Matemáticas III

Apuntes

Manuel E. Gegúndez Arias Manuel Maestre Hachero

Contents

1 Introducción a R-Commander y manejo de datos 7

1.1 Introducción 7

1.2 Instalación 8

1.2.1 Instalación de R 8
1.2.2 Instalación de la interfaz gráfica R-Commander 8
1.2.3 Instalación conjunta de de **R** y **R-Commander** 9
1.2.4 Instalación de paquetes adicionales 9

1.3 Arranque 10

1.4 La interfaz gráfica R-Commander 11

1.4.1 La barra de menús de R-Commander 12

1.5 Tipos de datos y operadores aritméticos y lógicos 12

1.6 Introducción de datos 13

1.6.1 Introducción de datos en línea de comandos 13
1.6.2 Introducción de datos desde el editor de R-Commander 14
1.6.3 Cambiar el conjunto de datos activo 16
1.6.4 Guardar datos 16

1.7 Abrir ficheros de Datos 17

1.7.1 Ficheros de datos con extensión *rda* o *RData* 17
1.7.2 Leer conjuntos de datos desde paquete adjunto 17
1.7.3 Ficheros de datos de *Excel* 17
1.7.4 Ficheros de datos de *texto plano* (*.txt*) 18
1.7.5 Ficheros de datos de *SPSS* (*.sav*) 19

1.8 Modificación del conjunto de datos 19

1.8.1 Eliminación de datos 19
1.8.2 Cálculo de variables 19
1.8.3 Recodificación de variables 19

1.8.4 Segmentar y recodificar una variable . . . 20
1.8.5 Convertir variable numérica a factor . . . 22
1.8.6 Filtrado de datos . . . 22
1.9 Manipulación de ficheros de resultados e instrucciones 25
1.9.1 Guardar los resultados . . . 25
1.9.2 Guardar las instrucciones . . . 25
1.9.3 Abrir un fichero de instrucciones . . . 26
1.9.4 Guardar el entorno de trabajo . . . 26
1.9.5 Cargar un entorno de trabajo . . . 26
1.10 Ayuda 26
1.11 Ejercicios propuestos 27

2 Estadística Descriptiva Unidimensional . . . 31
2.1 Introducción 31
2.2 Conceptos generales 31
2.3 Tabla de frecuencias. 32
2.4 Representaciones gráficas 35
2.5 Características asociadas a una distribución de frecuencias 37
2.5.1 Medidas de centralización . . . 37
2.5.2 Medidas de localización . . . 38
2.5.3 Medidas de dispersión . . . 38
2.5.4 Medidas de forma . . . 39
2.6 Diagramas de caja 42
2.7 Ejercicios Resueltos 44
2.8 Estadística Descriptiva Bidimensional 47
2.8.1 Conceptos generales . . . 47
2.8.2 Distribuciones marginales . . . 48
2.8.3 Distribuciones condicionadas . . . 49
2.8.4 Covarianza . . . 51
2.8.5 Coeficiente de correlación lineal . . . 51
2.8.6 Regresión . . . 54
2.9 Ejercicio resuelto 55
2.10 Ejercicios propuestos 58

3 Variables Aleatorias y Modelos de Distribuciones . . . 63
3.1 Variables aleatorias discretas 65
3.1.1 Distribución binomial . . . 65
3.1.2 Distribución de Poisson . . . 69
3.2 Variables aleatorias continuas 71
3.2.1 Variable aleatoria Normal . . . 71
3.3 Aproximaciones de variables 76
3.4 Distribución de una combinación lineal 78
3.5 Distribuciones en el muestreo 80
3.5.1 Variable t de Student . . . 80
3.5.2 Variable Chi^2 (χ^2) . . . 81

3.6 Ejercicios propuestos 81

4 Interv. de Confianza. Contrastes de hipótesis 85

4.1 Introducción 85

4.2 Contrastes de hipótesis e intervalos de confianza sobre medias 86

4.2.1 Contraste de hipótesis e intervalo de confianza sobre la media de una población 86

4.2.2 Contraste para la diferencia de medias de poblaciones independientes . . . 88

4.2.3 Contraste para la diferencia de medias para muestras relacionadas 91

4.3 Contrastes de hipótesis e intervalos de confianza sobre varianzas 92

4.3.1 Contraste de hipótesis e intervalo de confianza para el cociente de varianzas de dos poblaciones. 92

4.4 Contraste de hipótesis e intervalo de confianza para la varianza de una población 95

4.5 Ejercicios Propuestos 97

5 Regresión Lineal 99

5.1 Introducción 99

5.2 Correlación 99

5.2.1 Beneficios vs. C.agua 101

5.2.2 Beneficios vs. C.energía 102

5.2.3 Beneficios vs. empleados 102

5.3 Ajuste de la recta de regresión 103

5.3.1 Cálculo de la recta de regresión 103

5.4 Ejercicios Propuestos 105

6 Introducción a la Programación Lineal 107

6.1 Introducción 107

6.2 Algunos ejemplos de aplicación 107

6.2.1 El problema del transporte 107

6.2.2 Problema de planificación de la producción 108

6.3 El método gráfico 109

6.4 Tipos de soluciones óptimas 110

6.5 Resolución de un problema de optimización con R 111

6.5.1 Ejemplos 113

6.6 Ejercicios 114

1. R-Commander y manejo de datos

1.1 Introducción

R es un potente lenguaje de programación que incluye multitud de funciones para la representación el análisis de datos.

La ventajas de R frente a otros programas habituales de análisis de datos, como pueden ser SPSS, SAS, SPlus, Matlab o Minitab, son múltiples:

- Es software libre y por tanto gratuito. Puede descargarse desde la web `http://www.r-project.org/`.
- Es multiplataforma. Existen versiones para Windows, Macintosh, Linux y otras plataformas.
- Está avalado y en constante desarrollo por una amplia comunidad científica que lo utiliza como estándar para el análisis de datos.
- Cuenta con multitud de paquetes para todo tipo de análisis estadísticos y representaciones gráficas, desde los más habituales, hasta los más novedosos y sofisticados que no incluyen otros programas. Los paquetes están organizados y documentados en un repositorio CRAN (**C**omprehensive **R** **A**rchive **N**etwork) desde donde pueden descargarse libremente.
- Es programable, lo que permite que el usuario pueda crear fácilmente sus propias funciones o paquetes para análisis de datos específicos.
- Existen multitud de libros, manuales y tutoriales libres que permiten su aprendizaje e ilustran el análisis estadístico de datos en distintas disciplinas científicas como las matemáticas, la física, la biología, la psicología, la medicina, etc.

Por defecto el entorno de trabajo de R es en línea de comandos, lo que significa que los cálculos y los análisis se relizan mediante comandos o instrucciones que el usuario teclea en una ventana de texto. No obstante, existen distintas interfaces gráficas de usuario que facilitan su uso, sobre todo

para usuarios noveles. Una de las interfaces gráficas de usuario más extendidas es `R-commander`, desarrollada por John Fox, y será la que utilizaremos.

1.2 Instalación

1.2.1 Instalación de R

Linux En la distribución Debian y cualquiera de sus derivadas (Ubuntu, Kubuntu, etc.) basta con teclear en la línea de comandos

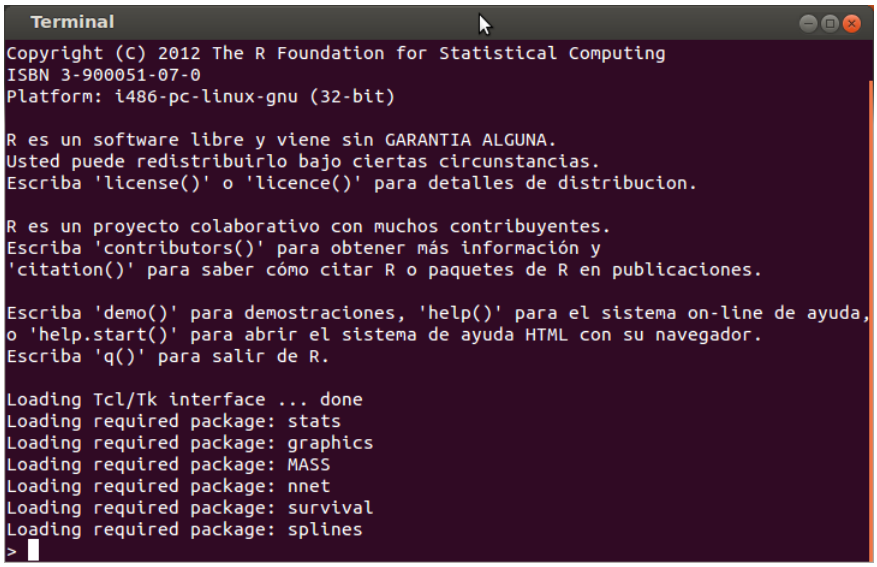
```
> sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc-html r-recommended
```

Windows Descargar de `http://cran.es.r-project.org/bin/windows/base/release.htm` el programa de instalación de R, ejecutarlo y seguir las instrucciones de instalación.

1.2.2 Instalación de la interfaz gráfica R-Commander

La interfaz gráfica de usuario `R-Commander` viene en el paquete **Rcmdr**, que puede instalarse como cualquier otro paquete para `R`.

En Linux, para instalarlo, una vez arrancado `R`, nos situamos en la ventana del terminal de `R`

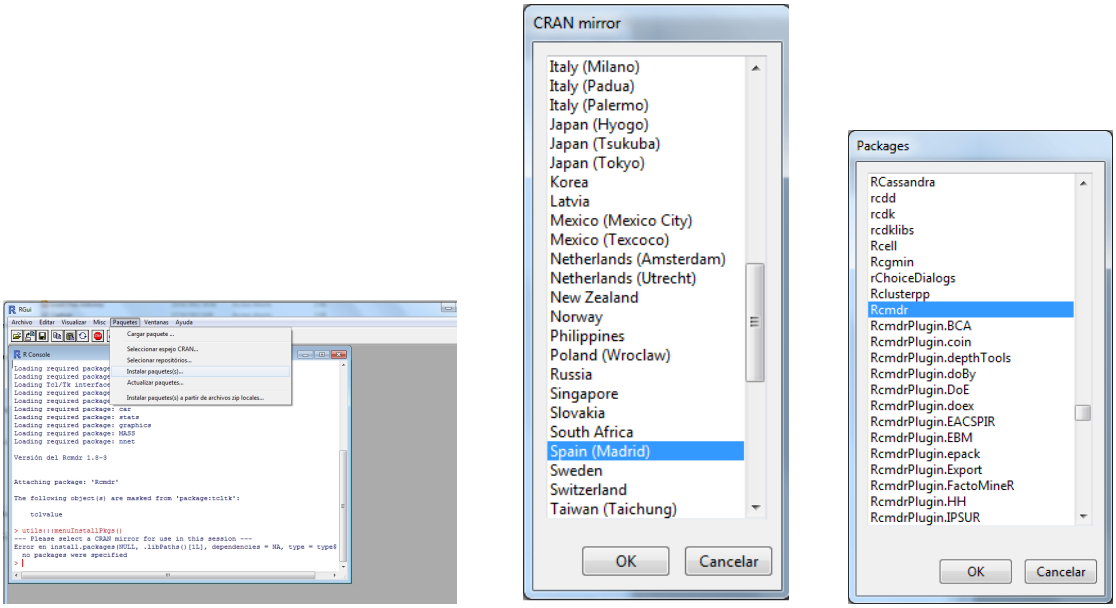


y tecleamos el comando

```
> install.packages("Rcmdr", dependencies=TRUE)
```

Para su instalación en Windows, seleccionamos en el menú superior de R: **Paquetes** → **Instalar Paquetes**(s) tal y como se observa en la Figura 1.1a.

Con esto se nos presentará una lista de *mirrors* desde los que podemos instalar el paquete que seleccionemos. Elejimos el mirror de **Spain (Madrid)** tal y como se puede observar en la Figura 1.1b.



(a) Instalar Paquetes (b) Seleccionar *mirror* (c) Seleccionar Rcmdr

Figure 1.1: Instalación de la interfaz gráfica `Rcmdr`.

Tras pulsar `Aceptar`, nos aparecerá una lista de paquetes. Seleccionamos **Rcmdr** tal y como se observa en la Figura 1.1c.

1.2.3 Instalación conjunta de de R y R-Commander

Una tercera opción es la instalación conjunta de `R` y `R-Commander`. Para ello nos vamos a la página del **Proyecto R-UCA** o `http://knuth.uca.es/R/doku.php?id=inicio` y en la parte de arriba en el centro, pulsamos sobre la versión que queramos descargar: (en la actualidad `Versión 2.14.2 Paquete R-UCA para Windows` o `Versión 14 Paquete R-UCA para Ubuntu`).

Ambos son paquetes autoinstables (**.exe** de Windows y un paquete **.deb** para Ubuntu). Cuentan con la ventaja de instalar en un único paso, `R`, `R-Commander`, y los otros paquetes recomendados, así como nos permite instalar `R` en un ordenador sin conexión a internet. Además, se configura a `R` para que inicie automáticamente `R-Commander` al iniciar `R` y en caso de desinstalación se borran todos los ficheros.

1.2.4 Instalación de paquetes adicionales

La instalación de cualquier otro paquete se realiza del mismo modo que hemos instalado *R-Commander*.

A lo largo del curso utilizaremos con cierta asiduidad los paquetes `HH` e `IPSUR`. Veamos cómo uno de ellos, por ejemplo, `HH`.

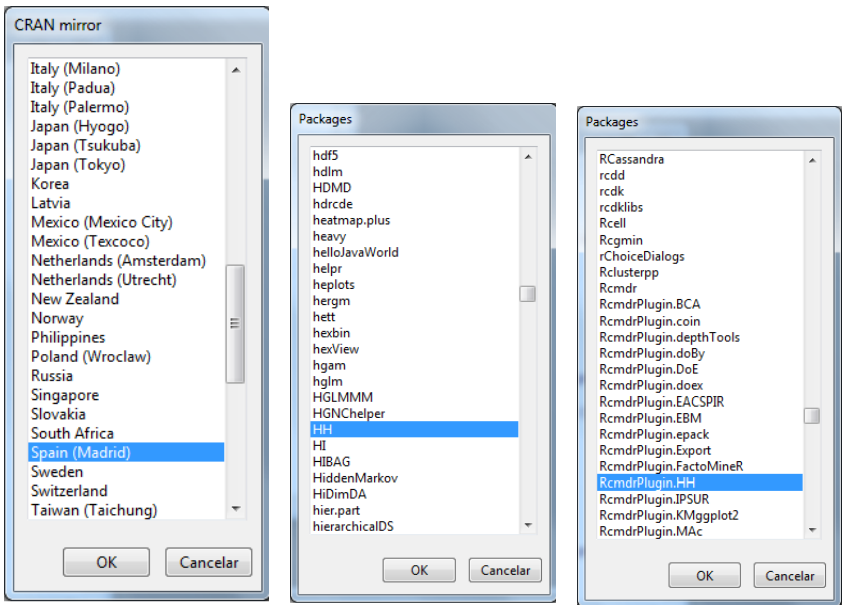
Para instalar `HH` en Linux (junto con el *Plugin* `RcmdrPlugin.HH`), una vez arrancado *R*, hay que teclear el comando

```
> install.packages("HH", dependencies=TRUE)
> install.packages("RcmdrPlugin.HH", dependencies=TRUE)
```

Para su instalaciónn en Windows, seleccionamos en el menú superior de R

Paquetes → **Instalar Paquetes**(s)

Con esto se nos presentará una lista de *mirrors* desde los que podemos instalar el paquete que seleccionemos. Elejimos el mirror de **Spain (Madrid)** tal y como se puede observar en la Figura 1.2a.



(a) Seleccionar *mirror* (b) Seleccionar HH (c) Seleccionar Plugin

Figure 1.2: Instalación del paquete HH y su correspondiente *Plugin*.

Tras pulsar Aceptar, nos aparecerá una lista de paquetes. Seleccionamos **HH** tal y como se observa en la Figura 1.2b. Por último instalamos el Plugin para R-Commander. repetimos la operación anterior y selecionamos RcmdrPlugin.HH tal y como se observa en la Figura 1.2c.

Nota 1.2.1 Una vez instalados los paquetes que deseemos, se escoge en el menú **Paquetes** la opción "**Cargar paquete**". Y elegimos otra vez los paquetes que hemos instalado. La opción **Cargar paquete** pone activos o disponibles para ejecución , los paquetes que tengamos instalados y queramos utilizar en una sesión de R.

Ejercicio 1.1 Instalar el paquete IPSUR junto con su correspondiente Plugin RcmdrPlugin.IPSUR. ■

1.3 Arranque

Como cualquier otra aplicación de Windows, para arrancar el programa hay que hacer click sobre la opción correspondiente del menú **Inicio** → **Programas** → **R**, o bien sobre el icono de escritorio

Una vez arrancado, para arrancar la interfaz gráfica de usuario R-commander hay que cargar el paquete **Rcmdr** y para ello hay que teclear el comando

```
> library(Rcmdr)
```

Para cargar cualquier otro paquete se utiliza el mismo comando, cambiando el nombre del paquete por el deseado.

Si se cierra R-Commander, para volver a cargarlo se utiliza el comando **Commander()**.

> **Nota 1.3.1** En el caso de optar por instalar el paquete R-UCA, al iniciar `R` se iniciará automáticamente `R-Commander`.

1.4 La interfaz gráfica R-Commander

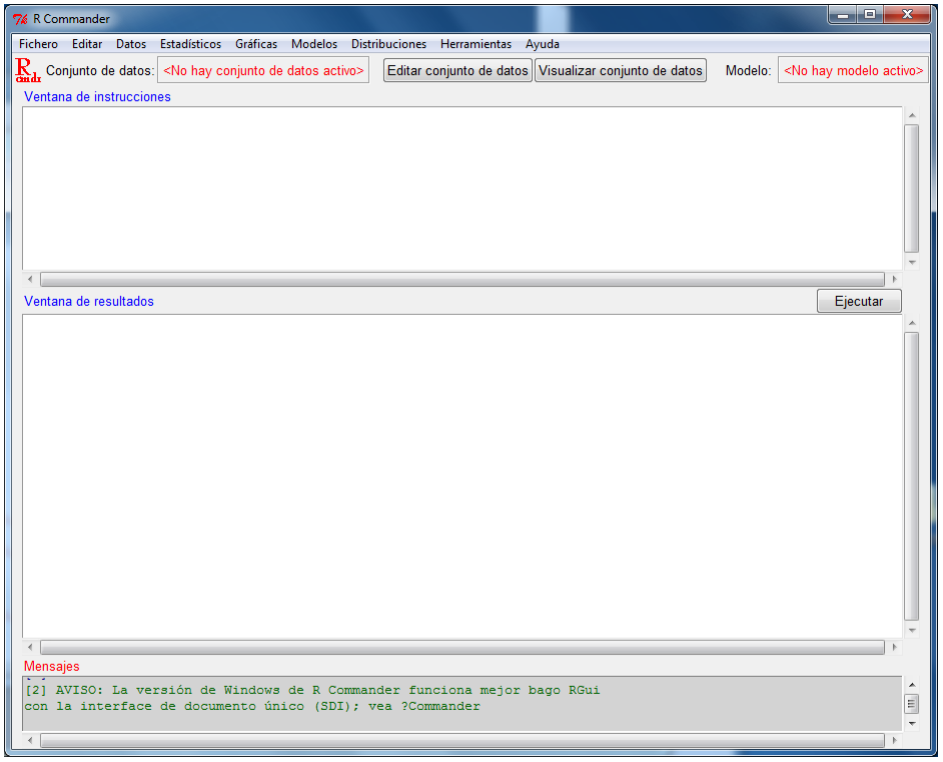


Figure 1.3: Interfaz gráfica de usuario de R-commander.

La interfaz gráfica de usuario R-commander se muestra en la Figura 1.3 por los siguientes elementos:

- **Barra de menús**. Contiene distintos menús con operaciones que pueden realizarse con R.
- **Barra de botones**. Contiene botones para seleccionar, editar o visualizar el conjunto de datos activos y también el modelo activo.
- **Ventana de instrucciones**. Es una ventana de texto similar a la línea de comandos de R donde se pueden introducir directamente comandos para R. Se puede ejecutar un comando o un bloque de comandos situándose en la línea que contiene el comando y pulsando sobre el botón **Ejecutar** o bien tecleando **Ctrl+r**.
  Cada vez que se seleccione un menú que lleve asociado la ejecución de algún comando, dicho comando aparecerá en esta ventana. Esto permite modificar fácilmente los parámetros del comando y volver a ejecutarlo rápidamente sin necesidad de volver al menú.
- **Ventana de resultados**. Es una ventana de texto en la que aparecerán las salidas que generen los comandos que se ejecuten en R. Las instrucciones aparecen en color rojo y los resultados en azul.

- **Mensajes**. Es una ventana de texto donde se muestra información adicional sobre errores, advertencias u otra información auxiliar al ejecutar un comando.

1.4.1 La barra de menús de R-Commander

En la barra de menú de R-Commander aparecen las siguientes opciones.

- **Archivo:** Opciones de menú para cargar y guardar archivos de instrucciones; para guardar resultados y el área de trabajo R; y para salir.
- **Editar:** Opciones de menú (Cortar, Copiar, Pegar, etc.) para editar los contenidos de las ventanas de instrucciones y de resultados. Al pulsar con el botón derecho en la ventana de instrucciones o de resultados también aparece un menú de edición.
- **Datos:** Submenús que contienen opciones de menú para leer y manipular datos.
- **Estadísticos:** Submenús que contienen opciones de menú para una variedad de análisis estadísticos básicos.
- **Gráficas:** Opciones de menú para crear gráficos estadísticos simples.
- **Modelos :** Opciones de menú y submenús para obtener resúmenes, intervalos de confianza, tests de hipótesis, diagnósticos y gráficas para un modelo estadístico, y para añadir cantidades diagnósticas, como residuos, a la serie de datos.
- **Distribuciones:** Probabilidades, cuantiles y gráficos para distribuciones estadísticas estándares (para usarse, por ejemplo, como sustituto de las tablas estadísticas) y ejemplos de estas distribuciones.
- **Herramientas:** Opciones de menú para cargar paquetes R no relacionados con el paquete R-commander (por ejemplo., para acceder a datos guardados en otro paquete), y para establecer algunas opciones.
- **Ayuda:** Opciones de menú para obtener información sobre R-Commander (incluyendo este manual). Además, cada cuadro de diálogo de R-Commander tiene un botón de Ayuda (ver más abajo).

1.5 Tipos de datos y operadores aritméticos y lógicos

En R existen distintos tipos de datos. Los más básicos son:

Numeric : Es cualquier número decimal. Se utiliza el punto como separador de decimales. Por defecto, cualquier número que se teclee tomará este tipo.

Integer : Es cualquier número entero. Para convertir un número de tipo Numeric en un entero se utiliza el comando `as.integer()`

Logical : Puede tomar cualquiera de los dos valores lógicos `TRUE` (verdadero) o `FALSE` (falso).

Character : Es cualquier cadena de caracteres alfanuméricos. Deben introducirse entre comillas. Para convertir cualquier número en una cadena de caracteres se utiliza el comando `as.character()`.

Los valores de estos tipos de datos pueden operarse utilizando distintos operadores o funciones predefinidas para cada tipo de datos. Los más habituales son:

Operadores aritméticos : + (suma), - (resta), * (producto), / (cociente), ^ (potencia).

Operadores de comparación : > (mayor), < (menor), >= (mayor o igual), <= (menor o igual), == (igual), != (distinto).

Operadores lógicos : & (conjunción y), | (disyunción o), ! (negación no).

Funciones predefinidas : `sqrt()` (raíz cuadrada), `abs()` (valor absoluto), `log()` (logarítmo neperiano), `exp()` (exponencial), `sin()` (seno), `cos()` (coseno), `tan()` (tangente).

Al evaluar las expresiones aritméticas existe un orden de prioridad entre los operadores de manera que primero se evaluan las funciones predefinidas, luego las potencias, luego los productos y cocientes, luego las sumas y restas, luego los operadores de comparación, luego las negaciones,

luego las conjunciones y finalmente las disyunciones. Para forzar un orden de evaluación distinto del predefinido se pueden usar paréntesis. Por ejemplo

```
> 2^2+4/2
[1] 6
> (2^2+4)/2
[1] 4
> 2^(2+4/2)
[1] 16
> 2^(2+4)/2
[1] 32
> 2^((2+4)/2)
[1] 8
```

También es posible asignar valores a variables mediante el operador de asignación =. Una vez definidas, las variables pueden usarse en cualquier expresión aritmética o lógica. Por ejemplo,

```
> x=2
> y=x+2
> y
[1] 4
> y>x
[1] TRUE
> x>=y
[1] FALSE
> x==y-2
[1] TRUE
> x!=0 & !y<x
[1] TRUE
```

1.6 Introducción de datos

Antes de realizar cualquier análisis de datos hay que introducir los datos que se quieren analizar.

1.6.1 Introducción de datos en línea de comandos

Existen muchas formas de introducir datos en R pero aquí sólo veremos las más habituales. La forma más sencilla de introducir datos es crear un vector de datos, mediante el comando `c()`. Por ejemplo, para introducir las notas de 5 alumnos se debe teclear

```
> nota = c(5.6,7.2,3.5,8.1,6.4)
```

Esto crea el vector nota con el que posteriormente se pueden realizar cálculos como por ejemplo la media

```
> mean(nota)
[1] 6.16
```

Otra forma habitual de introducir los datos de una muestra es crear un conjunto de datos mediante el comando `data.frame()`. Por ejemplo, para crear un conjunto de datos a partir de las notas anteriores, hay que teclear

```
> curso = data.frame(nota)
```

Esto crea una matriz de datos en la que cada columna se corresponde con una variable y cada fila con un individuo de la muestra. En el ejemplo la matriz curso sólo tendría una columna que se correspondería con las notas y 5 filas, cada una de ellas correspondiente a un alumno de la muestra. Es posible acceder a las variables de un conjunto de datos con el operador dolar $. Por ejemplo, para acceder a las notas hay que teclear

```
> curso$nota
[1] 5.6 7.2 3.5 8.1 6.4
```

Es fácil añadir nuevas variables a un conjunto de datos, pero siempre deben tener el mismo tamaño muestral. Por ejemplo, para añadir una nueva variable con el grupo (mañana o tarde) de los alumnos, hay que teclear

```
> curso$grupo = c("m","t","t","m","m")
```

Ahora el conjunto de datos curso tendría dos columnas, una para la nota y otra para el grupo de los alumnos. Tecleando el nombre de cualquier objeto, se muestra su información:

```
> curso
   nota   grupo
1   5.6      m
2   7.2      t
3   3.5      t
4   8.1      m
5   6.4      m
```

Cuando se introducen datos se puede utilizar el código `NA` (not available), para indicar la ausencia del dato.

