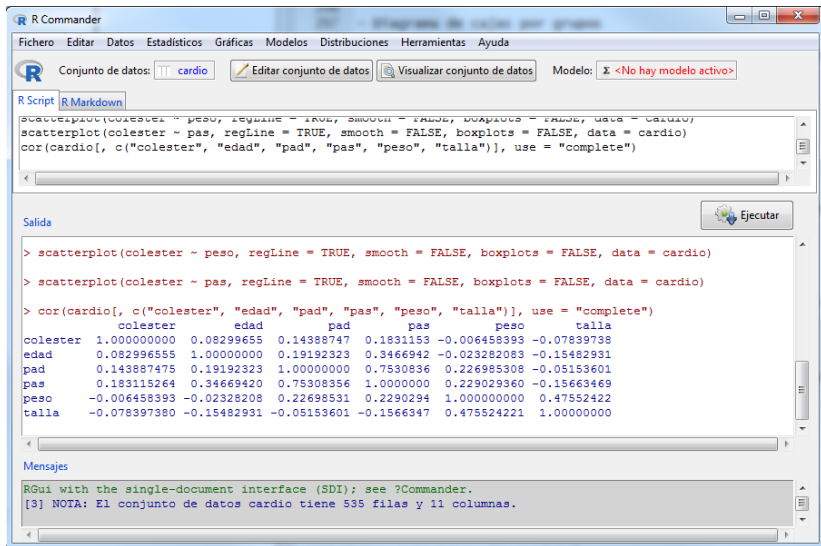
Para obtener el coeficiente de correlación lineal de Pearson entre las variables cuantitativas del estudio seleccionaremos en **R Commander**:

Estadísticos > Resúmenes > Matriz de correlaciones...

A continuación indicamos las variables entre las que queremos obtener el coeficiente de correlación y pulsamos **Aceptar**



Resultados



R Commander

Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda

Conjunto de datos: Editar conjunto de datos Visualizar conjunto de datos Modelo:

R Script R Markdown

```
scatterplot(colester ~ peso, regLine = TRUE, smooth = FALSE, boxplots = FALSE, data = cardio);
scatterplot(colester ~ pas, regLine = TRUE, smooth = FALSE, boxplots = FALSE, data = cardio)
cor(cardio[, c("colester", "edad", "pad", "pas", "peso", "talla")], use = "complete")
```

Salida

Ejecutar

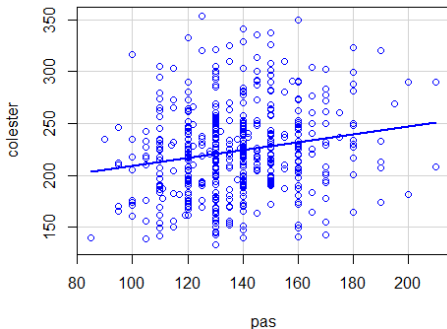
```
> scatterplot(colester ~ peso, regLine = TRUE, smooth = FALSE, boxplots = FALSE, data = cardio)
> scatterplot(colester ~ pas, regLine = TRUE, smooth = FALSE, boxplots = FALSE, data = cardio)
> cor(cardio[, c("colester", "edad", "pad", "pas", "peso", "talla")], use = "complete")
```

	colester	edad	pad	pas	peso	talla
colester	1.000000000	0.08299655	0.14388747	0.1831153	-0.006458393	-0.07839738
edad	0.08299655	1.00000000	0.19192323	0.3466942	-0.023282083	-0.15482931
pad	0.14388747	0.19192323	1.00000000	0.7530836	0.226985308	-0.05153601
pas	0.183115264	0.34669420	0.75308356	1.00000000	0.229029360	-0.15663469
peso	-0.006458393	-0.02328208	0.22698531	0.2290294	1.000000000	0.47552422
talla	-0.078397380	-0.15482931	-0.05153601	-0.1566347	0.475524221	1.000000000

Mensajes

```
RGui with the single-document interface (SDI); see ?Commander.
[3] NOTA: El conjunto de datos cardio tiene 535 filas y 11 columnas.
```