Las variables definidas en cada sesión de trabajo quedan almacenas en la memoria interna de R. Es posible obtener un listado de todos los objetos almacenados en la memoria mediante los comandos `ls()`. Si se desea más información, el comando `ls.str()` además de mostrar los objetos de la memoria indica sus tipos y sus valores.

```
> ls()
[1] "curso" "nota"  "x"     "y"
> ls.str()
curso : 'data.frame':   5 obs. of  2 variables:
 $ nota : num  5.6 7.2 3.5 8.1 6.4
 $ grupo: chr  " m " " t " " t " " m " ...
nota :  num [1:5] 5.6 7.2 3.5 8.1 6.4
x :  num 2
y :  num 4
```

Para eliminar un objeto de la memoria se utiliza el comando `rm()`.

```
> ls()
[1] "curso" "nota"  "x"     "y"
> rm(x,y)
> ls()
[1] "curso" "nota"
```

1.6.2 Introducción de datos desde el editor de R-Commander

Para introducir los datos desde `R-Commander` hay que ir al menu **Datos** → **Nuevo conjunto de datos**.

Con esto aparecerá una ventana donde hay que darle un nombre al conjunto de datos (por ejemplo **Notas**) y tras esto aparece la ventana de la figura 1.4 con una tabla en la que se pueden introducir los datos de la muestra parecido a una hoja de cálculo. Cada variable debe introducirse en una columna y cada individuo en una fila.

En la pantalla resultante, figura 1.4(a) (Editor de datos):

- Pulsamos el nombre de la primera variable, **V1**, y escribimos en su lugar el nombre de primera variable, **Nota**, como se visualiza en la Figura 1.5.

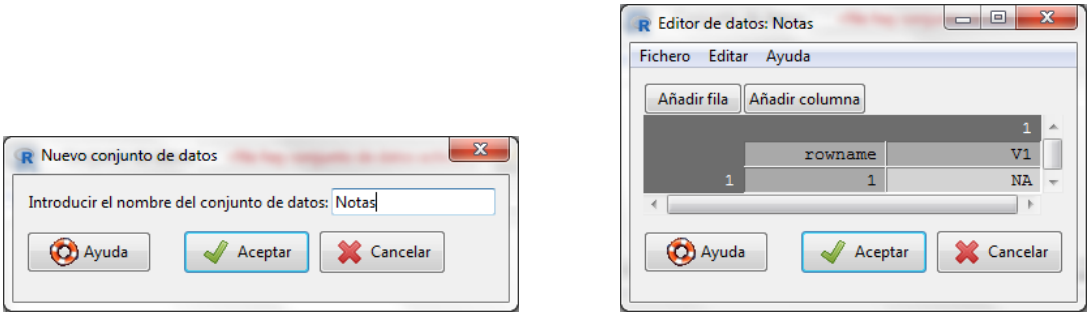


Figure 1.4: Ventana de introducción de datos (I)

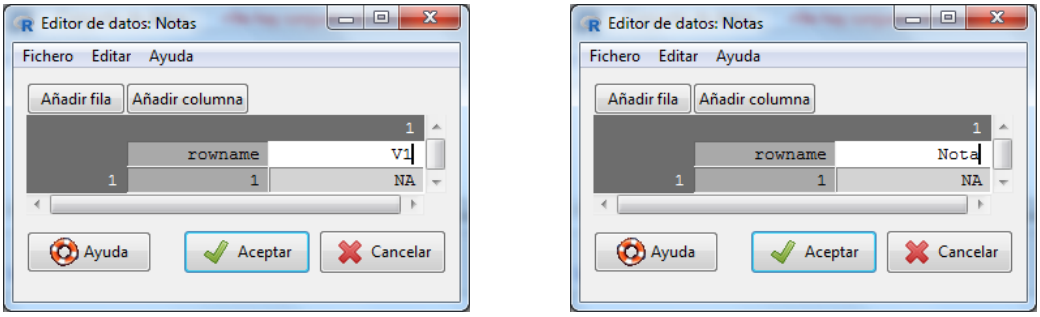


Figure 1.5: Ventana de introducción de datos (II)

- A continuación podemos pulsar **Añadir Fila**, para ir añadiendo los datos, hasta introducir todos los datos de esta variable, o bien pulsamos **Añadir Columna** para añadir una nueva variable y posteriormente pulsamos **Añadir Fila** para ir añadiendo los datos de las dos variables al mismo tiempo, hasta introducirlos todos, como puede observarse en la Figura 1.6.

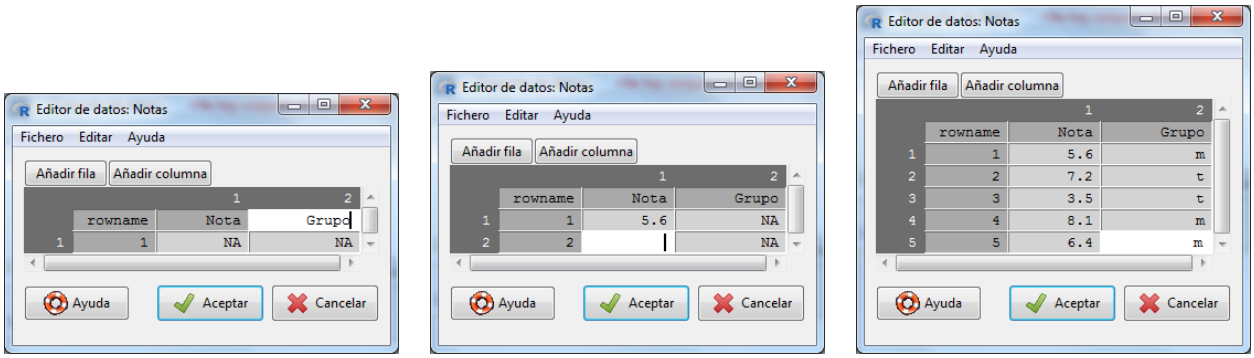


Figure 1.6: Ventana de introducción de datos (III)

Los **nombres de variables** deben comenzar con una letra o un punto y pueden contener cualquier letra, punto, subrayado (_) o número. En particular, **no se pueden utilizar espacios en blanco**. Además, R distingue entre mayúsculas y minúsculas.

Otra forma habitual de introducir datos, que utilizaremos a menudo en las prácticas, es por medio de una hoja de cálculo de Excel o un fichero de texto. Para ello basta con introducir los datos en la hoja de cálculo (o texto plano) siguiendo los mismos criterios de antes. Es conveniente poner los nombres de las variables en la primera fila de la hoja. Una vez guardada la hoja de cálculo es posible cargarla en R como se indica en la sección de *Abrir Datos*.

1.6.3 Cambiar el conjunto de datos activo

R permite definir más de un conjunto de datos en una misma sesión de trabajo, pero sólo puede haber uno activo en cada momento. Para cambiar el conjunto de datos activo se utiliza el menú **Datos** → **Conjunto de datos activo** → **Seleccionar conjunto de datos activo**.

Si tenemos, por ejemplo, dos conjuntos de datos activo (**Notas** y **EjemploDatos**) basta pulsar sobre el conjunto de datos activos (Figura 1.7(a)) y se nos abre una ventana desplegable con todos los conjuntos de datos que tenemos cargados en la memoria de RCommander, seleccionamos uno de ellos (Figura 1.7(b)), le damos a aceptar y ya tenemos activo el nuevo conjunto de datos (Figura 1.7(c)).

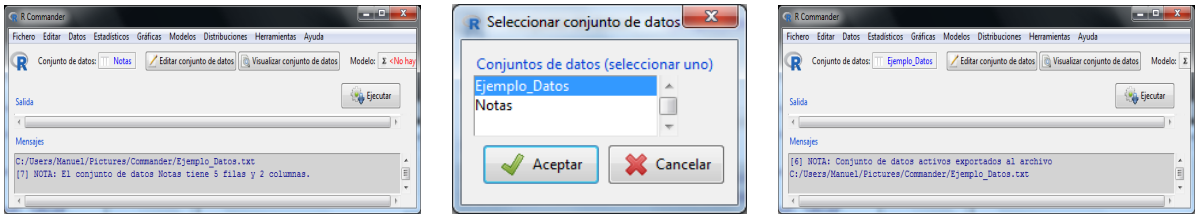


Figure 1.7: Cambiar conjunto de datos activos

1.6.4 Guardar datos

Una vez introducidos los datos, conviene guardarlos en un fichero para no tener que volver a introducirlos en futuras sesiones. Para guardar el conjunto de datos activo en un fichero, se utiliza el menú **Datos** → **Conjunto de datos activo** → **Guardar el conjunto de datos activo**.

Con esto aparece una ventana donde hay que darle un nombre al fichero y seleccionar la carpeta donde se guardará. Los conjuntos de datos se guardan siempre en ficheros de R con extensión **rda** o **RData**.

También es posible guardar los datos en un fichero de texto plano mediante el menú **Datos** → **Conjunto de datos activo** → **Exportar el conjunto de datos activo**.
Tras esto aparece una venta donde se puede indicar entre otras cosas el separador de los datos, que puede ser un espacio, tabuladores, comas u otro caracter, y tras aceptar aparece otra ventana donde hay que darle un nombre al fichero de texto y seleccionar la carpeta donde se guardará.

1.7 Abrir ficheros de Datos

1.7.1 Ficheros de datos con extensión *rda* o *RData*

Si los datos con los que se pretende trabajar ya están guardados en un fichero de R, entonces tendremos que abrir (**cargar**) dicho fichero. Para ello se utiliza el **Datos** → **Cargar conjunto de datos** y en la ventana que aparece se selecciona el fichero que se desea abrir. Automáticamente se cargará el conjunto de datos del fichero y pasará a ser el conjunto de datos activo.

1.7.2 Leer conjuntos de datos desde paquete adjunto

A continuación, abriremos uno de los ficheros de datos que vienen en la distribución de R.

Supongamos, por ejemplo, que queremos trabajar con un fichero llamado **Chile**, que está en el paquete **car**. Dicho fichero contiene datos de individuos de Chile para los cuales se indica: *sexo*, *edad*, *nivel educativo*, *ingresos*, posicionamiento sobre el *status quo* político, *opción de voto*, *región de residencia* y *población de la misma*.

Para abrir dicho fichero fichero de datos, desde R-Commander usamos el menú **Datos** → **Conjunto de de datos en paquetes** → **Leer conjunto de datos desde paquete adjunto**, y seleccionar el conjunto de datos que queremos estudiar tal y como se observa en la Figura 1.8(a).

(a) (b)

Figure 1.8: Leer conjunto de datos desde paquete adjunto.

Para obtener información sobre el conjunto de datos, pulsamos sobre **Ayuda** → **Ayuda sobre el conjunto de datos activo(si es posible)**, y nos da información sobre lo que representa cada variable, como se observa en la Figura 1.8(b).

1.7.3 Ficheros de datos de *Excel*

También es posible cargar datos de ficheros con otros formatos, como por ejemplo una hoja de cálculo Excel. Para ello se utiliza el menú **Datos** → **Importar datos** → **desde Excel** y en la ventana que aparece se selecciona el fichero con la hoja de cálculo Excel que se desea abrir. Si el fichero sólo contiene una hoja se cargará automáticamente, mientras que si tiene varias, aparecerá una ventana en la que habrá que seleccionar la hoja que se quiere abrir.

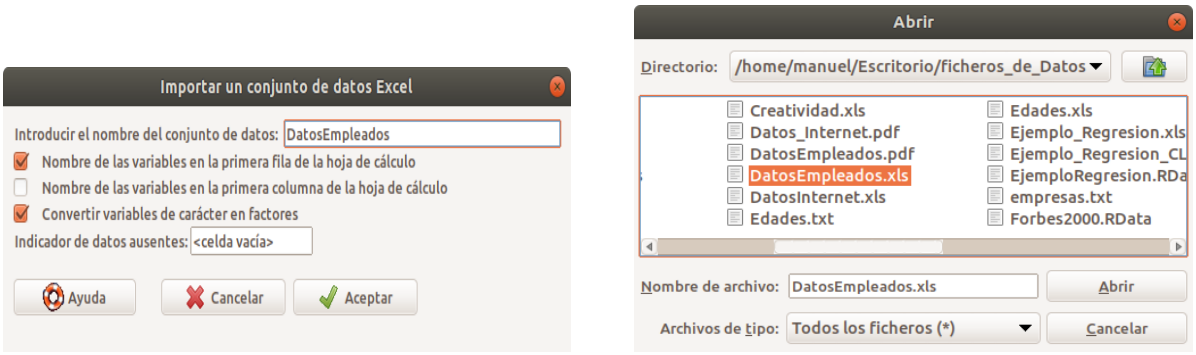


Figure 1.9: **Importar** datos desde un fichero Excel

1.7.4 Ficheros de datos de *texto plano* (*.txt*)

Para abrir un fichero de texto `.txt` se utiliza el menú **Datos** → **Importar datos** → **desde archivo de texto, URL o portapepeles**.

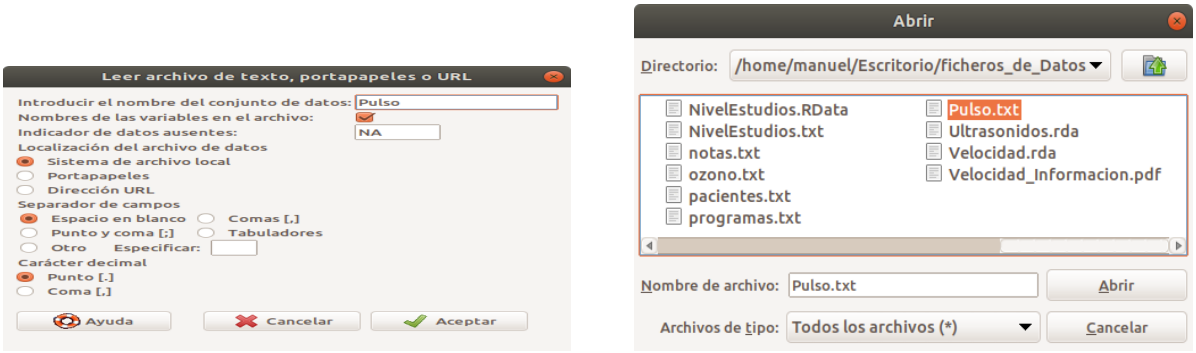


Figure 1.10: **Importar** fichero de datos en texto plano

1.7.5 Ficheros de datos de *SPSS* (*.sav*)

En el caso de ser un fichero de SPSS (extensión `.sav`), utilizaremos **Datos** → **Importar datos** → **desde SPSS**.

Figure 1.11: **Importar** datos desde un fichero de SPSS

1.8 Modificación del conjunto de datos

A menudo en los análisis hay que realizar transformaciones en los datos originales. A continuación se presentan las transformaciones más habituales.

1.8.1 Eliminación de datos

Para eliminar una variable del conjunto de datos se utiliza el menú **Datos** → **Modificar variables del conjunto de datos activo** → **Eliminar variables del conjunto de datos** y en el cuadro de diálogo que aparece basta con seleccionar la variable del conjunto de datos activo que se quiere eliminar.

Para eliminar individuos del conjunto de datos se utiliza el menú **Datos** → **Conjunto de datos activo** → **Borrar fila(s) del conjunto de datos activo** y en el cuadro de diálogo que aparece hay que indicar los números de las filas que se desean eliminar.

1.8.2 Cálculo de variables

Para calcular una nueva variable a partir de las ya existentes en el conjunto de datos activo se utiliza el menú **Datos** → **Modificar variables del conjunto de datos activo** → **Calcular una nueva variable**. Con esto aparece un cuadro de diálogo en la que hay que indicar el nombre de la nueva variable y la expresión a partir de la que se calculará la nueva variable. Aquí puede introducirse cualquier expresión aritmética o lógica de R, y dentro de las expresiones puede utilizarse cualquiera de las variables del conjunto de datos activo. Por ejemplo, para eliminar los decimales de la variable `nota` podría crearse una nueva variable `puntuacion` multiplicando por 10 las notas, tal y como se muestra en la Figura 1.12.

1.8.3 Recodificación de variables

Otra transformación habitual es la recodificación de variables que permite transformar los valores de una variable de acuerdo a un conjunto de reglas de reescritura. Normalmente se utiliza para convertir una variable numérica en una variable categórica que pueda usarse como un factor.

Para recodificar una variable se utiliza el menú **Datos** → **Modificar datos del conjunto de datos activo** → **Recodificar variable**. Con esto aparece una ventana en la que hay que seleccionar

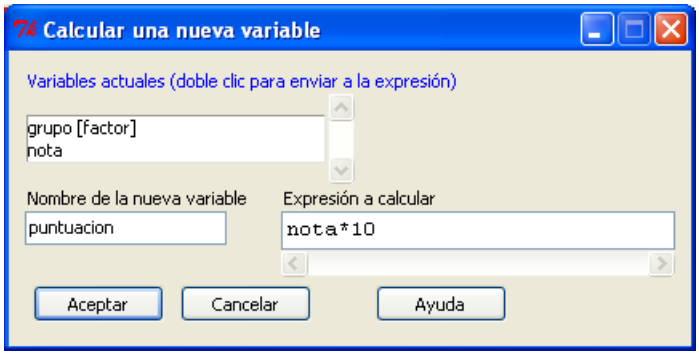


Figure 1.12: Ventana de cálculo de nuevas variables.

la variable que se desea recodificar, introducir el nombre de la nueva variable recodificada e introducir las reglas de recodificación en el cuadro de texto **Introducir directrices de recodificación**. Las reglas de recodificación siempre siguen la sintaxis `valor o rango de valores = nuevo valor` y pueden introducirse tantas reglas como se desee, cada una en una línea. Al lado izquierdo de la igualdad puede introducirse un único valor, varios valores separados por comas, o un **rango de valores** indicando el **límite inferior** y el **límite superior** del intervalo **separados por el operador** "**:**". Por ejemplo, para recodificar la variable `nota` en categorías correspondientes a las calificaciones ([0-5) Suspenso, [5,7) Aprobado, [7,9) Notable y [9,10] Sobresaliente), habría que introducir las reglas que se muestran en la Figura 1.13.

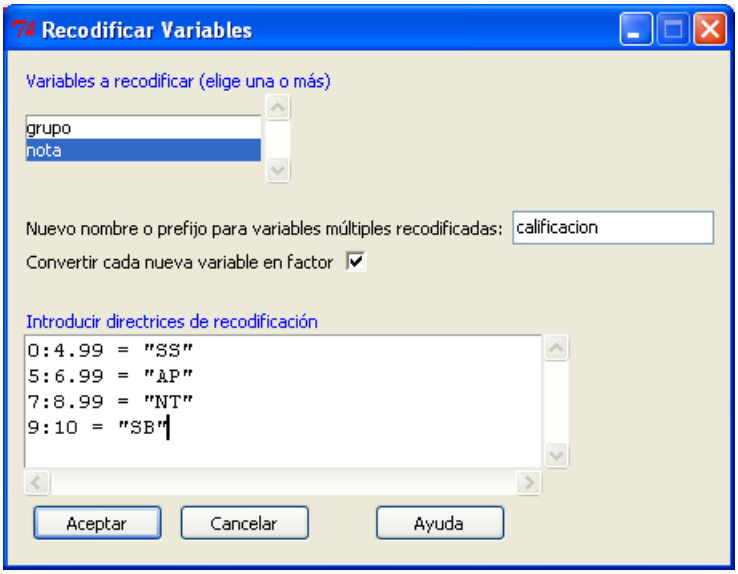


Figure 1.13: Ventana de recodificación de variables

1.8.4 Segmentar y recodificar una variable

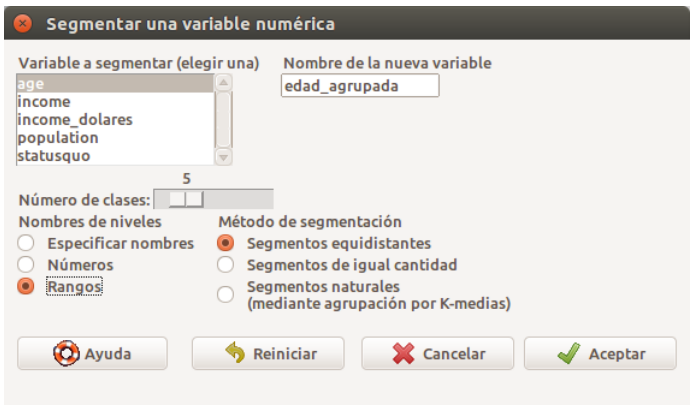
Queremos ahora **agrupar** los datos de la variable **edad** en 5 intervalos de igual amplitud, identificando cada intervalo con su **marca de clase**. Para ello, pulsamos sobre

Datos → **Modificar variables del conjunto de datos activo** → **Segmentar variable numérica**

Seleccionamos la variable **edad**, marcamos que el número de intervalos es 5 y le ponemos nombre a la variable de salida (**edad_agrupada**). Hemos de especificar en *Nombres de niveles* la opción

Rangos y en *Métodos de segmentación* la opción **Segmentos equidistantes**.

Con esto, `R-Commander` nos dividirá el conjunto de datos en 5 intervalos de igual amplitud, aunque no los habrá identificado con su marca de clase.



Esta variable creada (`edad_agrupada`) es una *variable cualitativa*. Para identificar los intervalos con su marca de clase (punto medio del intervalo), primero tendremos que saber cuáles son los intervalos. Para ello hay que calcular una tabla de frecuencias en la que se observen los intervalos:

Estadísticos → **Resúmenes** → **Distribuciones de frecuencias**

de la siguiente manera:

El resultados es (sólo visualizamos aquí las frecuencias absolutas:

```
edad_agrupada
(17.9,28.4] (28.4,38.8] (38.8,49.2] (49.2,59.6] (59.6,70.1]
        869         635         528         345         322
```

Una vez que tenemos los intervalos, hay que calcularles su marca de clase (punto medio del intervalo). Para ello, hay que **recodificar** la variable con los valores de las marcas de clase, para ello, haremos

Datos → **Modificar variables del conjunto de datos activo** → **Recodificar Variables**.

Seleccionamos la variable **edad_agrupada**, y la damos nombre a la variable recodificada con las marcas de clase (por ejemplo **edad_marca**). **Desmarcamos** la opción **Convertir cada nueva variable en factor**. En el recuadro **Introducir directrices de recodificación**, a cada intervalo, le asociamos su punto medio (sumamos ambos extremos y dividimos por 2), como sigue:

Nota 1.8.1 *Ya habíamos visto otra forma de recodificar variables, más concretamente en la Figura* **??** .

1.8.5 Convertir variable numérica a factor

Cuando una variable numérica representa, en realidad, a una variable cualitativa en la que los números son códigos correspondientes a las modalidades, es necesario indicarlo al programa (ya que éste no puede discernir si los números son cantidades o códigos).

En el fichero **Pulso.txt** (cargar los datos como se explica en el apartado 1.7.4, la variable **Smokes** tiene dos valores

1: si el paciente fuma
2: si el paciente no fuma

Si queremos convertir esta variable a factor utilizamos el menú **Datos** → **Modificar variables** → **Convertir variable numérica en factor**. Elegimos la variable que queremos convertir en factor (**Smokes**), marcamos la opción `Asignar nombres a los niveles` y escribimos los nombres que queramos, por ejemplo, "*Fuma*" y "*No Fuma*", tal como se indica en la Figura 1.14.

Figure 1.14: Convertir una variable numérica a factor.

1.8.6 Filtrado de datos

Cuando se desea realizar un análisis con un subconjunto de individuos del conjunto de datos activo que cumplen una determinada condición, o bien con un subconjunto de variables, es posible filtrar el conjunto de datos activo para quedarse con esos individuos y esas variables. Para ello se utiliza el menú **Datos** → **Conjunto de datos activo** → **Filtrar el conjunto de datos activo**. Con esto aparece un cuadro de diálogo en el que hay que indicar las variables que se desean y en el

cuadro **Expresión de selección** indicar la condición lógica que tienen que cumplir los individuos seleccionados. También se puede indicar el nombre del nuevo conjunto de datos ya que si se omite se sobreescribirá el conjunto de datos activo.

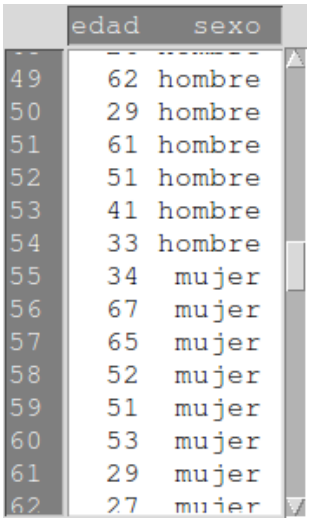
Por ejemplo, supongamos que el fichero de datos **Edades.txt** contiene datos que corresponde a las edades de un grupo de hombres y mujeres.

Supongamos que queremos tener ficheros separados: uno con las edades de los hombres y otro con las edades de las mujeres.

En primer lugar nos situamos en el directorio donde queramos trabajar.

Fichero → **Cambiar directorio de trabajo** y seleccionamos el directorio donde queremos trabajar.

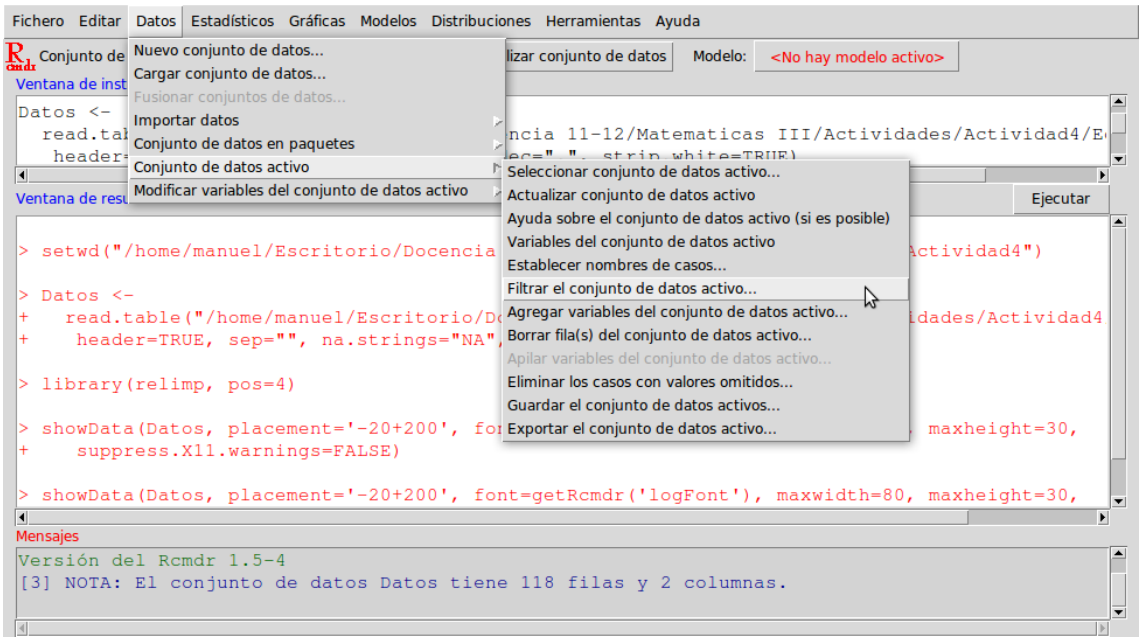
Posteriormente, cargamos el fichero de datos: **Datos** → **Importar datos** → **Desde fichero de texto** y seleccionamos nuestro fichero (**Edades.txt**). Al visualizar el fichero de datos, se tiene



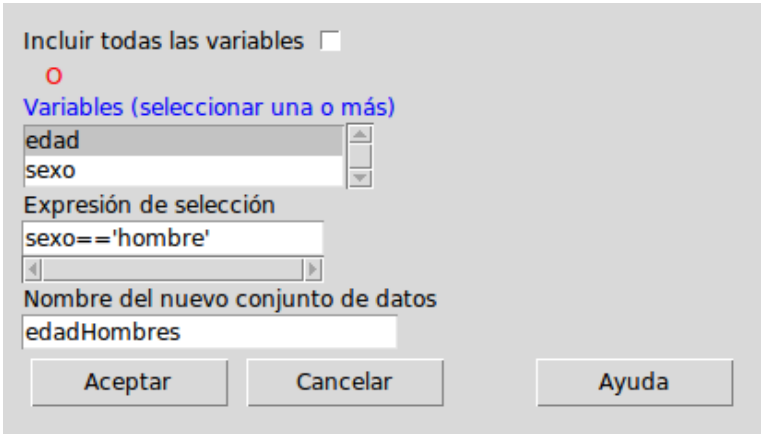
	edad	sexo
49	62	hombre
50	29	hombre
51	61	hombre
52	51	hombre
53	41	hombre
54	33	hombre
55	34	mujer
56	67	mujer
57	65	mujer
58	52	mujer
59	51	mujer
60	53	mujer
61	29	mujer
62	27	mujer

Vamos a filtrar este conjunto de datos y vamos a crear un nuevo conjunto de datos llamado **edadHombres**. Para ello seleccionamos **Datos** → **Conjunto de datos activos** → **Filtrar el**

conjunto de datos activo, tal como aparece en la siguiente Figura



En la ventana emergente, *desmarcamos* la opción **Incluir todas las variables**, seleccionamos la variable **edad** y en la *expresión de selección*, ponemos **sexo==’hombre’**, y le damos nombre al conjunto (**edadHombres**) tal y como aparece en la siguiente Figura



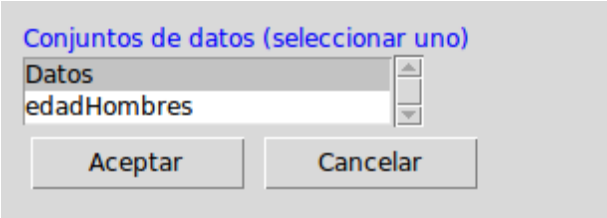
Al pulsar `Aceptar` vemos que ha cambiado el conjunto de datos activo a **edadHombres**. Podemos visualizar dicho conjunto de datos y vemos que ya sólo contiene los datos de la variable

edad, y contiene sólo las filas correspondientes a los hombres.

	edad
1	43
2	35
3	66
4	44
5	38
6	67
7	55
8	56
9	75
10	64
11	33

Si quiesiéramos ahora construir el conjunto de datos fichero **edadMujeres**, primero hemos de volver a activar el conjunto de datos **Datos**, que contenía las edades de hobres y mujeres.

Para ello seleccionamos **Datos** → **Conjunto de datos activos** → **Seleccionar el conjunto de datos activo** y en la ventana emergente elegimos el conjunto de datos **Datos** como se muestra en la Figura



Ya sólo basta repetir el procedimiento anterior para las edades de las mujeres.

1.9 Manipulación de ficheros de resultados e instrucciones

1.9.1 Guardar los resultados

Para guardar el contenido de la ventana de resultados en un fichero de texto plano se utiliza el menú **Fichero** → **Guardar los resultados**. Con esto aparece un cuadro de diálogo en el que hay que indicar el nombre del fichero y la carpeta donde se desea guardar.

También es posible copiar los resultados para después pegarlos en un procesador de texto como Word.

1.9.2 Guardar las instrucciones

Las instrucciones de la ventana de instrucciones también pueden guardarse en un fichero de texto plano mediante el menú **Fichero** → **Guardar las instrucciones** e indicando el nombre del fichero y la carpeta donde se guardará en el cuadro de diálogo que aparece. Esto opción es muy interesante de cara a repetir análisis o automatizar tareas repetitivas.

1.9.3 Abrir un fichero de instrucciones

Para abrir un fichero de texto con instrucciones de R se utiliza el menú **Fichero** → **Abrir fichero de instrucciones** y después seleccionar el fichero que se desea abrir en el cuadro de diálogo que aparece. Con esto las instrucciones del fichero se aparecen en la ventana de instrucciones y pueden seleccionarse y ejecutarse.

1.9.4 Guardar el entorno de trabajo

Durante cualquier sesión de trabajo, `R` guarda en memoria las variables o los modelos que se definan. Cuando cierra el programa esta información se pierde, de manera que si se desea continuar con un análisis conviene guardar el contenido del entorno de trabajo antes de acabar la sesión de trabajo. Para ello se utiliza el menú **Fichero** → **Guardar el entorno de trabajo de R** y como siempre hay que indicar el nombre del fichero y la carpeta donde se guardará.

1.9.5 Cargar un entorno de trabajo

Cuando se inice una nueva sesión de trabajo `R` cargará automáticamente en el entorno de trabajo cualquier fichero con extensión **.RData** que se encuentre en el directorio de trabajo. Para cambiar de directorio, y por tanto de entorno de trabajo, se utiliza el menú **Fichero** → **Cambiar directorio de trabajo** y se selecciona el nuevo directorio en el cuadro de diálogo que aparece.

1.10 Ayuda

Otra de las ventajas de `R` es que tiene un sistema de ayuda muy documentado. Es posible conseguir ayuda sobre cualquier función, prodecimiento o paquete simplemente tecleando el comando `help()`. Por ejemplo, para obtener ayuda sobre el comando `mean` se teclearía

```
> help("mean")
```

y con esto aparecerá una ventana de ayuda donde se describe la función y también aparecen ejemplos que ilustran su uso. Si no se conoce exactamente el nombre de la función o comando, se puede hacer una búsqueda aproximada con el comando `help.search()`. Por emplo, si no se recuerda el nombre de la función logarítmica, se podría teclear

```
> help("logarithm")
```

y con esto aparecerá una ventana con todos los ficheros de ayuda que contienen la palabra logarithm.

Finalmente, también es posible invocar la ayuda general de `R` con el menú **Ayuda** → **Iniciar ayuda de R** con lo que aparecerá una página web desde donde podremos navegar a la información deseada.

Para más información sobre `R` se recomienda visitar la página `http://www.r-project.org/`, y para más información sobre `R-Commander` se recomienda visitar la página `http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/`. También puede encontrarse un buen manual de introducción a `R-Commander` en el menú de ayuda de la propia interfaz gráfica de usuario.

1.11 Ejercicios propuestos

Ejercicio 1.1 Se realiza una encuesta en la que se pregunta ha preguntado a un grupo de 50 personas su altura, nivel de estudios, número de personas que conviven en su domicilio habitual y su peso.

He aquí los 15 primeros resultados:

Altura	Nivel_estudios	Num_personas_domicilio	Peso
1.63	bachiller	2	54.88
1.44	fp	3	33.5
1.9	bachiller	3	89.38
1.58	bachiller	3	53.28
1.38	bachiller	5	36.62
1.8	bachiller	6	79.47
1.64	superior	3	72.47
1.93	bachiller	5	85.26
1.95	bachiller	2	102
1.59	bachiller	2	62.28
1.78	primario	4	81.17
1.96	primario	4	100.61
1.89	bachiller	5	94.43
1.52	bachiller	4	57.96
1.74	diplomado	4	67.86

Se pide:

(a) Utilizar el *editor de datos* de `RCommander` para introducir los datos.
(b) Una vez introducidos los datos, guardarlos en un fichero llamado **Encuesta**.
(b) **Exportar** los datos a un fichero llamado **Encuesta.txt**.
(b) **Añadir** datos correspondientes a 3 encuestados más y salvar/guardar estos datos en los ficheros descritos anteriormente (reemplazándolos).
(c) **Añadir** una nueva variable, llamada **Sexo** e introducir los datos correspondientes al *sexo* del encuestado (*Hombre* o *Mujer*), salvando posteriormente los datos.
(d) **Eliminar** los datos correspondientes a los encuestados números **9** y **14**, y **Visualizar** los datos para comprobar que se han eliminado.
(e) **Cargar** los datos del fichero **Encuesta.RData** (o **Importar** los del fichero **Encuesta.txt** para volver a tener en memoria todos los datos.
(f) **Filtrar** el conjunto de datos activos para quedarnos con aquellos encuestados cuyo sexo sea "Hombre" y llamar **EncuestadosHombres** al nuevo conjunto de datos.
(g) **Cambiar el conjunto de datos activos** a **Encuesta** y **Filtrar** el conjunto de datos para quedarnos con aquellos que tengan estudios de **fp** o **primario** y llamar **fp_o_primario** a dicho conjunto de datos. **Visualizar** los datos.

■

Ejercicio 1.2 Crear un conjunto de datos, de la manera que estime más conveniente (editor de `R`, hoja de cálculo, texto plano, etc), con los datos de la siguiente muestra. Si se ha creado con el editor de textos de `R`, guardarlo con el nombre **colesterol.rda** (o **colesterol.RData**), en caso contrario, importar los datos y posteriormente guardarlo en formato propio de `R` (**colesterol.rda**, o **colesterol.RData**).

Nombre	Sexo	Edad	Peso	Altura	Colesterol
José Luis Martínez Izquierdo	H	18	85	179	182
Rosa Díaz Díaz	M	32	65	173	232
Javier García Sánchez	H	24	71	181	191
Carmen López Pinzón	M	35	65	170	200
Marisa López Collado	M	46	51	158	148
Antonio Ruiz Cruz	H	68	66	174	249
Cristóbal Campos Ruiz	H	44	70	178	220
Beatriz Márquez Pérez	M	51	66	161	232
Isabel López Pinto	M	38	65	167	178
José Antonio Carmona Pérez	H	33	89	167	237

Se pide:

(a) **Exportar** el conjunto de datos y guardarlo en un fichero de texto llamado **colesterol.txt**.

(b) **Crear** una nueva variable donde se calcule el `índice de masa corporal` (**imc**) de cada paciente mediante la formula:

$$\text{imc} = \frac{\text{Peso (en Kg)}}{\text{Altura (en mt)}^2}$$

(c) **Recodificar** la variable **imc** de acuerdo a las siguientes categorías:

Menor de 18.5	Bajo peso
De 18.5 a 24.9	Saludable
De 25 a 29.9	Sobrepeso
Mayor o igual que 30	Obeso

y **visualizar** el nuevo conjunto de datos obtenido [a].

(d) **Crear** una nueva variable llamada **altura_pulgadas** (1 pulgada = 2.54 *cms*).

(e) **Filtrar** el conjunto de datos para quedarnos con aquellas personas cuya altura sea superior o igual a 170 *cms*, llamar **Altos** al nuevo conjunto de datos y exportarlos a un fichero llamado **Altos.txt**.

(f) **Cambiar el conjunto de datos activos** al conjunto de partida y **Eliminar** los datos correspondientes a `Antonio Ruiz Cruz`.

(g) **Filtrar** el conjunto de datos para quedarnos con aquellas personas cuya *sexo* sea **M** llamando **Mujeres** al nuevo conjunto de datos y **guardarlos** en un fichero **Mujeres**.

(h) **Filtrar** el conjunto de datos creado en el apartado anterior con la condición de que el **colesterol** sea superior a 200.

(i) Partiendo del conjunto de datos **colesterol**, **Filtrar** dicho conjunto con las condiciones de los dos apartados anteriores en un solo paso (es decir, que la variable *sexo* sea **M** y la variable *colesterol* superior a 200).

(i) Volver a **cambiar el conjunto de datos activos** al conjunto de partida, **Eliminar** la variable **colesterol** y visualizar el nuevo conjunto de datos.

[a] **Nota:** Puede utilizarse en esta recodificación las variable **lo** (*lowest* –el valor más bajo) y **hi** (*highest* –el valor más alto).

Ejercicio 1.3 El fichero **Pulso.txt** contiene información sobre el pulso de un grupo de pacientes que han realizado distintos ejercicios: pulso en reposo (**Pulse1**), pulso después de hacer ejercicio

(**Pulse2**), tipo de ejercicio (**Ran**: 1= correr, 2= andar), sexo (**Sex**, 1= hombre, 2 = mujer), peso (**Weight**), altura (**Height**), si es o no fumador (**Smokes**, 1 = No, 2 = Si) y tipo de actividad (**Activity**, 1 = suave, 2 = Moderada, 3 = Alta). Se pide:

(a) **Importar** el fichero de datos.
(b) Editar el conjunto de datos y **renombrar** el nombre de las variables al castellano (**Pulso1**, **Pulso2**, **Correr**, **Fumar**, **Peso**, **Altura**, **Sexo** y **Actividad**.
(c) Cambiar el tipo de la variable **Correr** para que sea de tipo *Caracter*.
(d) **Recodificar** la variable correr en las categorías "*Si*" y "*No*".
(e) **Recodificar** la variable **Actividad** a "*Suave*", "*Media*" y "*Alta*".
(f) La variable **Peso** está expresada en libras (*pounds*). **Crear** una nueva variable (**Peso_kgs**) que indique el peso en kilogramos sabiendo que cada libra equivale a 454 gramos.
(f) **Calcular una nueva variable** llamada **Difer_pulsos** que indique la diferencia entre las variables **Pulso2** y **Pulso1**.
(g) La variable **Altura** está expresada en pulgadas. **Crear** una nueva variable (**Altura_cms**) que exprese la altura en centímetros sbiendo que 1 pulgada equivale a 2.54 ceentímetros.
(h) **Recodificar** la variable **Altura_cms** con de acuerdo a las siguientes categorías:

Menor de 160	Bajo
$160 \leq$ Altura_cms < 170	Altura media
$170 \leq$ Altura_cms < 180	Alto
Mayor o igual que 180	Muy Alto

llamando **AlturaC** a la nueva variable.
(i) **Filtrar** el conjunto de datos activo quedándonos sólo con aquellas personas cuya varable **AlturaC** sea *Alto*, llamando **Altos** a este nuevo conjunto de datos.
(j) **Cambiar el conjunto de datos activo** al conjunto de partida y **Segmentar** la variable **Peso_Kgs** en **tres segmentos equidistantes** a los que llamaremos "*Delgado*", "*Normal*" y "*Obeso*" (**Nota:** en este caso, cuando tengamos el menú desplegable marcaremos la opción **Especificar nombres** dándole los nombres anteriores a cada uno de los tres distintos niveles).

■

Ejercicio 1.4 El fichero de datos **trees** del paquete incluido en `R`, **datasets** contiene información sobre el *diámetro*, *altura* y *volumen de madera* de cerezos negros. Se pide:

(a) Leer los datos desde paquete (**Datos** $\rightarrow$ **Conjunto de datos en paquetes** $\rightarrow$) **Leer conjunto de datos desde paquete adjunto**.
(b) Editar el conjunto de datos y **renombrar** el nombre de las variables al castellano (**Girth = Diametro**, **Height = Altura**, **Volume = Volumen**.
(c) La variable **Diametro** esta en pulgadas. **Crear** una nueva variable, **Diametro_cms** que nos dé el diámetro en centímetros sabiendo que 1 pulgada equivale a 2.54 cms.
(d) **Segmentar** la variable **Diametro** en **tres segmentos equidistantes** a los que en la opción **Especificar nombre** llamaremos a "*Pequeño*", "*Normal*" y "*Grande*".
(e) **Segmentar** la variable **Altura** en **4 segmentos equidistantes**, marcando la opción **Rangos** en los *Nombres de niveles* llamando **Altura_agrupada** a esta variable. Posteriormente, **Recodificar** esta nueva variable y en las **directrices de recodificación** asignamos a cada segmento su **punto medio** (o **marca de clase**) como en la Sección 1.8.4 llamando **Altura_marca** a la nueva variable.
(f) **Recodificar** la variable **Volumen** de acuerdo a las siguientes categorías:

Menor de 20	Volumen Pequeño
$20 \leq$ Volumen < 45	Volumen medio
Mayor o igual que 45	Volumen Alto

llamando **VolumenC** a la nueva variable.

■

Ejercicio 1.5 Importar el conjunto de datos del fichero de texto **Nations.txt**. Se pide:

(a) **Visualizar** el conjunto de datos.

(b) **Filtrar** el conjunto de datos para quedarse únicamente con los países de Europa y salvar el nuevo conjunto de datos en un fichero llamado **Europa.RData**.

(c) Tabajar con dicho fichero de datos algunas de las cuestiones planteadas en los ejercicios anteriores.

■

2. Estadística Descriptiva Unidimensional

2.1 Introducción

La **Estadística** tiene por objeto el estudio de grandes cantidades de datos. Estos datos se obtienen como medidas de una o varias características de los individuos de una población. Un ejemplo sería la realización de un estudio de la población onubense utilizando sus valores de edad, sexo, peso, altura y color del pelo.

Dentro de la Estadística podemos distinguir dos grandes ramas:

- La **Estadística Inferencial**: trata de establecer conclusiones generales y hacer previsiones sobre una población a partir de los datos obtenidos de una muestra.
- La **Estadística Descriptiva**: tiene por objeto ordenar los datos, agruparlos, representarlos en tablas y gráficas, y resumirlos en una serie de parámetros que permitan captar las **características generales** de la población y así poder entenderla mejor o compararla con otras poblaciones.

2.2 Conceptos generales

Las características o variables bajo estudio se dividen en:

- *Cuantitativas*: si son medibles o cuantificables (peso, altura, producción de una fábrica, ...);
- *Cualitativas* o atributos: si no pueden ser medidas (profesión, estado civil, color de pelo, ...).

En las características cualitativas (atributos), las distintas observaciones reciben el nombre de **modalidades**. Por ejemplo el atributo *estado civil* da lugar, entre otras, a las modalidades: *soltero/a, casado/a y viudo/a*.

La función que a cada individuo le asocia el valor correspondiente de una determinada característica se denomina **variable estadística**. Tradicionalmente se utilizan las últimas letras del alfabeto (W, X, ...) para notar las variables estadísticas y sus valores concretos se denotan pon esas mismas letras en minúsculas y numeradas ($x_1, \ldots, x_m$) para distinguir la observación a la que pertenecen.

Atendiendo a la cantidad de valores que pueden tomar, se distinguen dos tipos de variables estadísticas:

- *Variables discretas*: son aquellas que sólo pueden tomar una cantidad finita o numerable de valores (altura en cm. de una persona, número de hijos de las familias españolas, etc.)
- *Variables continuas*: son aquellas que pueden tomar cualquier valor real dentro de un intervalo (distancia entre el 1^{er} y 2^{o} clasificado en una carrera de 100 metros).

$$\text{Clasificación} \begin{cases} \text{Cuantitativas} \begin{cases} \text{Continuas} \\ \text{Discretas} \end{cases} \\ \text{Cualitativas} \end{cases}$$

Ejercicio 2.1 — R.
El fichero de datos `acero2.rda` contiene los valores de las variables más relevantes de tres lineas de producción de una empresa siderúrgica. En total se disponen de 117 mediciones recogidas en las variables `consumo, pr.tbc, pr.cc, ..., naverias, sistema`. Determinar la clasificación de las variables presentes en el fichero de datos. ■

Al conjunto de individuos objeto del estudio de le llama **población**, al número de individuos de la población se le llama **tamaño poblacional**. Cuando no es posible estudiar todos los individuos de la población se estudia un subconjunto suyo que recibe el nombre de **muestra** y que, generalmente, se elige de manera aleatoria. La diferencia entre los valores máximo y mínimo de la muestra se denomina **recorrido**.

Durante este curso trabajaremos básicamente con variables cuantitativas discretas o continuas. Usualmente se denotará por X la variable de interés en el estudio, N el número total de observaciones y $x_1, x_2, \ldots, x_k$ los valores distintos observados de la variable ordenados en sentido creciente, es decir, $x_1 < x_2 < \cdots < x_k$. En el caso que X fuese un atributo, las observaciones $x_1, x_2, \ldots, x_k$ no se podrían ordenar.

Ejercicio 2.2 — R.
Basándose en el fichero de datos `acero2.rda` determinar la población bajo estudio y el número de observaciones. ■

2.3 Tabla de frecuencias.

El análisis de una muestra conlleva las siguientes tareas: recogida de datos, ordenación de los mismos, recuento, agrupación de datos, construcción de tablas estadísticas y representación gráfica adecuada.

De cara al análisis de la muestra, definimos:

- **Frecuencia absoluta** (n_i)**:** es el número de veces que aparece el valor x_i en la muestra.
- **Frecuencia absoluta acumulada** (N_i)**:** una vez ordenados los datos, generalmente en orden creciente, se llama así al número de veces que aparece un dato x_i y todos los anteriores. Esto es, si $x_1 < x_2 < \cdots < x_i < \ldots x_m$, entonces:

$$N_i = \sum_{k=1}^{i} n_k$$

- **Frecuencia relativa** (f_i)**:** es la razón entre la frecuencia absoluta del valor x_i y el tamaño de la muestra (N).

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

- **Frecuencia relativa acumulada (F_i):** Es la suma de la frecuencia relativa del dato x_i y la de todos los anteriores.

$$F_i = \sum_{k=1}^{i} f_k = \frac{N_i}{N}$$

Si la variable estadística es continua, o aún siendo discreta si el número de observaciones es muy grande, es aconsejable agruparla en intervalos. Cada uno de los intervalos $I_i = [L_i, L_{i+1})$ se denomina **intervalo de clase**, y se representa mediante un número m_i que se denomina **marca de clase**. Habitualmente se toma como marca de clase el punto medio del intervalo.

Agrupar los datos por intervalos facilita las operaciones posteriores, pero debemos tener presente que, en general, conlleva pérdida de información. El número apropiado de intervalos que debemos considerar dependerá de la naturaleza del problema y del grado de precisión deseado. Como norma, tomaremos un número de intervalos igual o ligeramente superior a $\sqrt{N}$.

■ **Ejemplo 2.1** Al lanzar un dado 10 veces se obtuvieron los siguientes resultados de forma consecutiva,

$$\texttt{puntuación} = [2, 6, 1, 1, 4, 3, 2, 1, 4, 5]$$

La variable `puntuación` es discreta. Resumimos la información de la muestra en la siguiente tabla:

x_i	n_i	N_i	f_i	F_i
1	3	3	0.3	0.3
2	2	5	0.2	0.5
3	1	6	0.1	0.6
4	2	8	0.2	0.8
5	1	9	0.1	0.9
6	1	10	0.1	1

La información recogida en la tabla nos permite afirmar que:
(1) En el 30% de las tiradas se obtuvo el 1.
(2) En el 10% de las tiradas se obtuvo el 3.
(3) En el 60% de las tiradas se obtuvo el 3 o un número menor que 3.
(4) En el 20% de las tiradas se obtuvo un 5 o un 6.

■

■ **Ejemplo 2.2 — R.** Basándose en el fichero de datos `acero2.rda` determinar la tabla de frecuencias de la variable `temperatura`.
Solución:
Para ello haremos **Estadísticos** → **Resúmenes** → **Distribución de frecuencias** y seleccionaremos la variable `temperatura`, obteniendo la frecuencia absoluta y la frecuencia relativa (multiplicada por 100).
Resultado:

```
counts:
temperatura
 Alta  Baja Media
   46    38    33

percentages:
temperatura
 Alta  Baja Media
39.32 32.48 28.21
```

Para obtener obtener también las frecuencias acumuladas podríamos, utilizando la función `cumsum`, modificar el código generado por la orden anterior del siguiente modo

```
local({
  .Table <- with(acero2, table(temperatura))
  cat("Frecuencia Absoluta:\n")
  print(.Table)
  cat("Frecuencia Relativa:\n")
  print(round(.Table/sum(.Table), 2))
  cat("Frecuencia Absoluta Acumulada:\n")
  print(cumsum(.Table))
  cat("Frecuencia Relativa Acumulada:\n")
  print(round(cumsum(.Table/sum(.Table)), 2))
})
```

y obtendríamos

```
Frecuencia Absoluta:
temperatura
 Alta  Baja Media
   46    38    33
Frecuencia Relativa:
temperatura
 Alta  Baja Media
 0.39  0.32  0.28
Frecuencia Absoluta Acumulada:
 Alta  Baja Media
   46    84   117
Frecuencia Relativa Acumulada:
 Alta  Baja Media
 0.39  0.72  1.00
```

■

■ **Ejemplo 2.3** Se mide la altura en centímetros de 10 personas, obteniéndose los siguientes resultados

$$156,\ 158,\ 165,\ 167,\ 173,\ 177,\ 179,\ 181,\ 185,\ 194$$

En este caso, la variable "*altura*" es discreta pero toma muchos valores. Lo que haremos es agrupar los datos en intervalos (ver Ejemplo **??**):

Intervalos de clase	Marcas de clase (x_i)	n_i	N_i	f_i	F_i
$[155, 165)$	160	2	2	0.2	0.2
$[165, 175)$	170	3	5	0.3	0.5
$[175, 185)$	180	4	9	0.4	0.9
$[185, 195]$	190	1	10	0.1	1.0

■

Ejercicio 2.3 — R.
Repetir el ejemplo anterior con la variable `hora`. ¿Por qué no se puede? ¿Y si ejecutamos el código del ejemplo anterior sustituyendo `temperatura` por `hora`? ¿Y si convertimos la variable

`hora` en un factor? ■

2.4 Representaciones gráficas

Aunque las tablas estadísticas encierran toda la información obtenida, su traducción a un gráfico proporciona una síntesis visual de dicha información y permite una interpretación mucho más intuitiva. No obstante, es necesario recalcar que la representación gráfica no es más que un elemento de análisis y representación, y que por sí sola no es suficiente para un estudio riguroso de la información contenida en los datos.

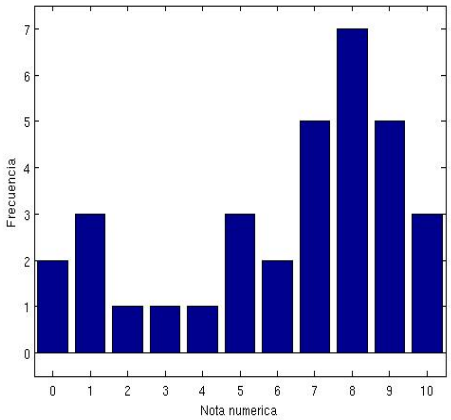
Una misma información puede ser representada gráficamente de muchas maneras. Existen distintos tipos de representación según el tipo de la variable, ya sea cuantitativa o cualitativa.

Supongamos que las notas de 33 alumnos de una clase vienen dadas por la siguiente tabla:

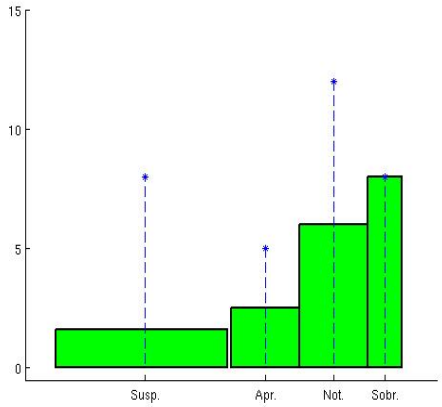
Nota	0	1	2	3	4	5	6	7	8	9	10
Nº Alumnos	2	3	1	1	1	3	2	5	7	5	3

estos datos pueden ser representados mediante un Diagrama de barras, un Histograma, un polígono de freciencias o un diagrama de sectores según lo consideremos más adecuado.