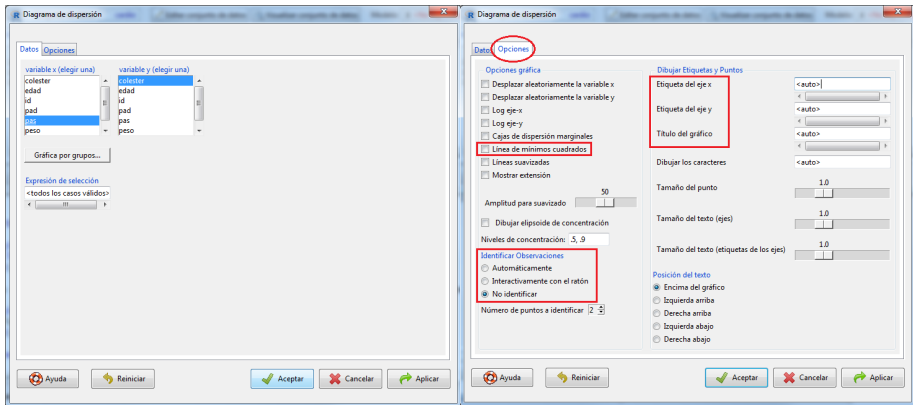
- Diagrama de dispersión



- Coeficiente de correlación lineal de Pearson entre las variables *colester* y *pas*: 0.18.

Para obtener el diagrama de dispersión entre dos variables cuantitativas seleccionaremos en **R Commander**: Gráficas > Diagrama de dispersión

A continuación, indicamos la variable que queremos representar en el “eje x” y la variable que queremos representar en el “eje y”. En la pestaña **Opciones** podemos personalizar la salida del gráfico (representar línea de mínimos cuadrados, identificar observaciones, título del gráfico y etiquetas de los ejes, etc.). Pulsamos **Aceptar**.



- Descripción de la relación entre **dos variables cualitativas**:
 - Tabla de contingencia de doble entrada y de entradas múltiples
 - Diagrama de barras por grupos
- Descripción de la relación entre **una variable cuantitativa y otra variable cualitativa**:
 - Resumen numérico por grupos
 - Gráfico de las medias
 - Diagrama de cajas agrupado
- Descripción de la relación entre **dos variables cuantitativas**:
 - Coeficiente de correlación lineal de Pearson
 - Diagrama de dispersión

Creación de informes con R Markdown

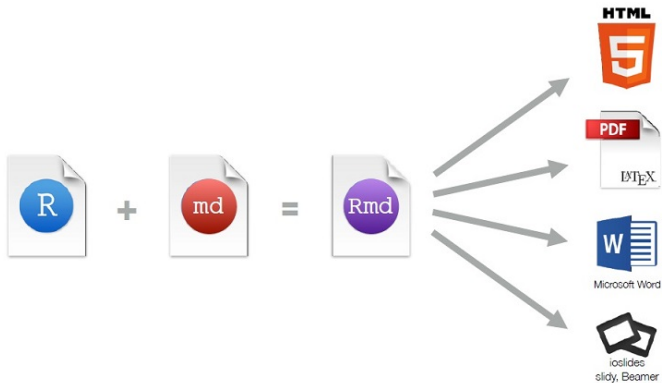
¿Qué es R Markdown?

R Markdown es una herramienta que permite generar informes desde R con los resultados obtenidos tras el análisis estadístico de unos datos.

Un archivo en formato R Markdown combina texto y código R de forma que al procesar el documento, el código se evalúa y los resultados obtenidos se presentan junto al texto en el informe.