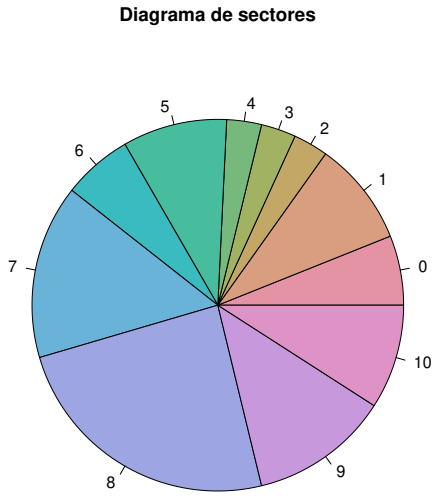
- **Diagramas de barras**: es adecuado para atributos y características cuantitativas de tipo discreto. Sobre el eje de abscisas se colocan las modalidades o los valores observados, siendo **la altura de las barras** proporcionales a las frecuencias absolutas o relativas correspondientes (ver Histograma).



- **Histogramas**: se utiliza para variables estadísticas de tipo continuo, o bien para discretas con un gran número de datos agrupados en intervalos. Se representan, sobre el eje OX los extremos de los intervalos de clase y sobre ellos se construyen rectángulos con una altura tal que **el área de cada rectángulo** sea proporcional a la frecuencia absoluta (o relativa) correspondiente.
  Si nos atenemos a intervalos de longitud 1 que contengan las calificaciones, obtendríamos la figura anterior ya que la base de las barras son todas iguales y de longitud 1, por lo que la altura es la que define el área
  Ahora bien, si dividiemos las notas en `Aprobado`, `Suspenso`, `Notable` y `Sobresaliente` el histograma es:



- **Diagramas de sectores**: son apropiados para atributos y para variables cuantitativas de tipo discreto. Se divide el círculo en sectores circulares de ángulo central proporcional a las frecuencias absolutas o relativas de los distintos atributos. De esta forma, el área de los sectores circulares es proporcional a dichas frecuencias.



Ejercicio 2.4 — R.
Calcular el histograma de la variable `consumo` con 12 barras equiespaciadas y el diagrama de sectores de la variable temperatura del fichero `acero2.rda`.

Para el Histograma utilizar la secuencia de comandos **Gráficas** → **Histograma** y en opciones seleccionar `frecuency` y 12 `breaks`; y para el diagrama de sectores **Gráficas** → **Gráfica de sectores** ■

2.5 Características asociadas a una distribución de frecuencias

Son parámetros que sintetizan el comportamiento de una distribución, como pueden ser las medidas de **centralización**, de **localización**, de **dispersión** y de **forma**.

2.5.1 Medidas de centralización

Una medida de centralización es un valor que es típico o representativo de un conjunto de datos y que tiende a situarse en el "*centro*" de éstos cuando están ordenados según su magnitud. Las más comunes on la *media*, la *moda* y la *mediana*.

En lo que sigue supondremos que tenemos un conjunto de datos $\{x_1,\dots,x_m\}$ con frecuencias absolutas $\{n_1,\dots,n_m\}$ tales que $n_1+\cdots+n_m=N$. (tenemos en total N datos que dan lugar a m valores diferentes)

Media aritmética: Se representa por $\overline{X}$ y se define como $\overline{X}=\dfrac{\sum_{i=1}^{m} x_i\cdot n_i}{N}$

Si la variable es continua o bien es discreta con valores agrupados en intervalos, se toma como x_i la marca de clase del intrevalo i–ésimo, y como n_i la frecuencia absoluta de dicho intervalo. La media aritmética representa el centro de gravedad de la distribución, aunque si los datos están muy dispersos o hay datos dispares, su valor puede ser poco representativo.

Media geométrica: Se define como: $G=\sqrt[N]{x_1^{n_1}\times\cdots\times x_n^{n_m}}$

Usualmente resulta más cómodo trabajar con el logaritmo neperiano de la media geométrica, pues de este modo

$$\ln G=\frac{1}{N}\sum_{i=1}^{m} n_i\cdot\ln x_i$$

con lo cual $G=e^{\ln G}$.

Media armónica Se define como: $H=\dfrac{1}{\frac{1}{N}\sum_{i=1}^{m}\frac{1}{x_i}\cdot n_i}$

Por tanto $\dfrac{1}{H}$ es la media aritmética de $\dfrac{1}{x_1},\dots,\dfrac{1}{x_m}$.

La relación que existe entre las tres medias es $H\le G\le\overline{X}$

Mediana se representa por $\widetilde{X}$ es el valor que divide el conjunto de los x_i en dos efectivos de igual tamaño.

A la hora de calcular la mediana de un conjunto de valores es necesario distinguir dos casos:

- **La variable es discreta:**

  Si para un cierto dato x_i ocurre que $N_i=\dfrac{N}{2}$ (para ello debe ser N un número par), la mediana es

  $$\widetilde{X}=\frac{x_i+x_{i+1}}{2}$$

  En el caso en que $N_{i-1}<\dfrac{N}{2}<N_i$, se tomará como mediana $M_e=x_i$.

  ■ **Ejemplo 2.4** Para la siguiente distribución de frecuencias absolutas acumuladas (N_i):

x_i	1	2	3	4	5	6
N_i	1	2	5	6	10	12

  la mediana es $\widetilde{X}=\dfrac{4+5}{2}=4.5$ ■

  ■ **Ejemplo 2.5** Para la distribución de frecuencias absolutas acumuladas (N_i):

x_i	1	2	3	4	5	6
N_i	1	3	7	12	15	18

la mediana es $\widetilde{X} = 4$, puesto que $\frac{N}{2} = 9$ y $N_3 < 9 < N_4$. ■

- **La variable es continua o discreta agrupada en intervalos:**
  primero hay que determinar el intervalo $[L_i, L_{i+1})$ que contiene la mediana, después se determina el valor de la mediana mediante interpolación. Esto es, si x_i es tal que $N_{i-1} < \frac{N}{2} \leq N_i$, entonces la mediana es

$$M_e = L_i + \frac{L_{i+1} - L_i}{n_i} \cdot \left(\frac{N}{2} - N_{i-1} \right)$$

Moda: Se representa por M_o, es el valor que aparece con más frecuencia en la distribución de los datos observados. Tiene sentido tanto en caracteres cuantitativos como en atributos. En general no es única, se denomina unimodal si presenta una única moda; bimodal si presenta dos modas; trimodal si presenta tres, etc.

En el caso de una variable discreta, es evidente cómo se calcula, pero cuando los datos están agrupados por intervalos, primero hay que determinar el intervalo $[L_i, L_{i+1})$ con mayor densidad de frecuencia (denominado intervalo modal) y a su altura h_i, el valor de la moda dentro de este intervalo es

$$M_o = L_i + \frac{(L_{i+1} - L_i)}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot (h_i - h_{i-1})$$

2.5.2 Medidas de localización

Son medidas que dan idea de la distribución de los datos.

- **Cuartiles:** Son los valores que dividen a la muestra en cuatro efectivos de igual tamaño. Hay tres cuartiles que denotaremos por Q_1, Q_2 y Q_3.
- **Deciles:** Son los valores que dividen al conjunto de datos observados en diez partes de igual tamaño. Hay nueve deciles que denotaremos por $D_1, D_2, \ldots, D_9$.
- **Percentiles:** Son los valores que dividen a la muestra en cien partes con igual número de observaciones. Hay 99 percentiles que denotaremos por $P_1, P_2, \ldots, P_{99}$.

Nota: Al igual que con la mediana, para calcular cada cuartil, decil o percentil tenemos que distinguir dos casos según el tipo de variable con el que estemos trabajando y actuar siguiendo el modelo de la mediana que coincide con el segundo cuartil, quinto decil y quincuagésimo percentil.

2.5.3 Medidas de dispersión

Son medidas que dan idea de la mayor o menor dispersión de los datos.

- **Recorrido o rango**: se representa por R, es la diferencia entre el valor más grande y el más pequeño de los datos.
- **Recorrido intercuartílico**: se representa IQR (sus siglas en inglés), es la diferencia entre el tercer y primer cuartil, $IQR = Q_3 - Q_1$[1]
- **Varianza**[2]

$$\mathrm{Var(X)} = \frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N-1}$$

[1] También existen los recorridos interdecílicos e interpercentílicos.

[2] Existen textos que llaman **varianza** a $\frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N}$ y **cuasivarianza** a $\frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N-1}$.

- **Desviación típica o estándar:**

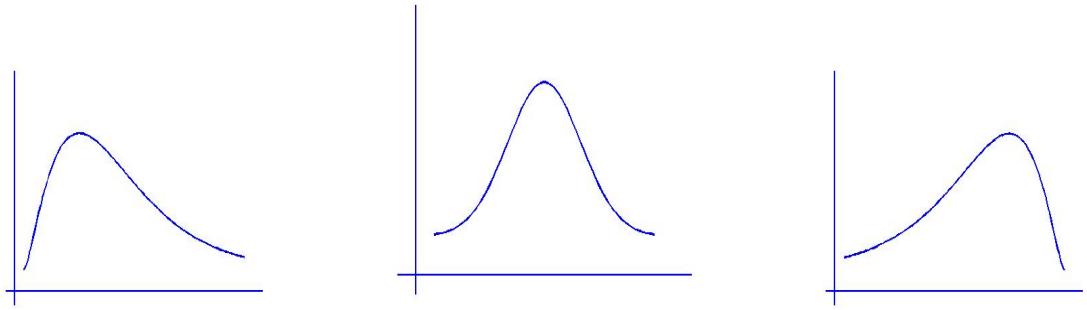
$$S = +\sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N-1}}$$

La desviación típica tiene la misma unidad que la variable bajo estudio. Si tenemos un conjunto de datos que corresponden a la altura en cm. de un grupo de personas, la desviación típica de esos datos también se mide en cm. Sin embargo, la unidad de la varianza es el cuadrado de la unidad en que se mide la magnitud bajo estudio. En el ejemplo anterior, la varianza de las alturas se mide en cm^2.

- **Coeficiente de variación de Pearson:** se define como $CV = \frac{S}{|\overline{X}|} \cdot 100\%$ (no tiene sentido definirlo cuando $\overline{X} = 0$).
  Es un número libre de influencia por cambio de escala en las medidas. Como puede observarse en la definición, el coeficiente de variación de Pearson se mide en tantos por ciento, permitiendo comparar la dispersión relativa de dos conjuntos de datos aunque las unidades de medida de esos datos sean diferentes.

2.5.4 Medidas de forma

Una vez iniciado el análisis estadístico de síntesis de información, para lo cual hemos estudiado las medidas de posición y dispersión de la distribución de una variable, necesitamos conocer más sobre su comportamiento. Las medidas de forma, como su propio nombre indica, nos proporciona información acerca de la forma de una distribución de valores sin necesidad de llevar a cabo su representación gráfica. Se clasifican en medidas de *asimetría* y medidas de *curtosis* o apuntamiento.



(a) Sesgada a la derecha (b) Insesgada (c) Sesgada a la izquierda

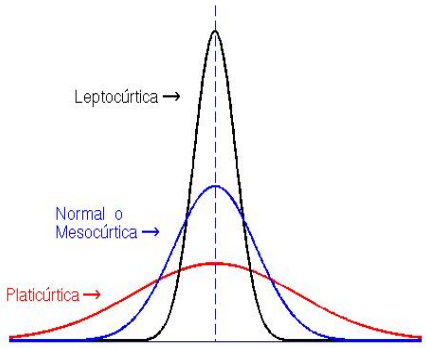
- **coeficiente de asimetría de Fisher**: Tienen como finalidad proporcionar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución.

$$g_1 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{x_i - \overline{X}}{S_1} \right)^3$$

donde $S_1 = S \cdot \frac{\sqrt{N-1}}{\sqrt{N}}$

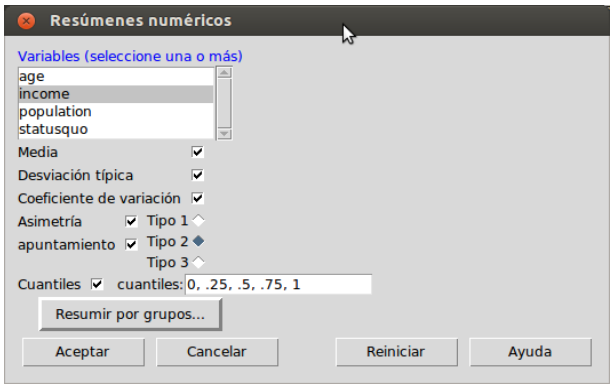
Si $g_1 = 0$ la distribución es simétrica, si $g_1 > 0$ la distribución es asimétrica positiva (sesgada a la derecha) y si $g_1 < 0$ la distribución es asimétrica negativa (sesgada a la izquierda).

- **coeficiente de Kurtosis**: se define como $g_2 = \frac{1}{N} \cdot \sum_1^N \left(\frac{x_i - \overline{X}}{S_1} \right)^4 - 3$ donde $S_1 = S \cdot \frac{\sqrt{N-1}}{\sqrt{N}}$ tiene como finalidad proporcionar un indicador de la conentración (o no) de los valores en el centro de la distribución.
  La medida de **apuntamiento** o **concentración** de una distribución de valores curtosis de una distribución la daremos en comparación con la distribución de una variable aleatoria **Normal**, que se corresponde con fenómenos muy habituales en la naturaleza, y cuya representación gráfica es la *campana de Gauss*. Esto es, con el coef. de Curtosis estudiamos la deformación, en sentido vertical, respecto a la Normal, de una distribución.



  Tomando la distribución Normal como referencia, diremos que una distribución puede ser más apuntada que la normal, $g_2 > 0$ (*leptocúrtica*), menos apuntada, $g_2 < 0$ (*platicúrtica*) o igual de apuntada, $g_2 = 0$ (*mesocúrtica*).

■ **Ejemplo 2.6 — R.** Con los datos del fichero **Chile**, calcular la media, desviación típica, cuartiles, coeficiente de variación, y medidas de forma para la variable **income** (*ingresos*)
Solución:
Para obtener los resultados solicitados haremos **Estadísticos** → **Resúmenes** → **Resúmenes numéricos**. A continuación, seleccionamos nuestra variable, marcamos la opción *cuantiles*, y seleccionamos los percentiles que deseemos estudiar, por defecto, aparecen los cuartiles (percentiles 25, 50 y 75), el mínimo y el máximo (0 y 1), tal como se muestra a continuación.



Resultado:

```
> numSummary(Chile[,"income"], statistics=c("mean", "sd", "quantiles", "cv",
+   "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")
```

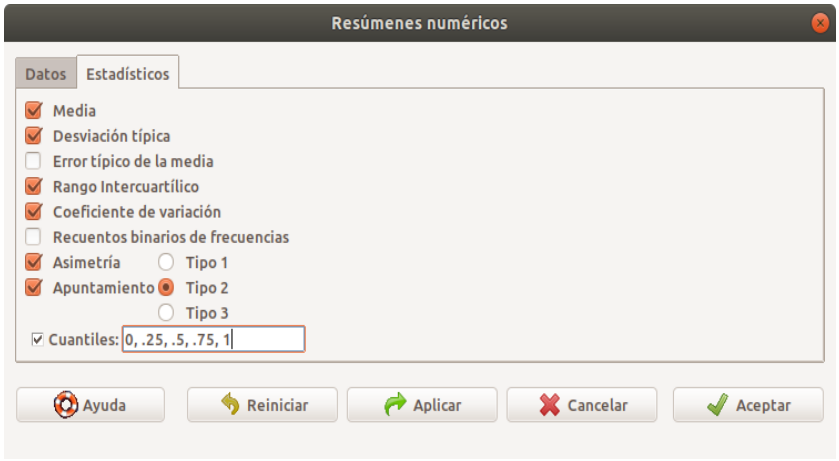


Figure 2.1: Opciones del Ejemplo 2.7

```
     mean       sd       cv skewness kurtosis   0%  25%   50%   75%  100%    n NA
 33875.86 39502.87 1.166107 2.587532 7.316215 2500 7500 15000 35000 2e+05 2602 98
```

La media es 33875.86 pesos, con una desviación típica de 39502.87 pesos, un coeficiente de asimetría de 2.587532 (asimétrica a la derecha), un coeficiente de curtosis de 7.316215 (platicúrtica), un mínimo de 2500 pesos, un máximo de 200000 pesos, y cuartiles $Q_1 = P_{25} = 7500$ pesos, $Q_2 = P_{50} = Me = 15000$ y $Q_3 = P_{75} = 35000$ pesos.

■

■ **Ejemplo 2.7 — R.** Basándose en el fichero de datos `acero2.rda` determinar:

a) la media, la mediana, los cuartiles, el rango, el recorrido intercuartílico, la varianza, la desviación típica, el coeficiente de variación de Pearson, el coeficiente de asimetría de Fisher y el coeficiente de Kurtosis de la variable `consumo`.

b) el recorrido interdecílico, el recorrido interpercentílico, el octavo decil y el percentil 73 de la variable `consumo`.

a) Solución:

Para ello haremos **Estadísticos** → **Resúmenes** → **Resúmenes numéricos** y seleccionaremos las opciones de la figura 2.1.

Resultado:

```
> numSummary(acero2[,"consumo", drop=FALSE], statistics=c("mean", "sd", "IQR", ...
                      "quantiles", "cv", "skewness", "kurtosis"), quantiles =  ...
                      c(0,.25,.5,.75,1), type="2")

   mean     sd   IQR    cv  skewness  kurtosis   0%   25%    50%    75%   100%   n
 139.456 55.185 83.39 0.395 0.003243 -0.345612 17.5 99.09 140.07 182.48 290.72 117
```

Si nos fijamos podemos comprobar que no aparecen ni la varianza ni el rango, aunque son muy fáciles de obtener sabiendo que la varianza es el cuadrado de la desviación típica y que el rango es el percentil 100% menos el percentil 0% por lo que bastaría con hacer

```
> varianza = 55.18525**2
> varianza
[1] 3045.412

> rango = 290.72 - 17.5
```

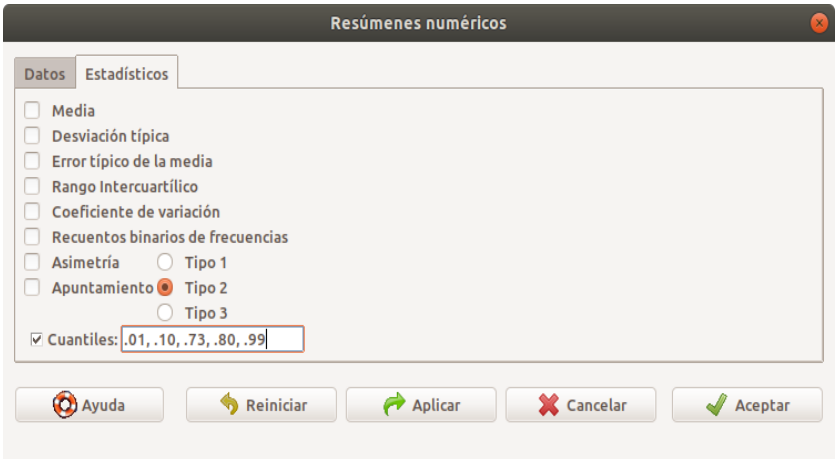


Figure 2.2: Opciones del Ejemplo 2.7

```
> rango
[1] 273.22
```

b) Solución:
Para ello haremos como en el apartado anterior, **Estadísticos** → **Resúmenes** → **Resúmenes numéricos**, desmarcamos todas las opciones y en la opción **Cuantiles** ponemos los valores "0.01, 0.10, 0.73, 0.80, 0.90, 0.99" (ver Figura 2.2), obteniendo
Resultado:

```
numSummary(acero2[,"consumo", drop=FALSE], statistics=c("quantiles"), ...
                      quantiles=c(0.01,0.10,0.73,0.80,0.90,0.99))
      1%      10%      73%      80%      90%      99%
 21.3276  66.3060 177.9976 189.9840 201.2020 250.0596
```

Con estos datos ya podemos determinar el octavo decil (189.9840) y el percentil número 73 (177.9976). Y para calcular los recorridos solicitados hacemos:

```
> RecIntDecil = 201.2020 - 66.3060
> RecIntDecil
[1] 134.896

> RecIntPerc  = 250.0596 - 21.3276
> RecIntPerc
[1] 228.732
```

Obteniendo 134.896 como recorrido interdecílico y 228.732 como recorrido interpercentílico.

■

2.6 Diagramas de caja

Un **diagrama de cajas (Box Plot o Whiskers Diagram)** es una representación gráfica de las principales características de un conjunto de datos. Esta representación nos permite, de un solo vistazo, obtener una imagen clara de su distribución. En ella podemos visualizar, entre otras características, cinco de los descriptores antes mencionados.

1. El mínimo del conjunto de datos.

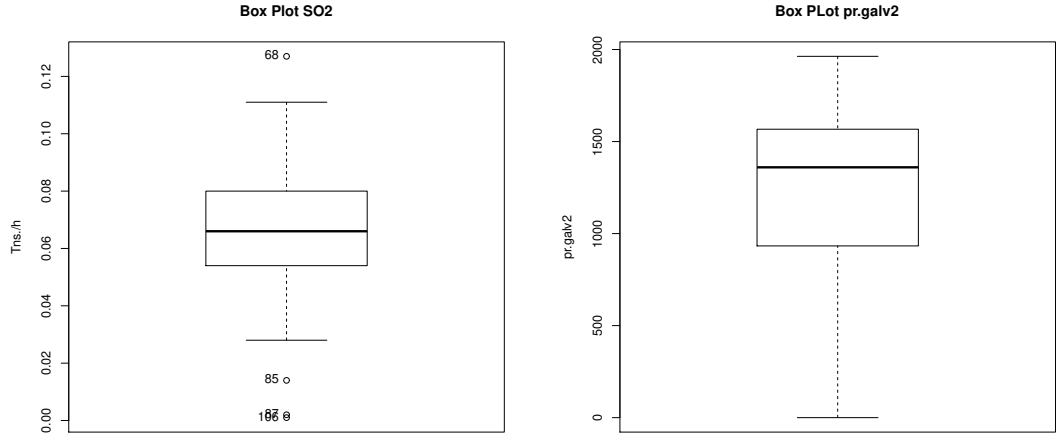


Figure 2.3: Ejemplos de diagramas de cajas.

2. El primer cuartil (Q_1).
3. La mediana (Q_2).
4. El tercer cuartil (Q_3).
5. El máximo.

El elemento central del diagrama de cajas, ver figura 2.3, es "la caja", rectángulo que va desde el primer cuartil hasta el tercer cuartil y dividido por una línea a la altura de la mediana. Es importante observar que en esa región se concentra la mitad de los datos del conjunto de datos; un cuarto por encima de la mediana y otro por debajo.

A cada extremo de la caja se encuentran los "bigotes" y se extienden, el bigote inferior, desde la caja hasta

$$B_{inf} = \max\{Q_1 - 1.5\,IQR, \min\{x_i / x_i \geq Q_1 - 1.5\,IQR\}\}$$

y el superior desde la caja hasta

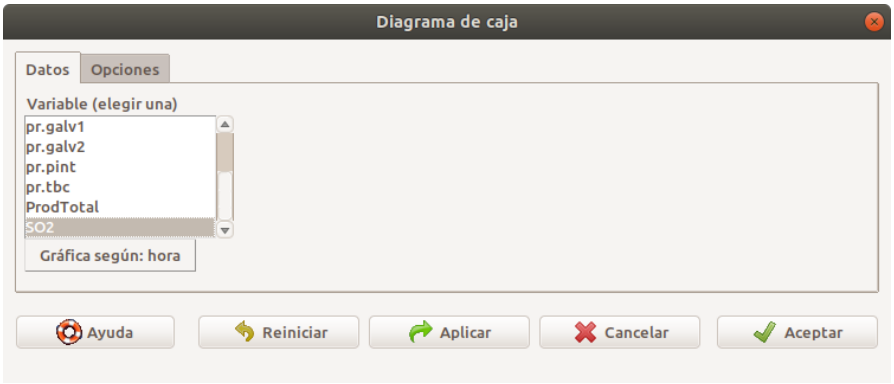
$$B_{sup} = \min\{Q_3 + 1.5\,IQR, \max\{x_i / x_i \leq Q_3 + 1.5\,IQR\}\},$$

donde $IQR = Q_3 - Q_1$ es el rango intercuartílico.

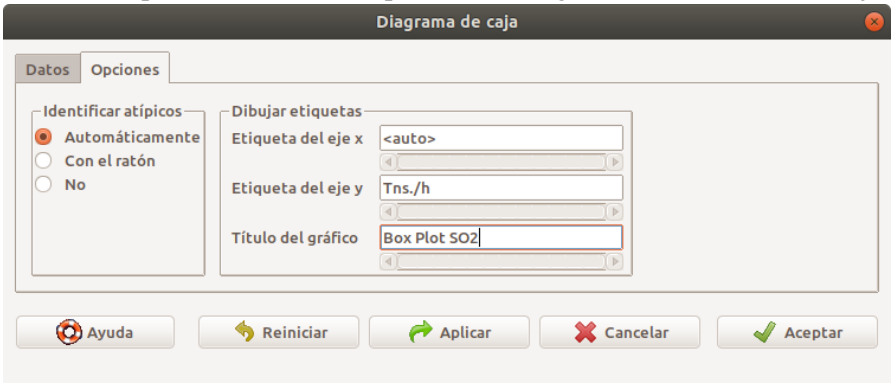
Los valores que están entre 1,5 y $3\,IQR$ de los extremos de la caja se denominan valores atípicos y los que están por encima de esa distancia son catalogados cómo outliers o valores extremos.

■ **Ejemplo 2.8 — R.** Basándose en el fichero de datos `acero2.rda` obtener el diagrama de caja de la variable `SO2`.

Solución:
Para ello haremos **Gráficas** → **Diagrama de cajas** y seleccionamos la variable `SO2`.



En la pestaña de **opciones** podemos configurar los nombres de los ejes y de la gráfica.



■

Ejercicio 2.5 — R.
De las variables `pr.tbc, NOX,pr.ca` y `naverías` del fichero de datos `acero2.rda` obtener los diagramas de caja e interpretarlos. Obtener el diagrama de caja de la variable `ProdTotal` agrupando por la variable `hora`. Interpretar la gráfica.
Ayuda: Para agrupar mediante una variable ésta ha de ser discreta o un factor.

■

2.7 Ejercicios Resueltos

■ **Ejemplo 2.9** Los datos siguientes, recogidos en el fichero `Creatinina.rda`, dan la cantidad de creatinina en miligramos por cien c.c. en muestras de orina correspondientes a 50 individuos:

1.61	1.51	1.65	1.58	1.54	1.65	1.40	1.08	1.81	1.38
1.56	1.83	1.69	1.22	1.22	1.68	1.47	1.68	1.49	1.80
1.33	1.83	1.50	1.96	1.67	1.60	1.23	1.54	1.73	1.43
2.18	1.46	2.34	1.90	2.02	1.31	1.52	1.47	1.43	2.29
1.75	1.90	1.66	1.26	1.75	1.66	1.66	2.30	2.00	1.20

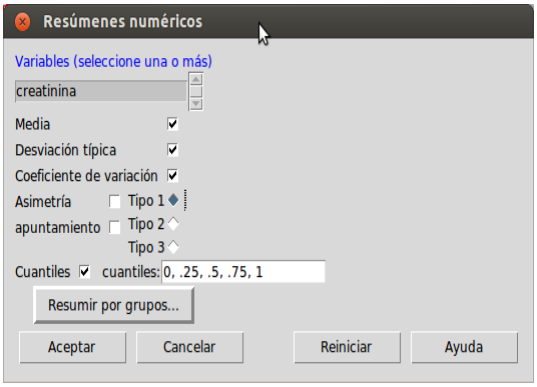
Calcular:
1. Media, desviación típica, mediana, cuartiles y coeficiente de variación de Pearson.
2. Dibujar el histograma asociado.
3. Agrupar los datos en 9 intervalos de igual longitud y dibujar su histograma.

■

Solución

En primer lugar hemos de introducir los dtos (**Datos** → **Nuevo conjunto de datos**, e introducirlos).

(a) Una vez introducidos los datos, se pulsa sobre **Estadísticos** → **Resúmenes** → **Resúmenes numéricos**, y podemos seleccionar la media y desviación típica de la variable introducida. Si nos fijamos en el cuadro de diálogo, nos dará además, los cuartiles, aunque se puede modificar para que nos de un percentil concreto.



El resultado que se obtiene es:

```
> numSummary(Creatinina[,"creatinina"], statistics=c("mean", "sd",
+   "quantiles", "cv"), quantiles=c(0,.25,.5,.75,1))
   mean        sd        cv   0%    25%  50%    75% 100%  n
 1.6346 0.2893992 0.1770459 1.08 1.4625 1.63 1.7875 2.34 50
```

(b) Pulsamos sobre **Gráficas** → **Histograma**, se selecciona la variable, en caso de no estarlo, y damos a aceptar.

(c) Para hacer una agrupación, hay que irse a **Datos**, → **Modificar datos** → **Segmentar una variable numérica**, darle un nombre a la nueva variable (la obtenida al realizar la agrupación), y marcamos el número de intervalos que nos interesa, en este caso, el número de clases, será 9. Si se marca "*Especificar nombres*", se le podrá asignar un nombre a cada nivel. Si marcamos "*Rangos*", nos asignará a cada valor, el intervalo numérico al que pertenece, y si le damos a "*Número*", nos dirá el número del intervalo al que pertenece (1 es el primer intervalo, "2" el segundo, ...).
Si una vez realizado, queremos obtener sus frecuencias, nos iríamos a **Estadísticos** → **Resúmenes** → **Distribución de frecuencias**.

■ **Ejemplo 2.10** Entre los ficheros que podemos descargarnos de *R*, se encuentra el fichero *Cars*93 del paquete `MASS`, sobre este conjunto de datos, se pide:

1. Representar la variable cualitativa Manufacturer (fabricante) mediante un diagrama de barras y mediante un diagrama de sectores.
2. Calcularle a la variable Precio Mínimo (Min.Price) la media, desviación típica, mediana, el primer cuartil y el percentil 55.
3. Agrupar la variable Precio (Price) en 7 intervalos de igual longitud y representarla en un histograma. Calcular a la variable agrupada los mismos parámetros que a la variable sin agrupar y comparar los resultados obtenidos.
4. Dadas las variables Min. Price y Max.Price, calcular una nueva variable que sea el resultado de hacer la media entre ambas, y calcular la media y desviación típica de la nueva variable.

■

Solución

Si pulsamos sobre Datos, Conjunto de datos en paquetes,Lista de conjuntos de datos en paquetes, obtenemos una lista de todos los ficheros de ejemplo que se han descargado con el programa, el siguiente ejemplo, lo haremos con el fichero *Cars*93, para ello, hay que pulsar sobre Datos, Conjunto de datos en paquetes,Leer de conjuntos de datos desde paquete adjunto.

(a) Diagrama de barras: Gráficas, Gráfica de barras, se selecciona la variable Manufacturer y Aceptar
Diagrama de sectores: Gráficas, Gráfico de sectores, se selecciona la variable **Manufacturer** y aceptar.

(b) La media, la mediana, la desviación típica, el primer cuartil y el percentil 55: Estadísticos, Resúmenes, Resúmenes numéricos. Se selecciona la variable Min.Price. En cuantiles, poner .25 (primer cuartil), .5 (mediana) y .55 (percentil 55). De los resultados, Mean es la media, SD es la desviación típica, y la mediana es el percentil asociado al 50%., el primer cuartil es el asociado al 25%, y el percentil 55 es el asociado al 55%.

(c) Agrupar: Primero, hacer la agrupación: Datos, Modificar variables del conjunto de datos activo, Segmentar variable numérica. Seleccionar la variable Precio (Price). Marcar que el número de intervalos es 7. Ponerle nombre a la variable de salida (PrecioAgrupada).
Segundo, pedirle una tabla de frecuencias en la que se observen los intervalos: Estadísticos, Resúmenes, Distribución de frecuencias, y seleccionar la variable PrecioAgrupada. Una vez que tenemos los intervalos, calcularles su marca de clase (punto medio del intervalo).
Tercero, recodificar la variable con los valores de las marcas de clase: Datos, Modificar variables del conjunto de datos activo, Recodificar Variables. Seleccionar la variable PrecioAgrupada, y darle un nombre a la variable resultado (MarcadeClase). Desmarcar la opción Convertir cada nueva variable en factor. En el recuadro Introducir directrices de recodificación, hacer la recodificación como sigue:

$$"(7.35, 15.1]" = 11.225$$

$$"(15.1, 22, 9]" = 19$$

$$"(22.9, 30.7]" = 26.8$$

$$"(30.7, 38.6]" = 34.65$$

$$"(38.6, 46.4]" = 42.5$$

$$"(46.4, 54.2]" = 50.3$$

$$"(54.2, 62]" = 58.1$$

Cuarto, calcular los datos que nos piden: Repetir el apartado B para la variable MarcadeClase

(d) Para calcular una nueva variable en base a variables que ya se tienen en el fichero de datos: Datos, Modificar variables del conjunto de datos activo,Calcular una nueva variable. Dar un nombre a la variable resultado (MediaPrecios), y en expresión a calcular, introducir,seleccionando las variables del recuadro superior: $(Min.Price + Max.Price)/2$ y darle a aceptar.
Una vez que se tiene la variable, operar como en los apartados A y B para obtener la media y la desviación típica asociada.

2.8 Estadística Descriptiva Bidimensional

2.8.1 Conceptos generales

La estadística descriptiva bidimensional tiene como objeto el estudio de poblaciones de individuos resultantes de la observación conjunta de dos características. Así, si tenemos una muestra de tamaño N y observamos las características X e Y, cada observación determinará una pareja de valores (x,y).

Si las observaciones de X dan lugar a $x_1,x_2,\ldots,x_n$ valores distintos y los de Y a $y_1,y_2,\ldots,y_m$, cada observación vendrá descrita por una pareja de valores (x_i,y_j). Como cada uno de los valores de una de las variables suele aparecer con distintos valores de la otra. Los datos se ordenan en una **tabla de doble entrada** como la de la Tabla 2.1. De cara al análisis de la muestra, definimos:

- **Frecuencia absoluta conjunta (n_{ij}):** es el número de veces que aparece el par (x_i,y_j) en la muestra.
- **Frecuencia relativa conjunta (f_{ij}):** es el número de veces que aparece el par (x_i,y_j) en la muestra dividido por el número total de elementos en la muestra, $f_{ij}=\dfrac{n_{ij}}{N}$.
- **Diagrama de dispersión:** es la nube de puntos que se obtiene con representación sobre el plano XY de los puntos de coordenadas (x_i,y_j). Esta representación se utiliza principalmente para obtener una primera impresión de la posible dependencia entre ambas variables.

Nota 2.8.1 Obviamente en la Tabla 2.1 se cumple que

$$\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}=N \qquad \text{y} \qquad \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}=\frac{1}{N}\cdot\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}=1$$

■ **Ejemplo 2.11 — R.** Basándose en el fichero de datos `acero2.rda` Obtener el diagrama de dispersión de la “Emisión de óxido nitroso por hora (NO2)” “Emisión de dióxido de carbono por hora (CO2)”.

Solución:
Para dibujar el diagrama de dispersión, nos vamos a **Gráficas** $\rightarrow$ **Diagrama de dispersión**. Seleccionamos como variable para el eje X a $CO2$, y como variable para el eje Y a $NO2$.

Si queremos ver sólo la nube de puntos, quitamos las opciones marcadas por defecto de "cajas de dispersión marginales", "línea de mínimos cuadrados", "líneas suavizadas", y "mostrar extensión", dejando lo demás como está. El resultado es el de la Figura 2.4: ■

■ **Ejemplo 2.12** (De tabla de doble entrada). El departamento de personal de una empresa encargó un estudio para conocer la posible relación existente entre la edad y el absentismo laboral de los trabajadores. Los datos obtenidos se reflejan en la siguiente tabla:

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	n_{n1}	n_{n2}	$\cdots$	n_{nj}	$\cdots$	n_{nm}

Table 2.1: Tabla de frecuencias de las variables X e Y.

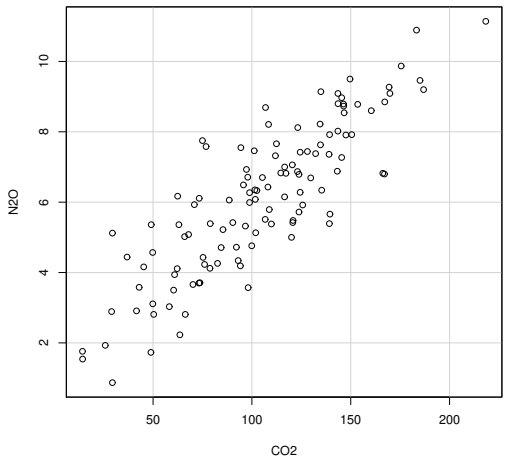


Figure 2.4: Diagrama de dispersión NO2 vs. CO2

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$\mathbf{n_{1*}}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	$\mathbf{n_{2*}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}	$\mathbf{n_{i*}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_n	n_{n1}	n_{n2}	$\cdots$	n_{nj}	$\cdots$	n_{nm}	$\mathbf{n_{n*}}$
	$\mathbf{n_{*1}}$	$\mathbf{n_{*2}}$	$\cdots$	$\mathbf{n_{*j}}$	$\cdots$	$\mathbf{n_{*m}}$	$\mathbf{n_{**}}$

Table 2.2: Tabla de frecuencias marginales de las variables X e Y.

Edad \ Absentismo	[0, 5)	[5, 10)	[10, 15)	[15, 20)
[15, 25)	–	3	7	1
[25, 35)	–	6	2	–
[35, 45)	1	3	2	–
[45, 55)	2	5	1	–
[55, 65)	3	2	–	1

■

2.8.2 Distribuciones marginales

En ocasiones puede ser interesante observar el comportamiento de una de las dos características independientemente del valor que tome la otra. En el ejemplo anterior podríamos estar interesados en estudiar el absentismo laboral sin tener en cuenta las edades.

- **Distribución marginal** de una variable es el conjunto de las sumas parciales para cada uno de los valores de esa variable, (ver Tabla 2.2).
- **La frecuencia marginal** de una variable es el conjunto de las sumas parciales para cada uno de los valores de esa variable (ver Tabla 2.2).

  Para x_i la frecuencia marginal es $n_{i*} = \sum_{j=1}^{m} n_{ij}$ y

para y_j la frecuencia marginal es $n_{*j} = \sum_{i=1}^{n} n_{ij}$.

- **La frecuencia relativa marginal** de una variable es el conjunto de las sumas parciales para cada uno de los valores de esa variable dividido por el número total de observaciones.
  Para x_i la frecuencia relativa marginal es $f_{i*} = \frac{n_{i*}}{N}$ y
  para y_j la frecuencia relativa marginal es $f_{*j} = \frac{n_{*j}}{N}$.

Nota 2.8.2 Es fácil observar que

$$n_{**} = \sum_{i=1}^{n} n_{i*} = \sum_{j=1}^{m} n_{*j} = \sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij} = N$$

y que

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

Naturalmente que para cada una de las distribuciones marginales cabe hacer un estudio de todos los aspectos que ha visto en las secciones anteriores para una variable. Nuestro objetivo ahora es hacer un estudio conjunto de ambas para analizar la posible "*dependencia*" entre ellas.

■ **Ejemplo 2.13** Siguiendo con el ejemplo 2.12, la tabla incluyendo las frecuencias marginales sería:

Edad \ Absentismo	[0, 5)	[5, 10)	[10, 15)	[15, 20)	
[15, 25)	–	3	7	1	11
[25, 35)	–	6	2	–	8
[35, 45)	1	3	2	–	6
[45, 55)	2	5	1	–	8
[55, 65)	3	2	–	1	6
	6	19	12	2	39

■

2.8.3 Distribuciones condicionadas

En el ejemplo 2.12 podríamos estar interesados en estudiar el absentismo laboral para para un determinado grupo de trabajadores, por ejemplo el grupo de edades comprendidas entre 35 y 45 años o la distribución por edades para el grupo de trabajadores que han faltado entre 10 y 15 dias. Generalizando las ideas apuntadas, supongamos que tenemos interés en estudiar la distribución de X para un cierto valor de Y, y_j.

Si nos restringimos a la submuestra de los n_{*j} individuos de la columna de y_j (ver Tabla 2.3) podemos definir

- La **frecuencia relativa de x_i condicionada por y_j, (f_{ij}^{j})** como la proporción de veces que aparece (x_i, y_j) de entre el total de veces que aparece y_j.

$$f_{ij}^{j} = \frac{n_{ij}}{n_{*j}}$$

En la Tabla 2.3 podemos observar la **distribución de X condicionada** por y_j, se denota por $X|_{y_j}$ o por $X|_{Y=y_j}$. Análogamente se puede definir Y condicionada por x_i.

$Y = y_j$		
X	frec. absoluta	frec. relativa
x_1	n_{1j}	f_1^j
x_2	n_{2j}	f_2^j
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_i^j
$\vdots$	$\vdots$	$\vdots$
x_n	n_{nj}	f_n^j
	$\sum_{j=1}^{m} n_{ij} = n_{*j}$	

Table 2.3: Distribución de X condicionada por y_j.

Nota 2.8.3 Se verifica que la **frecuencia relativa conjunta** es igual al producto de la **frecuencia relativa condicional** por la **frecuencia relativa marginal**.

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{\cdot j}} \cdot \frac{n_{\cdot j}}{N} = f_i^j \cdot f_{*j}.$$

Del mismo modo se verifica que $f_{ij} = f_j^i \cdot f_{i*}$

■ **Ejemplo 2.14** En el ejemplo 2.12, la tabla que muestra la distribución de las edades de los trabajadores que faltan entre 10 y 15 dias, $X|_{y\in[10,15)}$, es

$Y \in [10, 15)$		
X=edad	frec. absoluta	frec. relativa
[15, 25)	7	7/12
[25, 35)	2	2/12
[35, 45)	2	2/12
[45, 55)	1	1/12
[55, 65)	0	0
	12	

■

Independencia

Definición 2.8.1 El carácter X es independiente del carácter Y (la variable X es independiente de Y) si las distribuciones de X condicionadas por cada y_j son iguales, es decir, si $X|_{y_j}$ no depende de y_j.

La definición anterior equivale a exigir que las columnas de las frecuencias relativas de las distribuciones condicionadas por cada y_j sean iguales y por tanto que las columnas de las frecuencias absolutas sean proporcionales.

Proposición 2.8.4 Si X es independiente de Y, entonces las distribuciones condicionadas de X coinciden con la marginal de X, es decir, $f_i^j = f_{i*} \quad \forall j = 1, \ldots, m \quad \forall i = 1, \ldots, n.$

Proposición 2.8.5 Si X es independiente de Y entonces Y es independiente de X.

Proposición 2.8.6 X e Y son independientes si y sólo si $f_{ij} = \dfrac{f_{i*} \cdot f_{*j}}{N}$ $\left(\text{ó } n_{ij} = \dfrac{n_{i*} \cdot n_{*j}}{N}\right)$.

Ejercicio 2.6 Una muestra entre la población universitaria, sobre fumadores y no fumadores dió el resultado que se recoge el la siguiente tabla.

	Mujeres	Hombres	Totales
Fuma	200	150	350
No fuma	150	180	330
Totales	350	330	680

Estudiar si a la vista de los datos el hecho de fumar o no es independiente del sexo. ■

2.8.4 Covarianza

Definición 2.8.2 La **covarianza** de X e Y, S_{XY}, es la media del producto de las desviaciones de X respecto de $\overline{X}$ por las desviaciones de Y respecto de $\overline{Y}$.

$$S_{XY} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - \overline{X}) \cdot (y_j - \overline{Y}) \cdot n_{ij}}{N-1}$$

La covarianza sirve para medir el grado de dependencia entre las dos variables. Una covarianza positiva indica relación directa entre las variables (valores grandes de una variable están asociados con valores grandes de la otra). Una covarianza negativa indica relación inversa entre las variables (valores grandes de una variable están asociados con valores pequeños de la otra).

Propiedades de la covarianza

1. La covarianza es invariante por cambios de orígen. Es decir, si hacemos
   $X' = X - \theta_X$ y $Y' = Y - \theta_Y$, entonces $S_{X'Y'} = S_{XY}$
2. Si se hace un cambio de escala, esto es
   $X' = a \cdot X$ y $Y' = b \cdot Y$, entonces $S_{X'Y'} = a \cdot b \cdot S_{XY}$
3. De las dos anteriores se deduce que si hacemos $X' = \dfrac{X - \theta_X}{a}$ y $Y' = \dfrac{Y - \theta_Y}{b}$ entonces

$$S_{X'Y'} = \frac{1}{a} \cdot \frac{1}{b} \cdot S_{XY} \qquad \text{y} \qquad S_{XY} = a \cdot b \cdot S_{X'Y'}$$

4. Si X e Y son independientes, su covarianza es cero (el recíproco no es cierto, en general).

2.8.5 Coeficiente de correlación lineal

Definición 2.8.3 Coeficiente de correlación lineal r mide el grado de dependencia lineal entre las variables.

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

.

Propiedades del coeficiente de correlación lineal

- $-1 \leq r \leq 1$

- $r^2 = 1$ sí y sólo si las observaciones se encuentran perfectamente alineadas, esto es, sobre una recta.
- Si X e Y son independientes entonces $r = 0$. (El recíproco no es cierto).

Nota 2.8.7 La relación lineal entre X e Y será mayor cuanto más parecido sea r^2 a 1. Si $r = 0$, se dice que las variables son **incorreladas**, esto es, no existe relación lineal entre ambas.

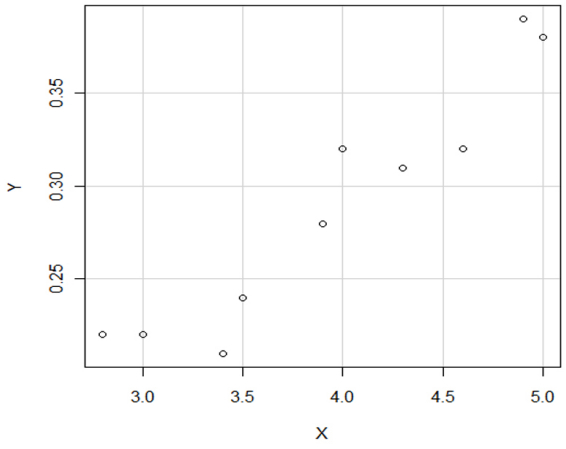
Un valor positivo de r nos indica que existe dependencia directa entre las variables (a medida que aumentan los valores de X también aumentan los de Y) mientras que cuando $r < 0$, la dependencia es inversa (al aumentar los valores de X disminuyen los de Y).

■ **Ejemplo 2.15 — Velocidad vs. Resistencia.** La velocidad de transmisión de ultrasonidos en una viga de hormigón es un indicador de su resistencia mecánica. Se miden las dos variables (velocidad y resistencia) en 10 vigas fabricadas con el mismo cemento, con los siguientes resultados:

Velocidad X Km/seg	2.8	3	3.4	3.5	3.9	4	4.3	4.6	4.9	5
Resistencia Y Tn/cm^2	0.22	0.22	0.21	0.24	0.28	0.32	0.31	0.32	0.39	0.38

a) Dibujar el diagrama de dispersión y comprobar gráficamente si las variables son independientes.
b) Calcular la correlación entre ellas.
Solución
a) Primero se introducen los datos definiendo dos variables en este caso, *Velocidad* e *Resistencia* (**Datos** → **Nuevo conjunto de datos ...**)

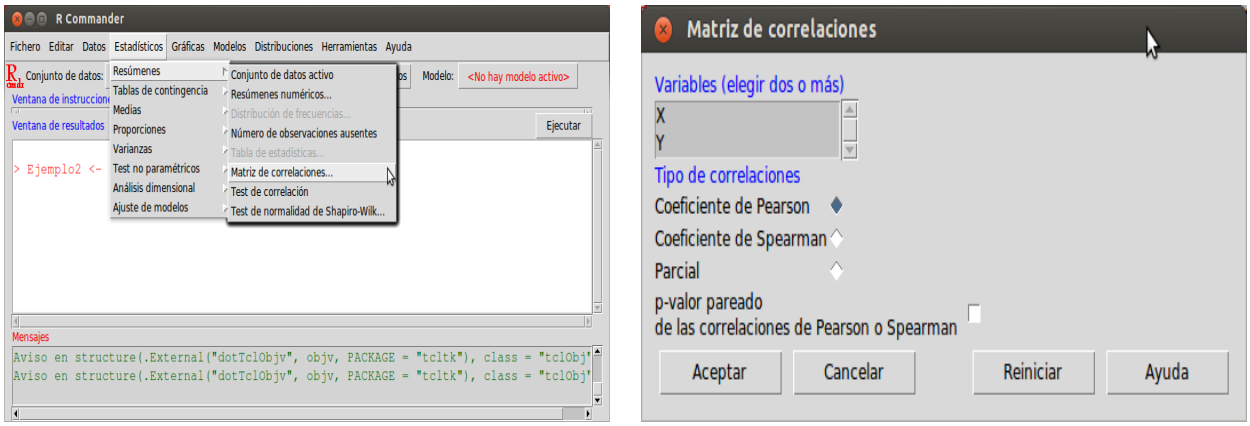
Para dibujar el diagrama de dispersión, nos vamos a **Gráficas** → **Diagrama de dispersión**. Seleccionamos como variable para el eje X la *Velocidad*, y como variable para el eje Y la *Resistencia*[3]. El resultado es:



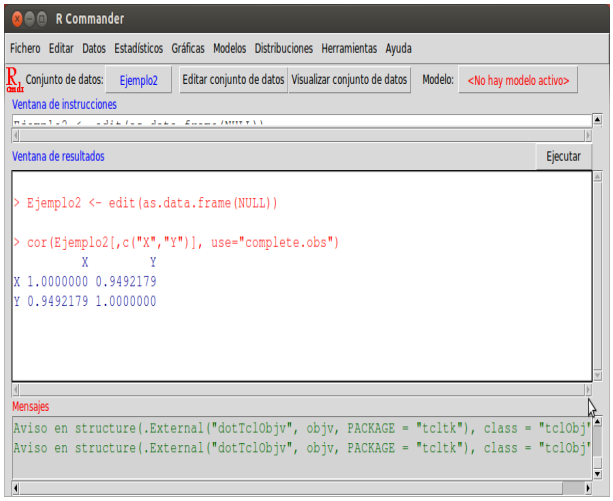
Gráficamente podemos ver que no son independientes, existe una relación entre ambas variables.

b) Para calcular el **coeficiente de correlación lineal** entre ambas variables, basta con seguir la secuencia **Estadísticos** → **Resúmenes** → **Matriz de correlaciones**; posteriormente seleccionamos las variables a las que queremos calcular el coeficiente de correlación, tal como se muestra en la Gráfica siguiente.

[3] Si queremos ver sólo la nube de puntos, quitamos las opciones marcadas por defecto de "cajas de dispersión marginales", "línea de mínimos cuadrados", "líneas suavizadas", y "mostrar extensión", dejando lo demás como está.

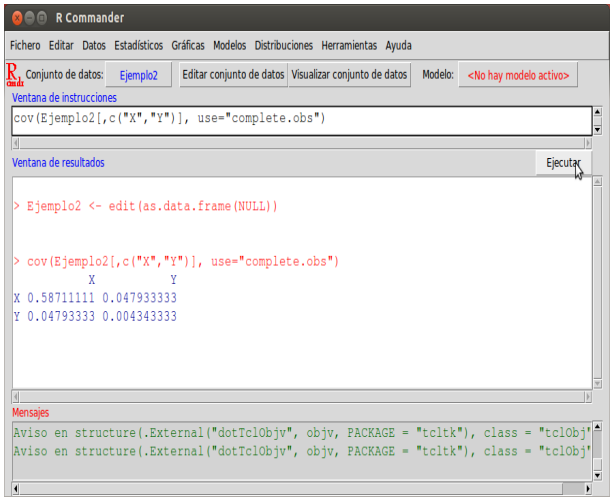


El resultado es una matriz de correlaciones simétrica, en cuya diagonal hay 1. El coeficiente de correlación en este caso es 0.949, que al ser próximo a 1, nos indica una fuerte relación lineal de tipo directo.



■

Nota 2.8.8 No hay orden en `RCommander` específica calcular la covarianza. Para calularla usaremos el *eco* del comando para calular la *matriz de correlaciones*, cambiamos la orden **cor** por **cov**, situamos el cursor al final de la línea y le damos a `Ejecutar`, tal y como aparece en la gráfica siguiente



2.8.6 Regresión

La recta que mejor aproxima los valores de Y respecto de los valores de X es la que se denomina recta de regresión de Y sobre X y se define como $\hat{y} = b_0 + b_1 \cdot x$ donde $b_1 = \dfrac{S_{XY}}{S_X^2}$ y $b_0 = \overline{y} - \dfrac{S_{XY}}{S_X^2} \cdot \overline{x}$.

Es importante hacer notar que la recta de regresión de Y sobre X no coincide con la recta de regresión de X sobre Y.