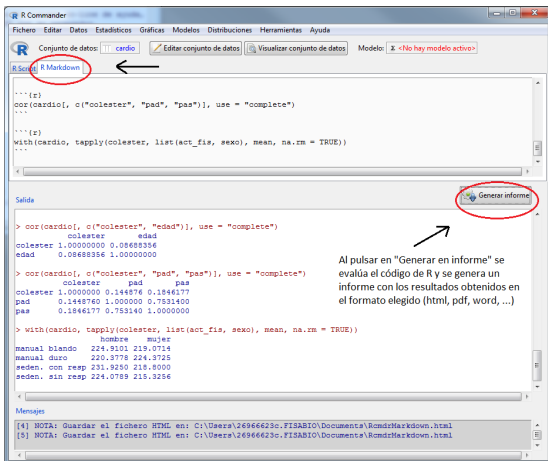
El texto se escribe en lenguaje *Markdown*, una forma sencilla de escribir texto simple (secciones, párrafos, listas, enlaces, imágenes, ...) que se convierte fácilmente a otros formatos (HTML, PDF, Word, ...).

R Markdown



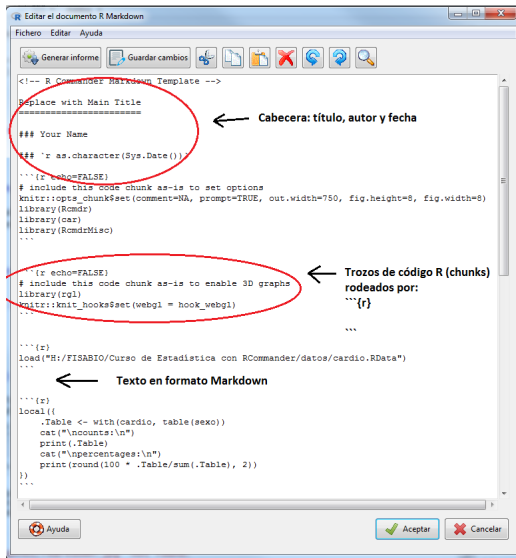
Creación de un documento R Markdown desde R Commander

R Commander incorpora una pestaña en la que se genera automáticamente un documento R Markdown con el código de R necesario para obtener cada uno de los análisis estadísticos solicitados durante la sesión de trabajo.



Componentes de un documento R Markdown

Hay 3 componentes generales en un archivo R Markdown:



Texto en language Markdown:

- Texto en **cursiva** —> Texto en *cursiva*
- Texto en ****negrita**** —> Texto en **negrita**
- Subíndice~2~, superíndice^^2^^ —> Subíndice₂, superíndice²
- ~~Tachado~~ —> Tachado
- Enlace a [R] (<https://cran.r-project.org/>) —> Enlace a R

- Imagen: ![] (figuras/RStudio-Ball.png): —> Imagen:



Texto en language Markdown:

Encabezado 1
Encabezado 2
Encabezado 3

Tablas:

```
Cabecera 1 | Cabecera 2
-----|-----
Celda 1 | Celda 2
Celda 3 | Celda 4
```

Cabecera 1	Cabecera 2
Celda 1	Celda 2
Celda 3	Celda 4

Texto en language Markdown:

Listas desordenadas:

- * Elemento 1
- * Elemento 2
 - Sub-elemento 1
 - Sub-elemento 2

- Elemento 1
- Elemento 2
 - Sub-elemento 1
 - Sub-elemento 2

Listas ordenadas:

1. Elemento 1
2. Elemento 2
 - Sub-elemento 1
 - Sub-elemento 2

- ① Elemento 1
- ② Elemento 2
 - Sub-elemento 1
 - Sub-elemento 2

Personalización de los resultados del código

Podemos personalizar los resultados del código con distintas **opciones** que se añaden como argumentos en la cabecera de los trozos de código.