■ **Ejemplo 2.16 — Continuación del ejemplo 2.15.** Calcular la recta de regresión de la *Resistencia* sobre la *Velocidad* y dibujarla junto con el diagrama de dispersión,

Solución

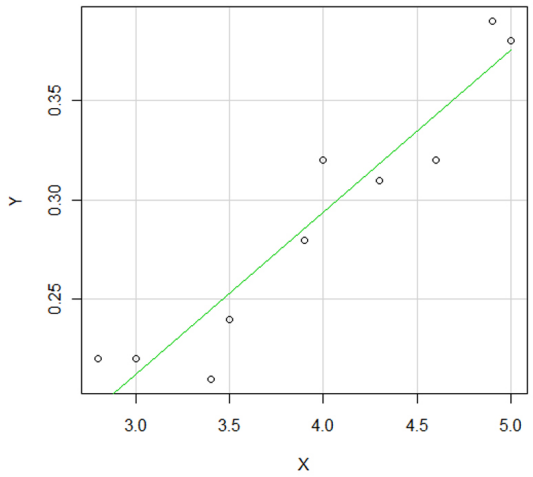
Nos vamos a **Estadísticos** → **Ajuste de Modelos** → **Regresión lineal**, en la casilla *Variable explicada* (*elegir una*), seleccionamos la variable Y (*Resistencia*), y en la casilla *Variables explicativas* (*elegir una o mas*), elegimos la variable X (*Velocidad*). Pulsamos aceptar, y obtenemos la siguiente salida:

```
Residuals:

      Min         1Q       Median         3Q          Max
  -0.034913  -0.011906  -0.000638    0.018903    0.026101$

Coefficients:
               Estimate      Std.    Error   t value   Pr(>|t|)
(Intercept)   -0.032672   0.038332  -0.852    0.419
     X         0.081643   0.009567   8.533   2.74e-05    ***


Residual standard error: 0.02199 on 8 degrees of freedom
Multiple R-squared: 0.901,  Adjusted R-squared: 0.8886
F-statistic: 72.82 on 1 and 8 DF,  p-value: 2.736e-05
```

La primera tabla, nos hace un estudio descriptivo de los errores comtidos al realizar el ajuste. La segunda tabla, nos da una estimación de los coeficientes del ajuste bajo la columna Estimate, la estimación para $b_0 = -0.032672$, y la de $b_1 = 0.081643$.

Al final, nos dan también el coeficiente de correlación al cuadrado (R-squared), que toma el valor 0.901.

Para realizar el gráfico que nos piden, nos vamos a Gráficas, Diagrama de dispersión, y dejamos marcado Línea de mínimos cuadrados tras seleccionar las variables correspondientes. El resultado es:



Como se observa en el gráfico, la recta de regresión es creciente, y los puntos se mantienen próximos a ella.

■

2.9 Ejercicio resuelto

A continuación vamos a tarbajar con un fichero de datos SPSS.

El fichero **bankloan.sav** es un fichero de datos de **SPSS** con datos hipotéticos sobre las iniciativas de un banco para reducir la tasa de moras de créditos. El archivo contiene información financiera y demográfica de 850 clientes anteriores y posibles clientes.

Tiene las siguientes variables (todas de tipo numérico) contenidas en el fichero de datos:

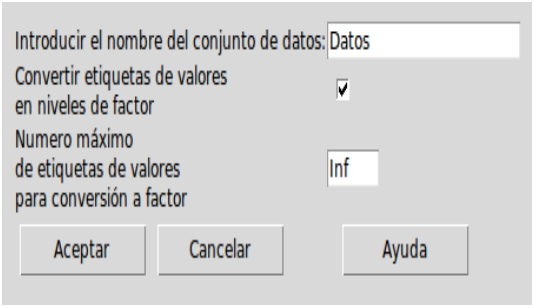
edad	edad en años
empleo	Años con la empresa actual
direccion	Años en la dirección actual
ingresos	Ingresos familiares en miles
deudaingr	Tasa de deuda sobre ingresos ($\times 100$)
deudacred	Deuda de la tarjeta de crédito en miles
deudaotro	Otras deudas en miles
morapred1	Impago pronosticado, modelo 1
morapred2	Impago pronosticado, modelo 2
morapred3	Impago pronosticado, modelo 3

Cuenta además con dos variable de tipo *carácter* con *etiquetas de valor*:

educ Nivel de educación
1 = "No completó el bachillerato"
2 = "Título de Bachiller"
3 = "Estudios Superiores iniciados"
4 = "Título Superior"
5 = "Título de Post-Grado"

impago Impagos anteriores
0 = "No"
1 = "Si"

R no trabaja con este tipo de variables. Si al importar datos desde SPSS, marcamos la opción **Convertir etiquetas de valores en niveles de factor** importará los caracteres ("No completó el bachillerato", ...).



Si desmarcamos esa opción, trabajará con las etiquetas de valores (1, 2, ...)

■ **Ejemplo 2.17** El fichero **bankloan.sav** (disponible en *moodle*), es un fichero producido con el programa *SPSS*. Queremos estudiar la relación existente entre algunas de sus variables. Se pide:

1. Calcular e interpretar el coeficiente de correlación entre las variables **Deuda de la tarjeta de crédito en miles** y la variable **Ingresos**.
2. Dibujar un diagráma de dispersión en el que se observe la variable **Deuda de la tarjeta de crédito en miles** con respecto a la variable **Ingresos**.
3. Dar la recta de regresión entre las variables **Tasa de deudas sobre ingresos** $x100$ (independiente), y **Impago pronosticado, modelo**3.
4. Dibujar un diagrama de dispersión en el que se observe la recta de regresión para las variables anteriores.

■

Solución

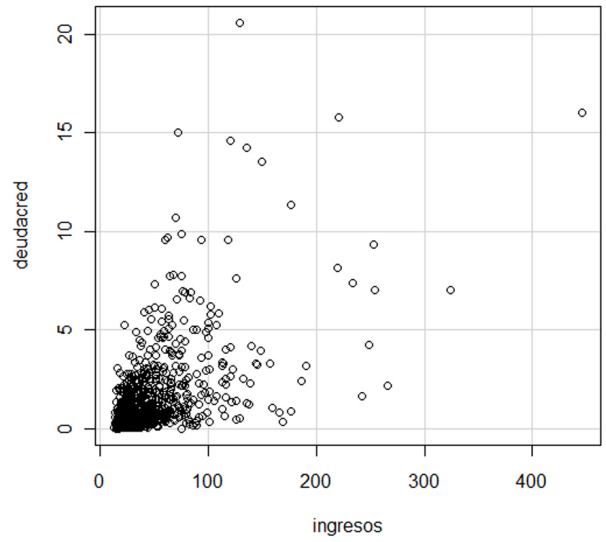
Primero, debemos abrir el fichero que nos piden, para ello, **Datos** → **Importar datos** → **desde datos SPSS**, y se introduce el nombre con el que se desea que se abra el fichero. A continuación, pulsamos sobre aceptar, y buscamos la carpeta en la que se encuentre el archivo.

(a) Al igual que en el ejemplo anterior, seguimos la secuencia: **Estadísticos** → **Resúmenes** → **Matriz de correlaciones**, seleccionamos las variables (en nuestro caso **deudacred** e **ingresos** y pulsamos `Aceptar`. Se obtiene

```
 > cor(Datos[,c("deudacred","ingresos")], use="complete.obs")
           deudacred  ingresos
deudacred 1.0000000 0.5515153
ingresos  0.5515153 1.0000000
```

El coeficiente de correlación es 0.55, lo que indica una baja relación lineal entre ambas variables de tipo creciente.

(b) Ir a **Gráficas** → **Diagrama de dispersión**. Introducir como variable *y* a **deudacred** y como variable *x* **ingresos**, eliminando las opciones que tenmos para que muestre distintas curvas. El resultado es:



Se observa claramente la tendencia creciente, y el bajo valor de *r* puede ser debido a que los puntos no se agrupan en torno a la forma de una recta, si no en la zona comprendida entre dos rectas.

(c) Pulsamos **Estadísticos** → **Ajuste de modelos** → **Regresión lineal**, y seleccionamos las variables en los recuadros adecuados (*deudaingr* en explicativa y *morapred*3 en explicada), dándole a *Aceptar* al terminar. La salida es:

Residuals:

Min	$1Q$	Median	$3Q$	Max
-0.089128	-0.021260	-0.007467	0.015433	0.067991

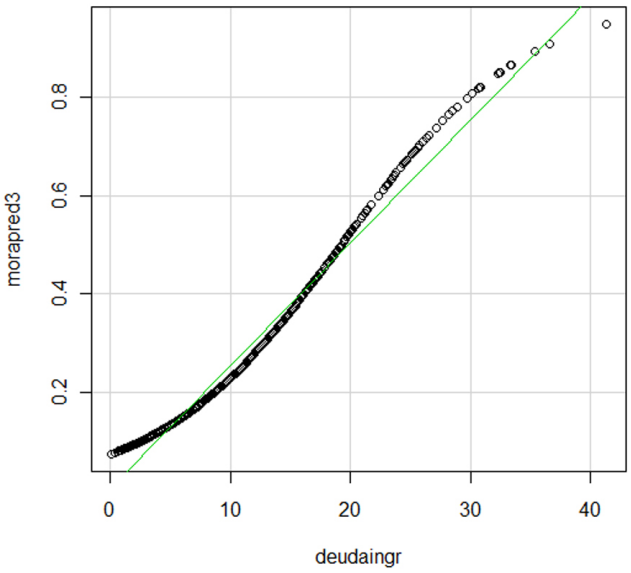
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0041446	0.0015346	2.701	0.00706 **
deudaingr	0.0250140	0.0001259	198.673	< 2e-16 ***

Residual standard error: 0.02465 on 848 degrees of freedom Multiple R-squared: 0.979, Adjusted R-squared: 0.9789 F-statistic: $3.947e+04$ on 1 and 848 DF, p-value: $< 2.2e-16$

Lo que nos lleva a concluir que la recta de regresión en: $\widehat{morapred3} = 0.0041446 + 0.0250140 * deudaingr$, con un coeficiente de correlación al cuadrado muy alto (0.979), lo que da idea de un buen ajuste lineal.

(d) Ir a **Gráficas** → **Diagrama de dispersión**. Introducir como variable *y* a **morapred3** y como variable *x* **deudaingr**, dejar marcado **línea de mínimos cuadrados**. El resultado es:



Nótese que, a pesar de lo que nos han dicho los datos analíticamente, el ajuste parece ajustarse mejor a otro tipo de gráfica que no es lineal.

2.10 Ejercicios propuestos

Ejercicio 2.7 La lluvia caída durante el año 1.983 en 10 estaciones meteorológicas de un cierto país fueron:

lluvias en mm^3	frecuencias
50 - 59	2
60 - 69	18
70 - 79	28
80 - 89	15
90 - 99	17
100 - 109	12
110 - 119	4
120 - 129	8
130 - 139	3
140 - 149	2
150 - 159	1

Calcular:

(a) Media, varianza y desviación típica.
(b) Coeficiente de variación de Pearson.
(c) Media, mediana y cuartiles.

■

Ejercicio 2.8 Las calificaciones de *Métodos Estadísticos* de un grupo de 20 alumnos en el examen de teoría y en el examen de prácticas han sido las siguientes:

Teoría	1.5	4	5	6	1	4	6	7.5	0	4.5
Prácticas	4	6	5	8.5	1	5	6	7.25	3.5	6.5

Teoría	5.5	8	1.8	4.25	6.6	9.5	4	6	6.75	7.75
Prácticas	6.5	9	4.5	5.6	5.25	7.5	2.5	5.4	5.75	9.5

(a) Crear las variable *NotTeor* (notas de teoría) y la variable *NotPrac* (notas de prácticas) con las notas de teoría y prácticas respectivamente.
(b) Calcular la variable *NotFinal* (nota final) que viene dada por la expresión

$$NotFinal = \frac{7.5 \cdot NotTeor + 2.5 \cdot NotPrac}{10},$$

(c) Dibujar un histograma de frecuencias absolutas para la variable *NotFinal*.
(d) Agrupar las notas finales en SUSPENSO, APROBADO, NOTABLE y SOBRESALIENTE en una variable de nombre *Calific*. Construir una tabla de frecuencias y dibujar un diagrama de sectores para esta variable.

■

Ejercicio 2.9 El peso de un objeto fué medido 30 veces obteniéndose los siguientes resultados:

6.120	6.129	6.116	6.114	6.112	6.119	6.119	6.121	6.124	6.127
6.113	6.116	6.117	6.126	6.123	6.123	6.122	6.118	6.120	6.120
6.121	6.124	6.114	6.121	6.120	6.116	6.113	6.111	6.123	6.124

1. Sin agrupar en intervalos, hallar la media, la mediana, la moda y la desviación típica.
2. Repetir los cálculos agrupando los datos, de forma apropiada, en 5 intervalos de longitud 0.004 y comparar los resultados obtenidos.

■

Ejercicio 2.10 Un investigador está interesado en el estudio de nuevos métodos para medir la concentración de hidrógeno. Se recogieron muestras de mediciones clasificándolas de acuerdo a las variables *concentración de hidrógeno determinada con un método de cromatografía de gases* (X) y *concentración determinada con un nuevo método de sensor* (Y):

X	47	114	118	62	65	124	198	221	127	140
Y	38	117	116	62	53	127	200	215	114	134

X	70	78	95	100	70	140	150	152	164	140
Y	84	79	93	106	67	142	170	149	154	139

(a) Calcular algunas de las medidas estudiadas (media, varianza, cuartiles, asimetría, etc) para cada una de las variables.
(b) Calcular la covarianza entre ambas variables.
(c) Calular el coeficiente de correlación lineal y el coeficiente de determinación (el cuadrado del coeficiente de correlación lineal) e interpreta los resultados.
(d) Mostrar en una misma gráfica el diagrama de dispersión y la recta de regresión.

■

Ejercicio 2.11 Se han tomado cinco muestras de glucógeno de una cantidad fija de cada una. Se les ha aplicado una cantidad de glucogenasa (en milimoles/litro) anotando en cada caso la

velocidad de reacción medida en μ–moles/minuto. Obteniéndose así la siguiente tabla:

Cantidad glucog.	1	2	3	0.2	0.5
Velocidad reacción	18	35	60	8	10

Se pide:

(a) Estudiar las medidas de posición central estudiadas para la velocidad de reacción.
(b) >En cuál de las dos variables la media es más representativa? Justificar la respuesta.
(c) >Se deduce de estos datos que la velocidad de reacción aumenta con la concentración de glucogenasa? Justificar detalladamente la respuesta.
(d) Si a una de las muestras le hubiésemos aplicado una concentración de glucogenasa de 5 milimoles/litro, >cuál hubiera sido la velocidad de reacción?

■

Ejercicio 2.12 El fichero `DatosEmpleados.xls` contiene información de los empleados de una empresa. EL fichero contiene una matriz de datos con la siguiente información (por columnas):

- **Columna 1:** Sexo del empleado (m = Mujer, h =Hombre)
- **Columna 2:** Nivel educativo del empleado
- **Columna 3:** Categoría laboral (0 = Ausente, 1 = Administrativo, 2 = Seguridad, 3 = Directivo).
- **Columna 4:** salario actual del empleado
- **Columna 5:** salario inicial
- **Columna 6:** Meses que lleva contratado
- **Columna 7:** Experiencia Previa (en meses)

Se pide:

(a) Representar la variable `sexo` mediante un diagrama de sectores.
(b) Crear la variable `aumento` que indique el porcentaje de aumento de sueldo de cada empleado respecto a salario inicial.

$$\text{aumento} = (\text{salario} - \text{salarioInicial})/\text{salarioInicial}$$

(c) Calcular la media de esta variable así como la media de aumento según el sexo del trabajador.
(d) Estudiar la relación lineal entre las variables `aumento` y la antigüedad. Calcular y dibujar la recta de regresión.

■

Ejercicio 2.13 Las calificaciones de 40 alumnos obtenidas en el examen parcial (X) y en el final (Y) de una asignatura han sido las siguientes:

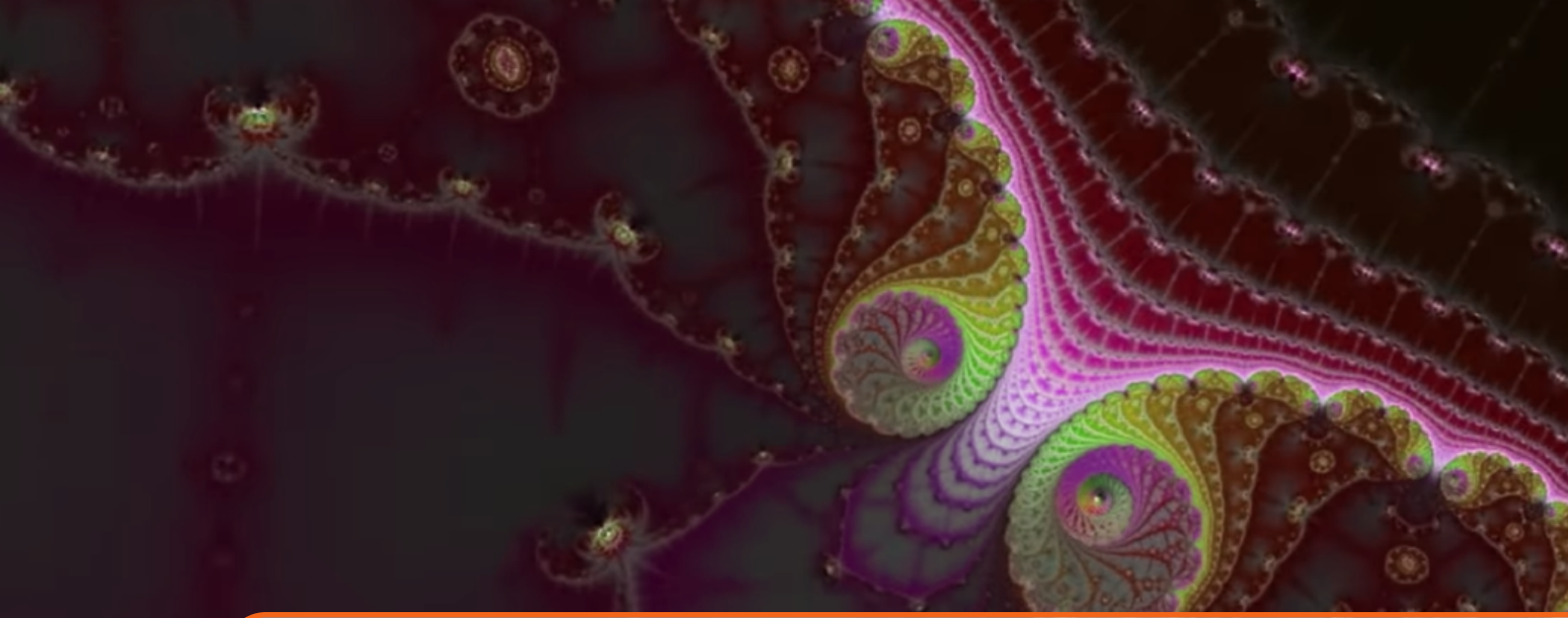
X	4	5	1	6	1	2	2	4	5	6	8	0	2	10	4	8	2	6	6	5
Y	2	7	3	3	3	1	0	2	7	6	9	3	3	10	8	7	0	3	6	3

X	8	9	9	8	5	2	4	3	0	2	2	5	4	7	6	10	6	3	9	0
Y	7	8	10	7	3	3	2	0	3	0	0	3	6	5	6	7	6	0	8	9

(a) Calcular, para cada una de las variables, la media, varianza, mediana y los deciles.
(b) Para la nota final, calcular también los coeficientes de asimetría y curtosis.
(c) Calcular la covarianza entre ambas variables y el coeficiente de correlación lineal. >Qué podemos decir de la relación entre ambas variables?

(c) Dibujar en una misma gráfica el diagrama de dispersión y la recta de regresión.

■



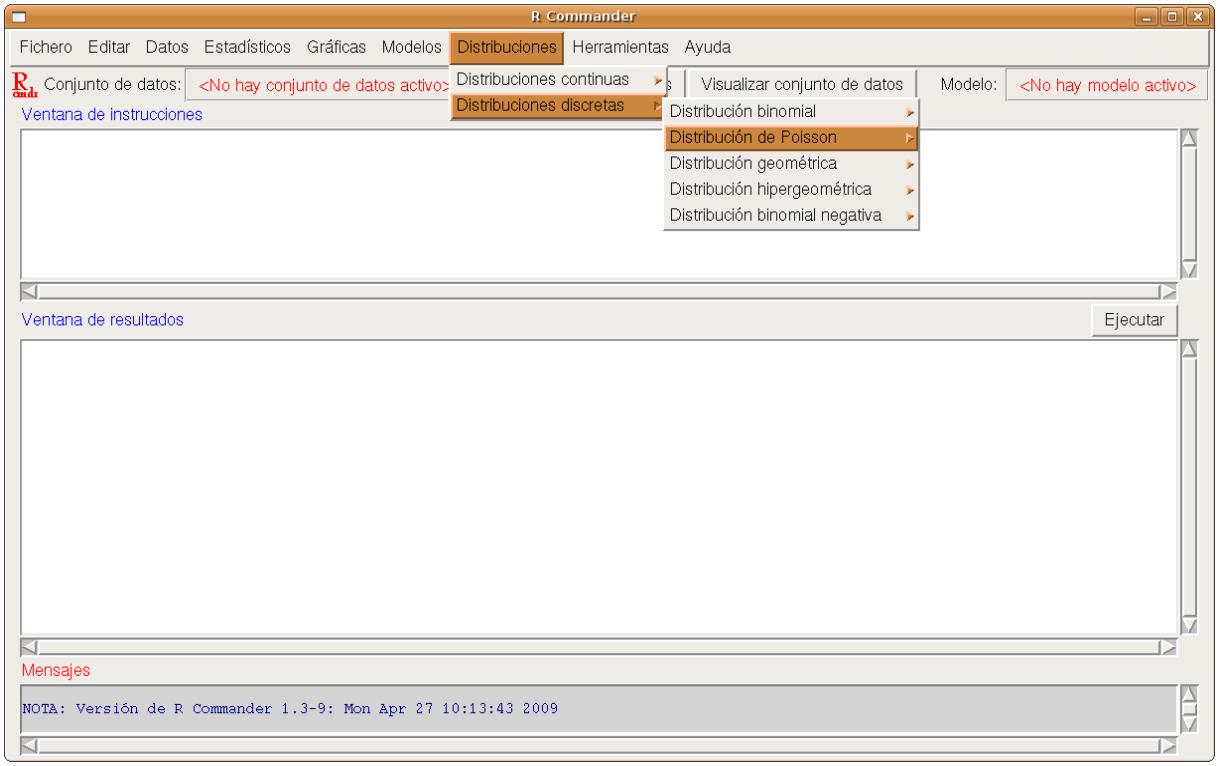
3. Variables Aleatorias y Modelos de Distribucione

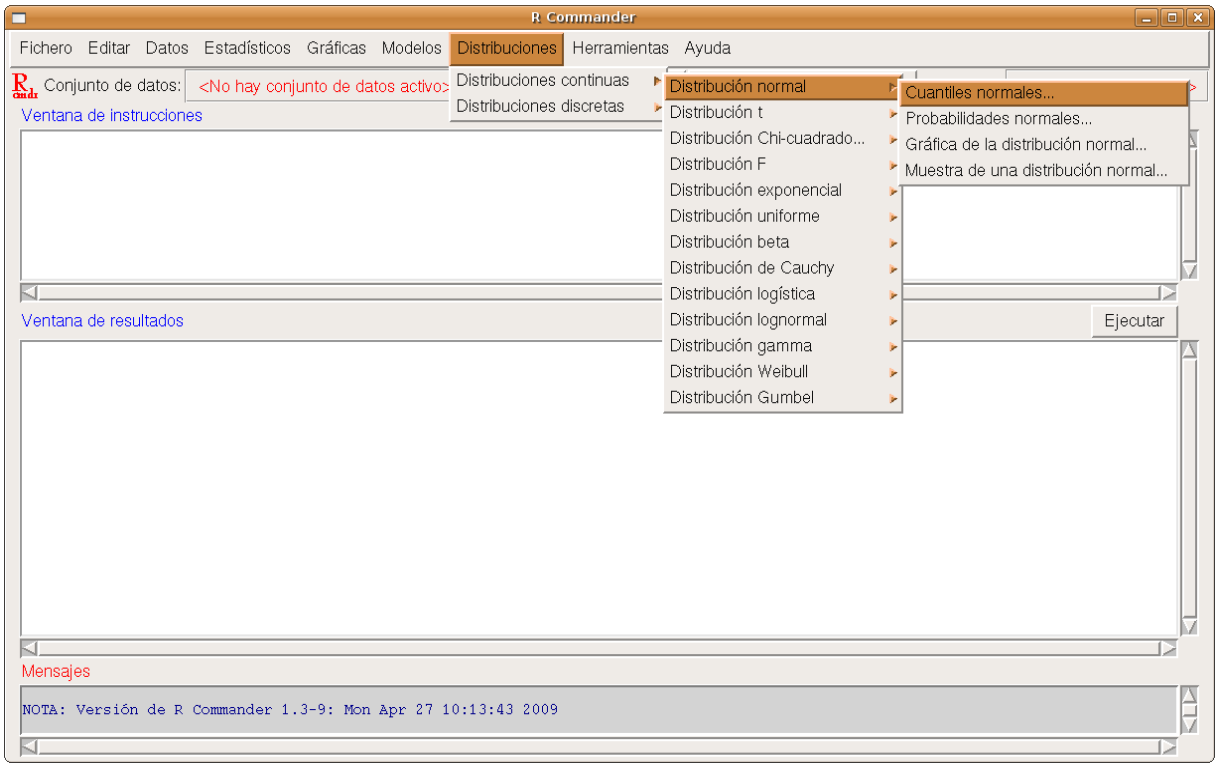
Existen un conjunto de funciones R que gestionan el cálculo de la función de densidad o probabilidad, de la función de distribución, de los cuantiles (que son los valores de la función inversa de la función de distribución), o de una muestra aleatoria de una variable aleatoria discreta o continua.

En la siguiente tabla exponemos algunas de las funciones de las distribuciones disponibles en R.

Distribuciones de probabilidad			
Nombre en R	**Distribución**	**Parámetros**	**Valores por defecto de los parámetros**
`unif`	Uniforme	min, max	0, 1
`norm`	Normal	mean, sd	0, 1
`t`	t de Student	df (*gr. libertad*)	
`exp`	Exponencial	rate (*tasa*)	1
`f`	F de Snedecor	df1, df2 (*grs. libertad*)	
`chisq`	Chi–cuadrado	df (*gr. libertad*)	
`binom`	Binomial	size, prob	
`pois`	Poisson	lambda	
`geom`	Geométrica	prob	

Estas funciones, junto con otras, es lo que nos muestra `R-Commander` tanto en el caso continuo como en el discreto.


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: <No hay conjunto de datos activo>
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
Distribuciones continuas
Distribuciones discretas
Distribución binomial
Distribución de Poisson
Distribución geométrica
Distribución hipergeométrica
Distribución binomial negativa
Ventana de instrucciones
Ventana de resultados
Ejecutar
Mensajes
NOTA: Versión de R Commander 1.3-9: Mon Apr 27 10:13:43 2009


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: <No hay conjunto de datos activo>
Distribuciones continuas
Distribuciones discretas
Distribución normal
Distribución t
Distribución Chi-cuadrado...
Distribución F
Distribución exponencial
Distribución uniforme
Distribución beta
Distribución de Cauchy
Distribución logística
Distribución lognormal
Distribución gamma
Distribución Weibull
Distribución Gumbel
Cuantiles normales...
Probabilidades normales...
Gráfica de la distribución normal...
Muestra de una distribución normal...
Ventana de instrucciones
Ventana de resultados
Ejecutar
Mensajes
NOTA: Versión de R Commander 1.3-9: Mon Apr 27 10:13:43 2009

El nombre de dichas funciones `R` debe comenzar por **d**, **p**, **q** o **r**. El significado de estas letras es el siguiente:

d **Función de probabilidad o densidad** (`dbinom, dnorm, dpois, dt, ...`).
Ejemplo: `dbinom(4, size=6, prob=0.3)`
Nos da el valor $P[X = 4]$ siendo $X \sim \text{Bi}(6, 0.3)$.

p **Pobabilidades acumuladas o función de distribución** (`pbinom, pnorm, ...`).
Para un valor dado c de una variable aleatoria calcula $P(x \leq c)$ ó $P(x > c)$
(**¡OJO!** A la derecha el mayor es **estricto**).
Ejemplo `pnorm(c(12), mean=11, sd=2, lower.tail=TRUE)`
Nos da el valor $P[X \leq 12]$ siendo $X \sim N(11, 2^2)$. El argumento `lower.tail=TRUE` es quién nos indica $P[X \leq 12]$. En caso contrario, `lower.tail=FALSE`, se obtendría $P[X > 12]$ (mayor estricto).

q **Cuantiles** (`qbinom, qnorm, qpois, qt, ...`).
Es el valor c_p tal que para una probabilidad dada p: $P[X \leq c_p] = p$.
Ejemplo `qpois(c(0.9), lambda=4, lower.tail=TRUE)`.
Nos da el valor k de la variable Poisson de media $\lambda = 4$ ($X \sim \mathscr{P}(4)$), tal que $P[X \leq k] = 0.9]$.

r **Genera números aleatorios** (`rbinom, rnorm, rpois, rt, ...`).
No lo estudiaremos.

También se pueden obtener las **gráficas** de la función de densidad (caso continuo) o de la de probabilidad (caso discreto), así como de las funciones de distribución de probabilidades acumuladas en ambos casos. Las estudiaremos más adelante.

3.1 Variables aleatorias discretas

3.1.1 Distribución binomial

Ejercicio 3.1 Un examen consta de 10 preguntas a las que hay que contestar `Verdadero` o `Falso`. Suponiendo que una persona contesta al azar las 10 preguntas, se pide:

1. Representar gráficamente las funciones de probabilidad y de distribución de la variable aleatoria que mide el número de aciertos.
2. La probabilidad de obtener 5 aciertos.
3. La probabilidad de obtener algún acierto.
4. Probabilidad de obtener al menos 5 aciertos.
5. Calcular el valor correspondiente al tercer cuartil.

■

Solución

1. *Representar gráficamente las funciones de probabilidad y de distribución de la variable aleatoria que mide el número de aciertos.*
   La variable aleatoria que mide el número de aciertos es una variable aleatoria binomial de parámetros $n = 10$ y $p = P(\text{acertar la pregunta}) = \dfrac{1}{2} = 0.5$, es decir, $X \sim B(10, 0.5)$.
   Para representar gráficamente la función de probabilidad, seleccionamos desde el menú de `R-Commander` las siguientes opciones **Distribuciones** → **Distribuciones discretas** → **Distribución binomial** → **Gráficas de la distribución binomial**, tal como puede verse en la gráfica de la Figura 3.1
   En la ventana emergente, hecemos la selección, `Ensayos binomiales:` 10, `Probabilidad de éxito:` 0.5, y como gráfica seleccionamos la `Gráfica de la función de probabilidad`, como aparece en la Figura 3.2(a)
   Pulsamos aceptar y obtenemos la representación gráfica de la $B(10, 0.5)$, como se puede observar en la Figura 3.3(a).

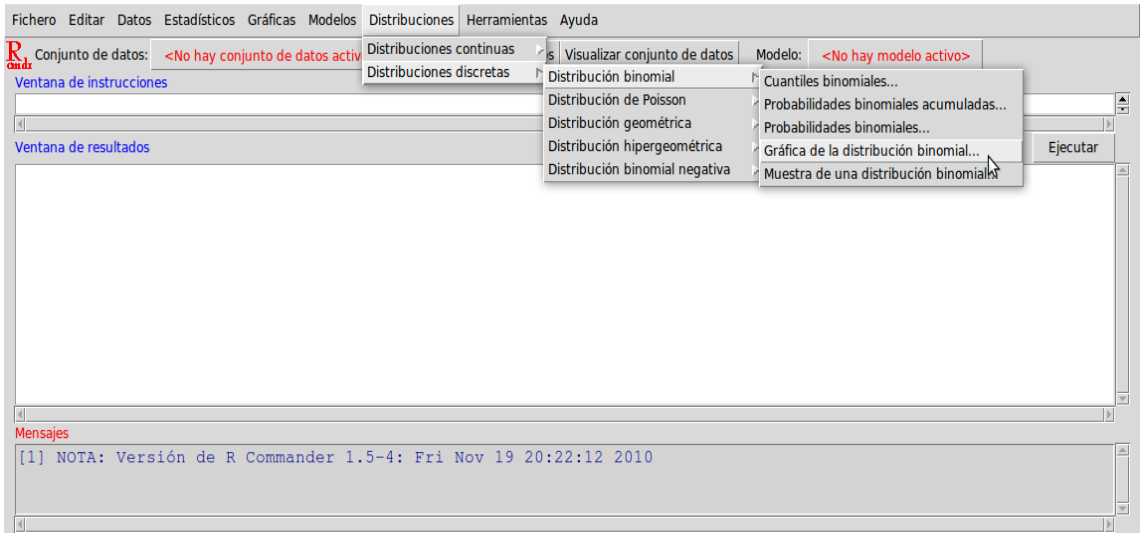
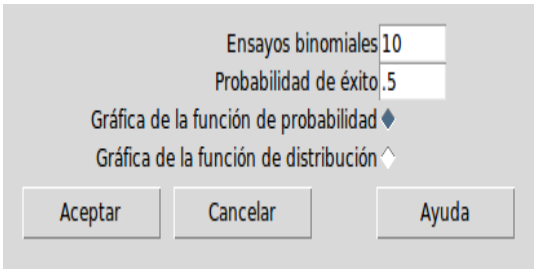
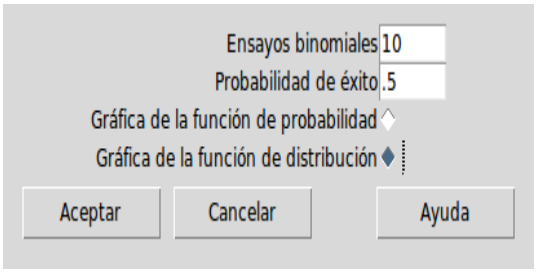


Figure 3.1: Selección para la gráfica de la distribución binomial.



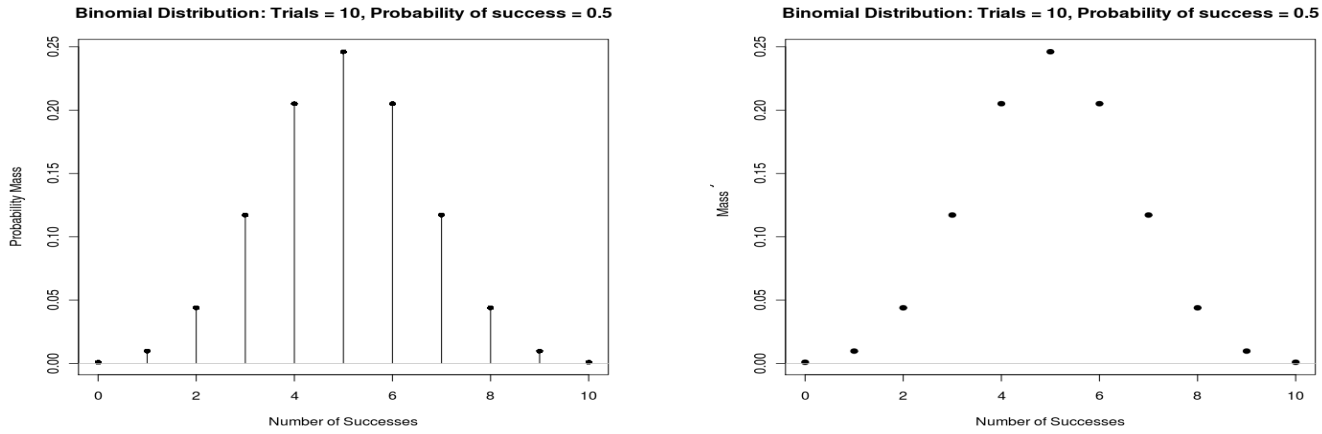
(a)



(b)

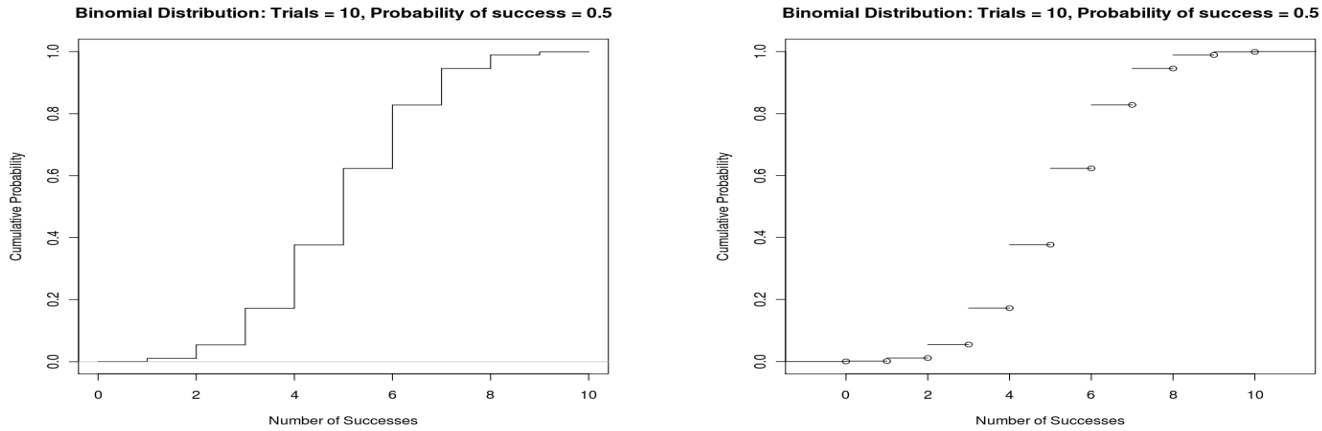
Figure 3.2: Selección de la gráfica de la distribución $B(10, 0.5)$.

Si no queremos que aparezcan las líneas verticales, simplemente cambiamos en el código generado por `R-Commander`, donde pone **type="h"** por **type="p"**, seleccionamos todo el código, ejecutamos y aparece la gráfica de la Figura 3.3(b).
Para representar la **función de distribución** repetimos el procedimiento, pero seleccionando la opción `Gráfica de la función de distribución` (ver Figura 3.2(b)). Obtenemos así la gráfica de la Figura 3.4(a).
Esta gráfica no es realmente la gráfica de una función (no deben aparecer las líneas verticales). Para que la gráfica sea efectivamente la de una función, podemos introducirnos en la ventana de código de `R-Commander` e introducir algunos cambios para utilizar la función **stepfun()** de `R` y dibujar de manera más elegante la función de distribución.
Los posibles valores de X son $\{0, 1, \ldots, 10\}$ (en `R` simplemente **0:10**). La función **stepfun** necesita dos vectores. En el primero de ellos ponemos los valores que toma la variable. El segundo vector de **stepfun()** ha de tener un elemento más (a la izquierda del primero) y pondremos el **-1** (`c(-1,0:10)`), es decir, formaremos un nuevo vector con el número -1 y los números comprendidos entre 0 y 10. Hemos de eliminar también la referencia **[-length(x)]** y la referencia al tipo de gráfica (**type="l"**). No nos hemos de olvidar de añadir los argumentos **right=TRUE** y **verticals=FALSE** par que los intervalos sean cerrados a la derecha y para que no dibuje las líneas verticales en los puntos de discontinuidad.
En resumen, nos situamos en la ventana de códigos de R-Commander y las instrucciones



(a) (b)

Figure 3.3: Función de probabilidad de la distribución $B(10, 0.5)$.



(a) (b)

Figure 3.4: Función de distribucion de la distribución $B(10, 0.5)$.

generadas por `R-Commander` al dibujar la gráfica anterior.

```
plot(.x[-1], pbinom(.x, size=10, prob=0.5)[-length(.x)],
  xlab="Number of Successes", ylab="Cumulative Probability", main="Binomial
  Distribution: Trials = 10, Probability of success = 0.5", type="l")
abline(h=0, col="gray")
```

las modificamos como sigue

```
plot(stepfun(0:10, pbinom(c(-1,0:10), size=10, prob=.5), right=TRUE), verticals=FALSE,
   xlab="Number of Successes", ylab="Cumulative Probability",
      main = "Binomial Distribution: Trials = 10, Probability of success = 0.5")
```

Seleccionamos con el ratón estas instrucciones y pulsamos sobre Ejecutar en `R-Commander`. Obtenemos así la gráfica de la función de distribución como se observa en la Figura 3.4(b).

2. *La probabilidad de obtener 5 aciertos.*
   Para ello seleccionamos desde el menú de `R-Commander` **Distribuciones** → **Distribuciones discretas** → **Distribución binomial** → **Probabilidades binomiales**
   Se obtienen así todos los valores de la función de probabilidad. Seleccionamos el valor deseado:

0	0.0009765625
1	0.0097656250
2	0.0439453125
3	0.1171875000
4	0.2050781250
5	**0.2460937500**
6	0.2050781250
7	0.1171875000
8	0.0439453125
9	0.0097656250
10	0.0009765625

`R-Commander` nos devuelve la probabilidades completa. Si quisiéramos sólo el valor corespondiente para $X = 5$, en la ventana de comandos de `R-Commander` escribimos el correspondiente código `R`

```
> dbinom(c(5), size=10, prob=0.5)
  [1] 0.2460938
```

3. *La probabilidad de obtener algún acierto.*
   Es decir, $P[X \geq 1]$ (o, lo que es lo mismo, $P[X > 0]$)
   Seleccionamos **Distribuciones** → **Distribuciones discretas** → **Distribución binomial** → **Probabilidades binomiales acumuladas (cola derecha y valor 0)**.

```
> pbinom(c(0), size=10, prob=0.5, lower.tail=FALSE)
  [1] 0.9990234
```

Nota 3.1.1 También se podría haber hecho por el suceso contrario. Valiéndonos de la tabla calculada en el apartado anterior, se tiene:
$P[X > 0] = 1 - P[X = 0] = 1 - 0.0009766 = 0.9990234$ donde hemos redondeado el valor $P[X = 0]$ al mismo número de cifras decimales que el valor obtenido a través de las probabilidades acumuladas en la cola derecha. Repetimos los mismos comandos, pero ahora con cola izquierda y después modificamos la orden en la ventana de comandos. Al ejecutar se obtiene:

```
> 1-pbinom(c(0), size=10, prob=0.5, lower.tail=TRUE)
  [1] 0.9990234
```

4. *Probabilidad de obtener al menos 5 aciertos.*
   Nos piden $P[X \geq 5]$ (o lo que es lo mismo, $P[X > 4]$).
   procedemos como en el apartado anterior: **Distribuciones** → **Distribuciones discretas** → **Distribución binomial** → **Probabilidades binomiales acumuladas (cola derecha y valor 4)**.

```
> pbinom(c(4), size=10, prob=0.5, lower.tail=FALSE)
  [1] 0.6230469
```

5. *Calcular el valor correspondiente al tercer cuartil.*
   El tercer cuartil se corresponde con el percentil 75. Se sigue la secuencia: **Distribuciones** → **Distribuciones discretas** → **Distribución binomial** → **Cuantiles binomiales**.
   A continuación se rellenan los valores solicitados: en probabilidad ponemos 0.75, en ensayos, 10 y en probabilidad de éxito, 0.5, obteniéndose.

```
> qbinom(c(0.75), size=10, prob=0.5, lower.tail=TRUE)
[1] 6
```

es decir, es el valor 6.

3.1.2 Distribución de Poisson

■ **Ejemplo 3.1** La centralita telefónica de un hotel recibe un nº de llamadas por minuto que sigue una ley de Poisson con parámetro $\lambda = 0.5$. Determinar las probabilidades:

(a) De que en un minuto al azar, se reciba una única llamada.
(b) De que en un minuto al azar se reciban un máximo de dos llamadas.
(c) De que en un minuto al azar, la centralita quede bloqueada, sabiendo que no puede realizar más de 3 conexiones por minuto.
(d) Se reciban 5 llamadas en dos minutos.

■

Solución

Teneos una variable X que sigue una distribución de Poisson de parámetro $\lambda = 0.5$, $X \sim \mathscr{P}(0.5)$.

(a) Hemos de calcular $P[X = 1]$. Mediante la secuencia: **Distribuciones** → **Distribuciones discretas** → **Distribución de Poisson** → **Probabilidades de Poisson** no se obtiene el valor de `Pr[X = 1]`, sino una tabla:

```
> .Table <- data.frame(Pr=round(dpois(0:5, lambda=0.5), 4))
> rownames(.Table) <- 0:5
> .Table
      Pr
0 0.6065
1 0.3033
2 0.0758
3 0.0126
4 0.0016
5 0.0002
> remove(.Table)
```

Si sólo se quiere la `Pr[X=1]`, simplemente escribiendo en la ventana de comandos de `R-Commander`, el comando `dpois` (de `R`), seleccionándolo y pulsando sobre Ejecutar en la ventana de `R-Commander`. Se obtiene

```
> dpois(1, lambda=0.5)
[1] 0.3032653
```

(b) Hay que calcular `P[X` ≤ 2`]`. La secuencia es: **Distribuciones** → **Distribuciones discretas** → **D. Poisson** → **Probabilidades acumuladas.**. En la ventana emergente, le damos el valor `lambda0=.5`, el valor 2 a la variable, y seleccionamos `cola izquierda`.
La instrucción `R` y el resultado que se genera:

```
> ppois(c(2), lambda=0.5, lower.tail=TRUE)
[1] 0.9856123
```

(c) Hay que calcular: `P[X` > 3`]`. La secuencia es: **Distribuciones** → **Distribuciones discretas** → **D. Poisson** → **Probabilidades acumuladas.**. En la ventana emergente, le damos el valor `lambda=0.5`, el valor 3 a la variable y marcamos `cola derecha`.
El comando en código `R` que aparece en la ventana de comandos es:

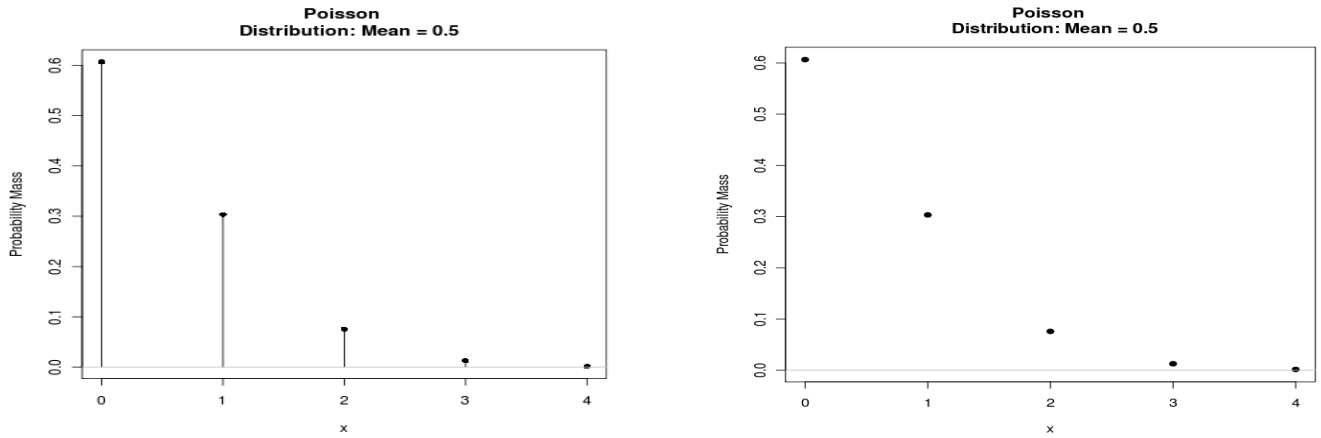
```
> ppois(c(3), lambda=0.5, lower.tail=FALSE)
[1] 0.001751623
```

(d) En este último caso, tenemos dos variables Poisson, ambas $\mathscr{P}(0.5)$, es decir, $X_1, X_2 \sim \mathscr{P}(0.5)$. Nos piden la probabiliadd de que entre ambas (la suma) haya 5 llamadas. Por la propiedad de reproductividad: $Y = X_1 + X_2 \sim \mathscr{P}(1)$, con lo que la instrucción R para la respuesta:

```
> dpois(5, lambda=1)
[1] 0.003065662
```

Gráfica de la distribución de Poisson

Para representar gráficamente la función de probabilidad, seleccionamos desde el menú de `R-Commander` las siguientes opciones **Distribuciones** → **Distribuciones discretas** → **Distribución Poisson** → **Gráficas de la distribución Poisson**, y posteriormente, en la ventana emergente, damos el valor `lambda=0.5` y marcamos la opción `Gráfica de la función de probabilidad`, obteniéndose la gráfica de la Figura 3.5(a). Al igual que hicimos con la binomial, si cambiamos **type="h"** por **type="p"**, obtenemos la gráfica sin líneas vericales como se muestra en la Figura 3.5(b)[1].



(a) (b)

Figure 3.5: Función de distribucion de la distribución $\mathscr{P}(0.5)$.

Para representar la **función de distribución**, repetimos el procedimiento, pero seleccionando la opción `Gráfica de la función de distribución`. Al igual que ocurría con la distribución binomial, la representación no es de una función (por las líneas verticales) por lo que utilizaremos la función **stepfun()**.

En el primer vector de **stepfun**, ponemos los valores de la variable que `R` considera que tienen probabilidad significativa, en nuestro caso $\{0,1,2,3,4\}$), (en `R`, **0:4**). El segundo vector de **stepfun()** ha de tener un elemento más. Dado que las variables Poisson no pueden tomar valores negativos, lo elegimos a la derecha del último, o sea, el **5** (el segundo vector es `c(0:4,5)`). Hemos de eliminar también la referencia **[-length(x)]** y la referencia al tipo de gráfica (**type="l"**). No nos hemos de olvidar de añadir los argumentos **right=TRUE** y **verticals=FALSE** como en la binomial. En resumen, nos situamos en la ventana de códigos de `R-Commander` y las instrucciones

```
plot(.x, dpois(.x, lambda=0.5), xlab="x", ylab="Probability Mass",
main="Poisson  Distribution: Mean = 0.5", type="h")
points(.x, dpois(.x, lambda=0.5), pch=16)
```

[1] En la gráfica observamos que aunque una variable aleatoria que sigue una distribución de Poisson puede tomar infinitos valores, en éste caso, a partir de la probabilidad $P[X = 5]$, estos valores son tan pequeños que `R` no los considera.

```
abline(h=0, col="gray")
```

las modificamos como sigue

```
plot(stepfun(0:4, ppois(c(0:4,5), lambda=0.5), right=TRUE), verticals=FALSE,
    xlab="x", ylab="Probability Mass",
  main="Poisson Distribution: Mean = 0.5")
```

Seleccionamos con el ratón estas instrucciones y pulsamos sobre Ejecutar en R-Commander. Obtenemos así la gráfica de la función de distribución

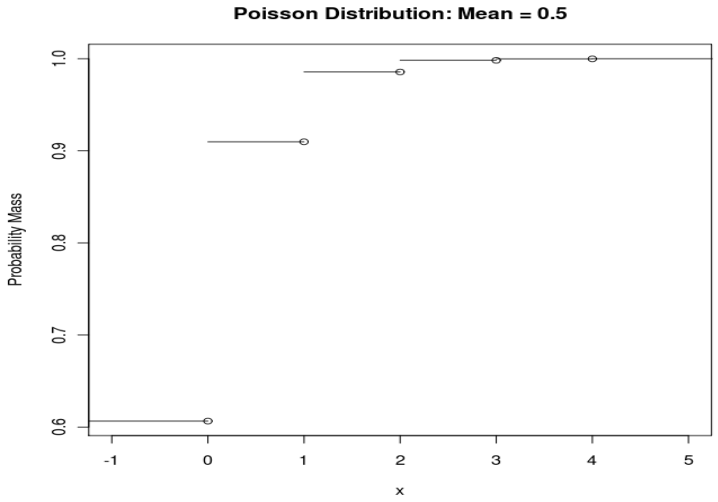


Figure 3.6: Función de distribucion de la distribución $\mathscr{P}(0.5)$.

3.2 Variables aleatorias continuas

3.2.1 Variable aleatoria Normal

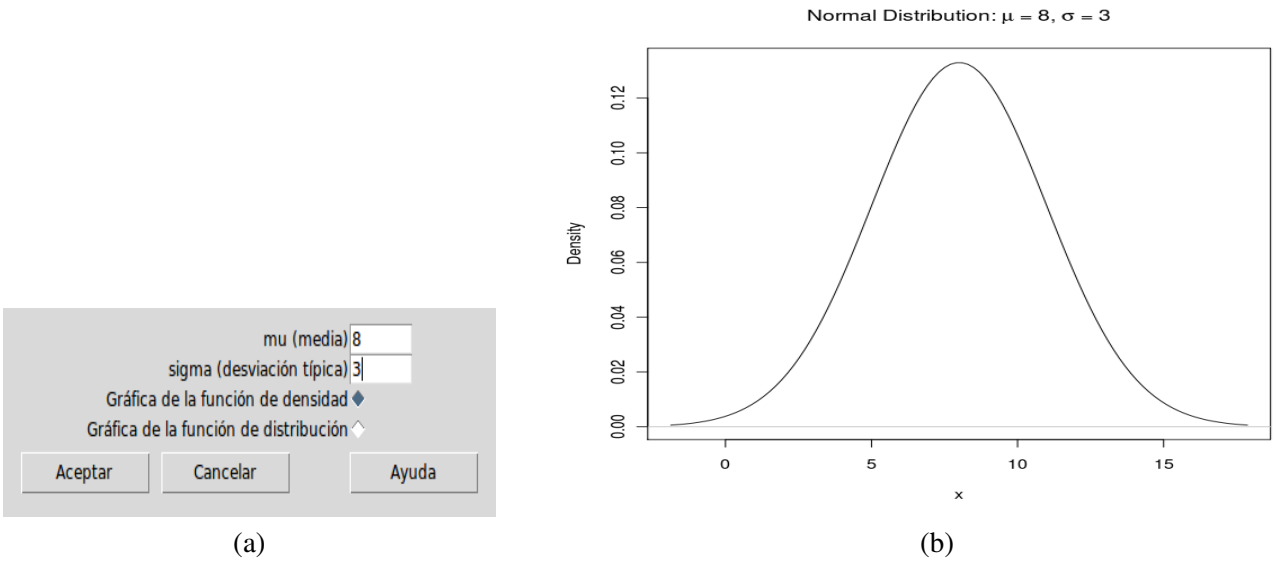
Ejercicio 3.2 Se supone que la estancia de los enfermos en un hospital sigue una distribución normal de media 8 días y desviación típica 3 días. Se pide

(a) Representar gráficamente sus funciones de densidad y de distribución.
(b) Calcular la probabilidad de que la estancia de un enfermo sea inferior a 7 días.
(c) Calcular la probabilidad de que la estancia de un enfermo sea superior a 3 días.
(d) Calcular la probabilidad de que la estancia de un enfermo esté comprendida entre 10 y 12 días.
(e) Calcular los percentiles 5 y 87.