Ejemplo:

```
```{r, fig.align = 'center', fig.width = 4, fig.height = 3}  
scatterplot(colester ~ pas, regLine = FALSE, smooth = FALSE, boxplots = FALSE, data = cardio)
```
```

Personalización de los resultados del código

En la siguiente tabla se muestran algunos de los argumentos disponibles para personalizar los resultados del código:

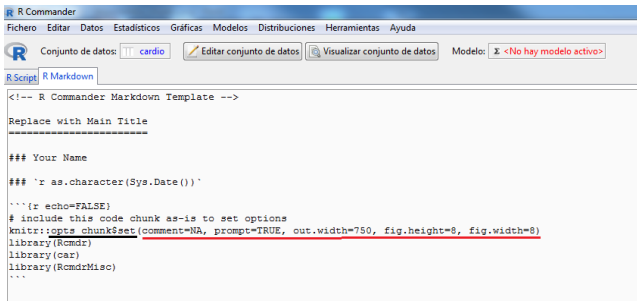
| Argumento | Valor por defecto | Descripción |
|------------|-------------------|---|
| echo | TRUE | Mostrar código en el documento final |
| warning | TRUE | Mostrar advertencias generadas por el código |
| error | TRUE | Mostrar errores generados por el código |
| message | TRUE | Mostrar mensajes generados por el código |
| fig.height | 7 | Altura de las figuras creadas por el trozo de código (en pulgadas) |
| fig.width | 7 | Anchura de las figuras creadas por el trozo de código (en pulgadas) |
| fig.align | 'default' | Alineación de las figuras en el documento ('left', 'right', 'center') |
| tidy | FALSE | Visualizar el código de forma ordenada (espacios entre operadores, separando en varias líneas las líneas de código largas) |
| results | 'markup' | 'markup': mostrar resultados con cierto formato, 'hide': ocultar los resultados en el documento, 'hold': visualizar todos los resultados al final del trozo de código, 'asis': mostrar los resultados tal cual están en R |

Más opciones en: <https://yihui.name/knitr/options/>

Personalización de los resultados del código

Opciones globales:

Pueden establecerse opciones globales que se apliquen a todos los trozos de código del documento R Markdown. Para ello dichas opciones deben especificarse dentro de la función **opts_chunk\$set()** en el primer trozo de código que aparece en el documento R Markdown.



```
<!-- R Commander Markdown Template -->

Replace with Main Title
=====

### Your Name

### `r as.character(Sys.Date())`

```{r echo=FALSE}
include this code chunk as-is to set options
knitr::opts_chunk$set(comment=NA, prompt=TRUE, out.width=750, fig.height=8, fig.width=8)
library(Rcmdr)
library(car)
library(RcmdrMisc)
...

```

# Tablas

Por defecto, R Markdown muestra las tablas con los resultados tal y como aparecen en el programa R.

```
> local({
+ .Table <- with(cardio, table(sexo))
+ cat("\ncounts:\n")
+ print(.Table)
+ cat("\npercentages:\n")
+ print(round(100 * .Table/sum(.Table), 2))
+ })
```

```
counts:
sexo
hombre mujer
 255 280

percentages:
sexo
hombre mujer
 47.66 52.34
```

Existen varias funciones en R que permiten mostrar los resultados mejorando la presentación de las tablas. Por ejemplo, la función `kable()` del paquete `knitr`.

# Tablas

Para mejorar la presentación de las tablas haciendo uso de la función `kable()` debe establecerse el argumento `results='asis'` en la cabecera del trozo de código, cargarse el paquete `knitr` con el comando `library(knitr)` y escribir el código que genera cada tabla dentro de la función `kable()` tal y como se muestra a continuación:

```
> library(knitr)
> local({
+ .Table <- with(cardio, table(sexo))
+ cat("\ncounts:\n")
+ print(kable(.Table))
+ cat("\npercentages:\n")
+ print(kable(round(100 * .Table/sum(.Table), 2)))
+ })
```

counts:

sexo	Freq
hombre	255
mujer	280

percentages:

sexo	Freq
hombre	47.66
mujer	52.34

# ¿Qué nos permite R Markdown?

- **Organizar texto, código y resultados en un mismo documento.**
- **Forma sencilla de dar formato al texto** de nuestro informe mediante el lenguaje *Markdown*.
- **Generación de tablas y figuras automáticamente en el documento**, evitándo tener que escribir e insertar manualmente los resultados.
- **Gran variedad de formatos de presentación de informes:** HTML, Word, PDF, ...

Guía de referencia R Markdown

## Ejercicio práctico