■

Solución

(a) *Representar gráficamente sus funciones de densidad y de distribución.*
Para representar gráficamente la función de densidad de la variable aleatoria $N(8,3)$ seleccionamos: **Distribuciones** $\rightarrow$ **Distribuciones continuas** $\rightarrow$ **Distribución Normal** $\rightarrow$ **Gráficas de la distribución Normal**, rellenando la ventana emergente con los datos de la Figura 3.7(a). Obtenemos así la gráfica de la función de densidad (Figura 3.7(b)).



(a) (b)

Figure 3.7: Selección de parámetros y gráfica de la función de densidad de la distribución $N(8, 3)$.

Si en lugar de haber elegido en los datos de la Figura 3.7, *Gráfica de le función de densdad* hubiésemos elegido **Gráfica de la función de distribución** se obtiene

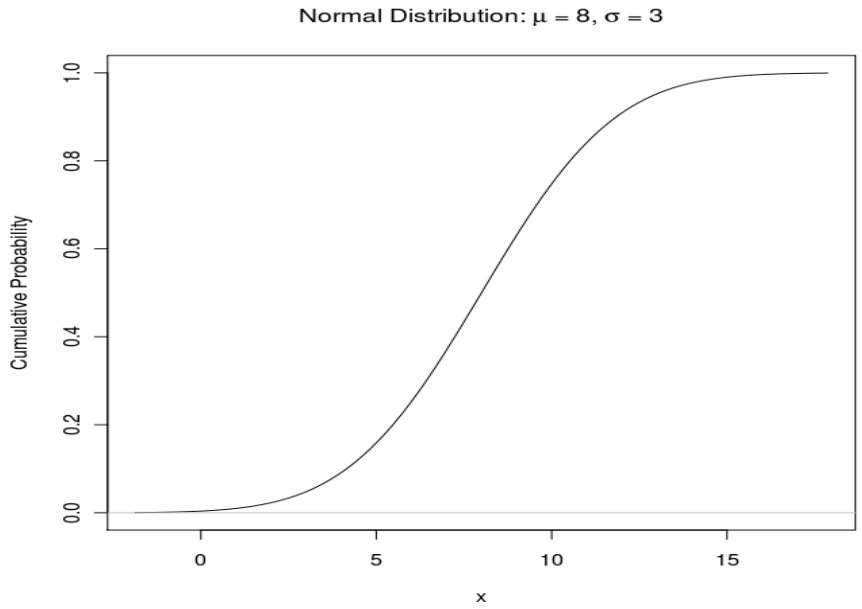


Figure 3.8: Función de distribución de la variable aleatoria $N(8, 3^2)$.

(b) *Calcular la probabilidad de que la estancia de un enfermo sea inferior a 7 días.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** A continuación, introducimos los datos como en la Figura 3.9.

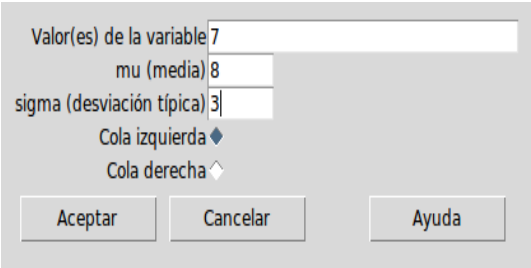


Figure 3.9: Selección de parámetros para el cálculo de $P[X \leq 7]$.

Se obteniene el valor **0.3694413**.

(c) *Calcular la probabilidad de que la estancia de un enfermo sea superior a 3 días.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** A continuación, introducimos los datos como en la Figura 3.10 obteniéndose el valor **0.9522096**.

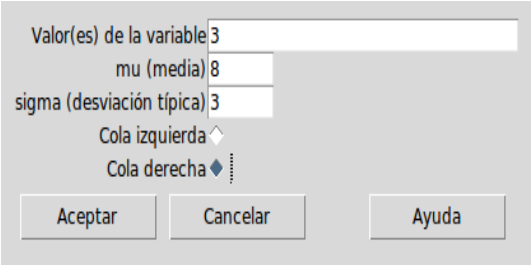


Figure 3.10: Selección de parámetros para el cálculo de $P[X > 3]$.

(d) *Probabilidad de que la estancia de un enfermo esté comprendida entre 10 y 12 días.*
$P[10 \leq X \leq 12] = P[X \leq 12] - P[X < 10]$. Calculamos cada una de ellas por separado, mediante **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** y seleccionando los parámetros de la Figura 3.11.

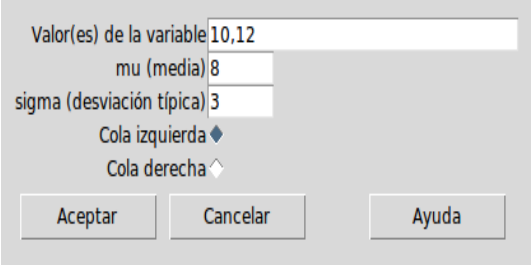


Figure 3.11: Selección de parámetros para el cálculo de $P[X \leq 12]$ y $P[X < 10]$.

A la vista de los resultados $P[X \leq 12] - P[X < 10] = 0.9087888 - 0.7475075 = 0.1612813$.
O bien, ayudándonos del código R que se genera en la ventana de comandos, escribir

```
> pnorm(12, mean=8, sd=3, lower.tail=TRUE) - pnorm(10, mean=8, sd=3, lower.tail=TRUE)
[1] 0.1612813
```

(e) *Calcular los percentiles 5 y 87.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **cuantiles normales** y marcamos los valores correspondientes a los cuantiles 0.05 y 0.87, tal como se indica en la Figura 3.12

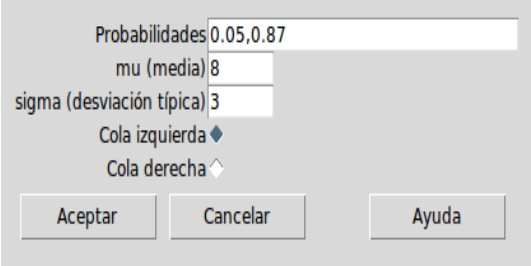


Figure 3.12: Cuantiles de la distribución normal.

La orden en R (y su respuesta) que aparecen en la ventana de comandos es:

```
> qnorm(c(0.05,0.87), mean=8, sd=3, lower.tail=TRUE)
[1]  3.065439 11.379173
```

Ejercicio 3.3 El diámetro de los árboles (pulg.) de un bosque sigue una distribución normal con $\mu = 8.8$ y $\sigma = 2.8$.

a. >Cuál es la probabilidad de que el diámetro de un árbol sea por lo menos 10 pulg?
b. >Cuál es la probabilidad de que el diámetro de un árbol sea mayor que 20 pulg?
c. >Cuál es la probabilidad de que el diámetro de un árbol esté entre 5 y 10 pulg?
d. >Qué valor de c es tal que el intervalo (8.8-c,8.8+c) incluya el 98% de los valores de diámetro?
e. Si se elige a cuatro árboles de forma independiente, >cuál es la probabilidad de que por lo menos uno tenga un diámetro mayor que 10 pulg.?

■

Solución

(a) *Probabilidad de que el diámetro de un árbol sea por lo menos 10 pulgadas.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** A continuación, introducimos los datos (no debemos olvidarnos marcar **Cola derecha**) como en la Figura 3.13 obteniéndose el valor

```
 > pnorm(c(10), mean = 8.8, sd = 2.8, lower.tail = FALSE)
[1] 0.3341176
```

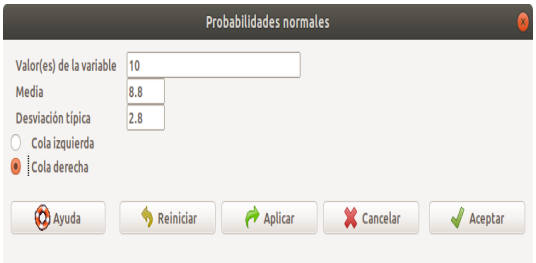


Figure 3.13: Selección de parámetros para el cálculo de $P[X \geq 10]$.

(b) *Probabilidad de que el diámetro de un árbol sea por mayor que 20 pulgadas.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** A continuación, introducimos los datos como en la Figura 3.14 obteniéndose el valor

```
 > pnorm(c(20), mean = 8.8, sd = 2.8, lower.tail = FALSE)
[1] 0.00003167124
```

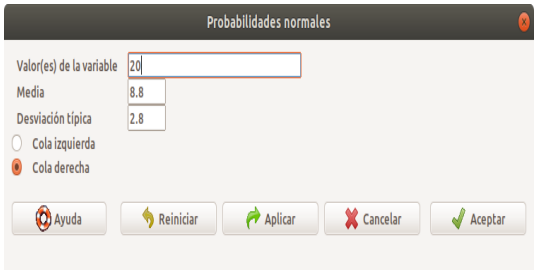


Figure 3.14: Selección de parámetros para el cálculo de $P[X > 20]$.

(c) *Probabilidad de que el diámetro de un árbol esté entre 5 y 10 pulgadas.*
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales** A continuación, introducimos los datos como en la Figura 3.15, y posteriormente calculamos la diferencia entre las dos probabilidades, obteniéndose:

```
 > pnorm(c(10, 5), mean = 8.8, sd = 2.8, lower.tail = TRUE)
[1] 0.66588243 0.08736791

> 0.66588243 - 0.08736791
[1] 0.5785145
```

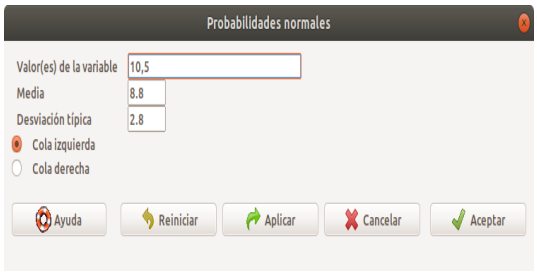


Figure 3.15: Selección de parámetros para el cálculo de $P[5 \leq X \leq 10]$.

(d) *Hallar el valor de c tal que el intervalo* $(8.8-c, 8.8+c)$ *incluya el 98% de los valores de diámetro.*
Si el intervalo $(8.8-c, 8.8+c)$ incluye el 98% de los valores de diámetro siginifica que a la izquierda de $8.8-c$ hay sólo un 1% de dichos valores. Igualmente, hay un 1% de dichos valores a la dercha de $8.8+c$ (o, lo que es lo mismo, un 99% a la izquierda de $8.8+c$).
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Cuantiles normales** A continuación, introducimos los datos como en la Figura 3.16 obteniéndose:

```
 > qnorm(c(0.01, 0.99), mean = 8.8, sd = 2.8, lower.tail = TRUE)
[1]  2.286226 15.313774
```

, es decir, obtenemos el intervalo $(2.286226, 15.313774)$.

Figure 3.16: Selección de parámetros para obtener el valor de c tal que $(8.8-c, 8.8+c)$ contiene el 98% de los valores.

(e) *Se eligen cuatro árboles de forma independiente. Hallar la probabilidad de que por lo menos uno tenga una diámetro mayor de 10 pulgadas.*
Llamemos E al suceso: "*el diámetro del árbol de mayor que 10 pulgadas*".
Llamemos Y a la variable: "*número de árboles entre cuatro cuyo diámetro es mayor que 10 pulgadas*".
Entonces $Y \sim B(n, p)$ con $n = 4$ y $p =$.
Nos piden calcular $P[Y \geq 1]$.
Seleccionamos **Distribuciones** $\rightarrow$ **Distribuciones discretas** $\rightarrow$ **Distribución Binomial** $\rightarrow$ **Probabilidades binomiales acumuladas** A continuación, introducimos los datos (no debemos olvidarnos que en las variables discretas en `RCommander` a la deracha es estrictamente mayor por lo que $P[X \geq 1] = P[X > 0]$ y marcar **Cola derecha**) como en la Figura 3.17 obteniéndose el valor

```
 > pbinom(c(0), size = 4, prob = 0.3341176, lower.tail = FALSE)
[1] 0.803397
```

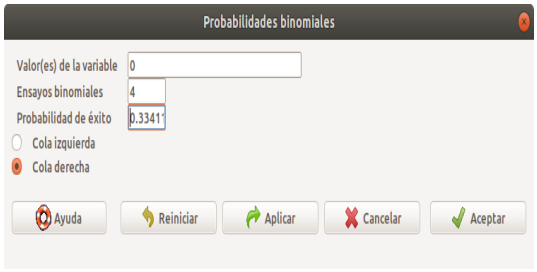


Figure 3.17: Selección de parámetros para calcular la probabilidad de que por lo menos un árbol tenga diámetro mayor que 10 pulgadas.

3.3 Aproximaciones de variables

En la teoría de la *Inferencia Estadística*, algunas cuestiones sobre intervalos de confianza y contraste de hipótesis exigen la normalidad de las poblaciones. Hay estadísticos como la media muestral que se pueden suponer con distribución normal aún no siéndolo la distribución poblacional, por el Teorema del Límite Central. Ello exige que el tamaño muestral sea mayor cuanto menor sea el comportamiento normal de la población. En general se puede considerar que en una muestra la media muestral sigue una distribución normal a partir un tamaño de 25 o 30 datos.

El *Teorema del Límite Central* justifica que se puedan calcular la probabilidad binomial y la de Poisson con aproximaciones mediante la normal.

Proposición 3.3.1 Sea $X \sim B(n,p)$, si se cumple que $np > 5$ y $nq > 5$, entonces $F_X(x) = P[X \leq x]$ se puede aproximar por $F_Y(x) = P[Y \leq x]$, donde $Y \sim N(np, npq)$ (distribución normal de media np y varianza npq).

Proposición 3.3.2 Sea $X \sim \mathscr{P}(\lambda)$ entonces si λ es grande ($\lambda > 20$) se cumple que $F_X(x) = P[X \leq x]$ se puede aproximar por $F_Y(x) = P[Y \leq x]$, donde $Y \sim N(\lambda, \lambda)$ (distribución normal de media λ y varianza λ.)

Aproximación de Binomial.

■ **Ejemplo 3.2** Una pieza es defectuosa con probabilidad 0.06. Hallar la probabilidad de que en una muestra de 100 piezas tomadas al azar, 8 sean defectuosas utilizando la aproximación normal. ■

Solución

Identificamos primeramente la distribución que rige el experimento:

Una pieza es perfecta o defectuosa (éxito o fracaso) con una probabilidad de $p = 0.06$ luego el número de defectuosas en 100 extracciones es una v. a. binomial $B(n = 100, p = 0,06)$.

La aproximación se hará mediante una variable normal de media $n \cdot p = 100 \times 0.06 = 6$ y varianza $n \cdot p \cdot (1-p) = 100 \times 0.06 \times 0.94 = 5.64$. Su desviación típica $\sigma = \sqrt{5.64} = 2.37487$.

Con la $N(6, 2.37487^2)$ tendremos que aproximar $P[X = 8]$, para lo que, aplicando la **corrección de continuidad** en la v.a. Normal, hallaremos $P(7.5 < X < 8.5) =$
$= P(X < 8.5) - Pr(X < 7.5) = 0.8537583 - 0.7361803 = 0.117578$

```
> pnorm(c(8.5,7.5), mean=6,  sd=2.37487, lower.tail=TRUE)
 [1] 0.8537583 0.7361803
```

Después hacemos la diferencia

```
> pnorm(8.5, mean=6, sd=2.3, lower.tail=TRUE) - pnorm(7.5, mean=6, sd=2.3, lower.tail=TRUE)
[1] 0.1186165
```

Calculándolo exactamente con la binomial:

```
> dbinom(8, size=100, prob=0.06)
 [1] 0.1053636
```

Aproximación Poisson.

■ **Ejemplo 3.3** En un proceso de fabricación se sabe que el nº aleatorio de unidades defectuosas producidas diariamente, viene dado por la ley de probabilidad de Poisson de media $\lambda = 10$. Determinar la probabilidad de que en 150 días, el nº de unidades defectuosas producidas supere 1.480 unidades. ■

Solución

Identificación del problema: nos dicen que el nº de piezas defectuosas generadas diariamente sigue una v.a. de Poisson ($\lambda = 10$).

Para un período de 150 días, el número de defectuosas será $P(150 \cdot 10) = P(1500)$.

Para la pregunta $P(X > 1480)$, haremos la corrección de continuidad, calculando $P(X > 1480 + 0,5)$ con la distribución normal.

La aproximación normal para la distribución de Poisson será: $P(\lambda) \to N(\lambda, \lambda)$

Luego, pasando a la aproximación normal tendremos que trabajar con: $N(10 \cdot 150, 10 \cdot 150)$

La desviación típica (necesaria en el comando `R`) es $\sqrt{10 \cdot 150} = 38.7298$.

La llamada a la función de `R`:

```
> pnorm(c(1480.5), mean=1500, sd=sqrt(10*150), lower.tail=FALSE)
 [1] 0.6926893
```

Si se hace sin corrección de continuidad:

```
> pnorm(c(1480), mean=1500, sd=sqrt(10*150), lower.tail=FALSE)
  [1] 0.6972117
```

Operando con la distribución de Poisson exacta:

```
> ppois(c(1480), lambda=1500, lower.tail=FALSE)
 [1] 0.6915581
```

Hay que recordar que para una variable discreta X, la opción `lower.tail=FALSE` calcula la $P[x > k]$, aquí $k = 1480$, es decir, con mayor estricto.

3.4 Distribución de una combinación lineal

Ejercicio 3.4 Sean X_1, X_2 y X_3 los tiempos necesarios para llevar a cabo tres tareas de reparación sucesivas en cierto taller. Suponga que los tiempos son variables aleatorias independientes, normales, con valores esperados μ_1, μ_2 y μ_3 y varianzas σ_1^2, σ_2^2 y σ_3^2, respectivamente.

a. Si $\mu_1 = \mu_2 = \mu_3 = 60$ y $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 15$, calcule $P(X_1 + X_2 + X_3 \leq 200)$ y $P(150 \leq X_1 + X_2 + X_3 \leq 200)$

b. Por medio de las μ_i, y las σ_i del inciso (a), calcule $P(55 \leq \overline{X})$ y $P(58 \leq \overline{X} \leq 62)$.

c. Con las μ_i, y las σ_i del inciso (a), calcule $P(-10 \leq X_1 - 0.5X_2 - 0.5X_3 \leq 5)$.

d. Si $\mu_1 = 40$, $\mu_2 = 50$, $\mu_3 = 60$, $\sigma_1^2 = 10$, $\sigma_2^2 = 12$ y $\sigma_3^2 = 14$, calcule $P(X_1 + X_2 + X_3 \leq 160)$ y $P(X_1 + X_2 \geq 2X_3)$.

■

Solución

(a) La combinación lineal $X_1 + X_2 + X_3$ sigue una distrubución normal de media $\mu = \mu_1 + \mu_2 + \mu_3 = 180$ y varianza $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = 45$.

(a.1) $P[X_1 + X_2 + X_3 \leq 200]$.

Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales acumuladas**. A continuación, introducimos los datos (`RCommander` no acepta el valor *sqrt*45, luego hemos de calularlo y en poner en el campo el valor de $\sqrt{45} = 6.708204$.). Los datos pueden visualizarse en la la Figura 3.18(a). Se obtiene:

```
 > pnorm(c(200), mean = 180, sd = 6.708204, lower.tail = TRUE)
[1] 0.9985654
```

(a) (b)

Figure 3.18: Cálculo de: (a) $P[X_1 + X_2 + X_3 \leq 200]$; (b) $P[150 \leq X_1 + X_2 + X_3 \leq 200]$.

(a.2) $P[150 \leq X_1 + X_2 + X_3 \leq 200]$.

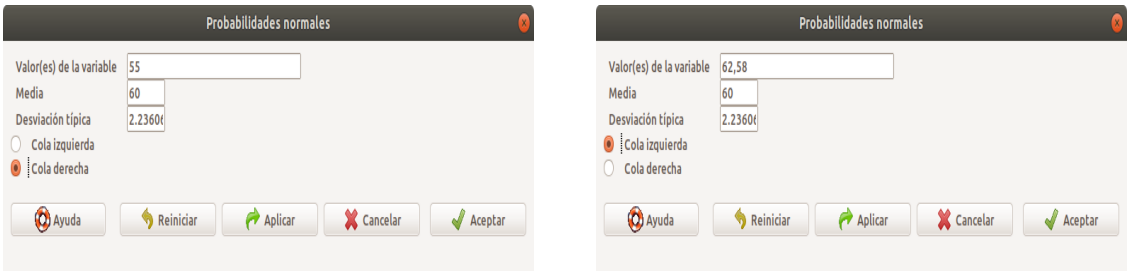
Procedemos de forma análoga, introduciendo los datos como en la Figura 3.18(b), y posteriormente calculamos la diferencia entre los dos valores calculados, obteniéndose:

```
 > pnorm(c(200, 150), mean = 180, sd = 6.708204, lower.tail = TRUE)
[1] 0.998565443463 0.000003872109

> 0.998565443463 - 0.000003872109
[1] 0.9985616
```

(b) La **media muestral** $\overline{X}$ sigue una distrubución normal de media $\mu_{\overline{X}} = \mu_1 = \mu_2 = \mu_3 = 60$ y varianza $\sigma^2_{\overline{X}} = \sigma^2_1/3 = \sigma^2_2/3 = \sigma^2_3/3 = 5$.

(b.1) $P[55 \leq \overline{X}] = P[\overline{X} \geq 55]$.
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales acumuladas**. Introducimos los datos (calculando previamente $\sqrt{5}$) como en la Figura 3.19(a). Se obtiene:

```
 > pnorm(c(55), mean = 60, sd = 2.236068, lower.tail = FALSE)
[1] 0.9873263
```



(a) (b)

Figure 3.19: Cálculo de: (a) $P[55 \leq \overline{X}]$; (b) $P[58 \leq \overline{X} \leq 62]$.

(a.2) $P[58 \leq \overline{X} \leq 62]$.
Procedemos de forma análoga, introduciendo los datos como en la Figura 3.19(b), y posteriormente calculamos la diferencia entre los dos valores calculados, obteniéndose:

```
 > pnorm(c(62, 58), mean = 60, sd = 2.236068, lower.tail = TRUE)
[1] 0.8144533 0.1855467

> 0.8144533 - 0.1855467
[1] 0.6289066
```

(c) $P[-10 \leq X_1 - 0.5X_2 - 0.5X_3 \leq 5]$
La combinación lineal $X_1 - 0.5X_2 - 0.5X_3$ sigue una distrubución normal de media $\mu = \mu_1 - 0.5 \cdot \mu_2 - 0.5 \cdot \mu_3 = 0$ y varianza $\sigma^2 = \sigma^2_1 + (-0.5)^2\sigma^2_2 + (-0.5)^2\sigma^2_3 = 22.5$.
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales acumuladas**. A continuación, introducimos los datos (calculando previamente $\sqrt{22.5}) = 4.743416$) como en la Figura 3.20. Posteriormente calculamos la diferencia entre los dos valores calculados, obteniéndose:

```
 > pnorm(c(5, -10), mean = 0, sd = 4.743416, lower.tail = TRUE)
[1] 0.85407975 0.01750748

> 0.85407975 - 0.01750748
[1] 0.8365723
```

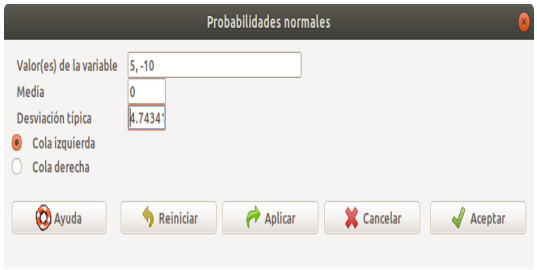


Figure 3.20: Cálculo de $P[-10 \leq X_1 - 0.5X_2 - 0.5X_3 \leq 5]$.

(d) (d.1) $P[X_1+X_2+X_3 \leq 160]$.
La combinación lineal $X_1+X_2+X_3$ sigue una distrubución normal de media $\mu = \mu_1+\mu_2+\mu_3 = 40+50+60 = 150$ y varianza $\sigma^2 = \sigma_1^2+\sigma_2^2+\sigma_3^2 = 36$.
Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Normal** → **Probabilidades normales acumuladas**. A continuación, introducimos los datos como en la Figura 3.21(a). Se obtiene:

```
 > pnorm(c(160), mean = 150, sd = 6, lower.tail = TRUE)
[1] 0.9522096
```

(a) (b)

Figure 3.21: Cálculo de: (a) $P[X_1+X_2+X_3 \leq 160]$; (b) $P[X_1+X_2 \geq 2X_3]$.

(a.2) $P[X_1+X_2 \geq 2X_3]$.
$X_1+X_2 \geq 2X_3 \Leftrightarrow X_1+X_2-2X_3 \geq 0$
La combinación lineal $X_1+X_2-2X_3$ sigue una distrubución normal de media $\mu = \mu_1+\mu_2-2\mu_3 = 40+50-120 = -30$ y varianza $\sigma^2 = \sigma_1^2+\sigma_2^2+(-2)^2\sigma_3^2 = 78$.
Procedemos de forma análoga, introduciendo los datos (calculando previamente $\sqrt{78} = 8.831761$) como en la Figura 3.21(b), obteniéndose:

```
 > pnorm(c(0), mean = -30, sd = 8.831761, lower.tail = FALSE)
[1] 0.0003408552
```

3.5 Distribuciones en el muestreo

En la teoría de muestreo suelen aparecer las variable t de Student, χ^2 y F de Snedecor. Al resolver problemas de intervalos de confianza, hemos de calcular los cuantiles de estas distribuciones. Veamos cómo hacerlo con `R-Commander`.

3.5.1 Variable t de Student

■ **Ejemplo 3.4** Hallar el valor crítico de t para el que el área bajo la cola derecha de la f. de densidad de la variable aleatoria t de Student sea 0.05, para el caso de que la v.a. **t** tenga 16 grados de libertad (g. l.). ■

Solución

Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución t** → **cuantiles t** y marcamos los valores 0.05 la Figura 3.22

El comando `R` correspondiente es

```
> qt(c(0.05), df=16, lower.tail=FALSE)
[1] 1.745884
```

Si el valor buscado de la v.a. t deja a la derecha un área de 0.05, a la izquierda el área será $1-0.05$, con lo que también se podría haber marcado el valor `0.95` y `cola izquierda`.

```
> qt(c(0.95), df=16, lower.tail=TRUE)
[1] 1.745884
```

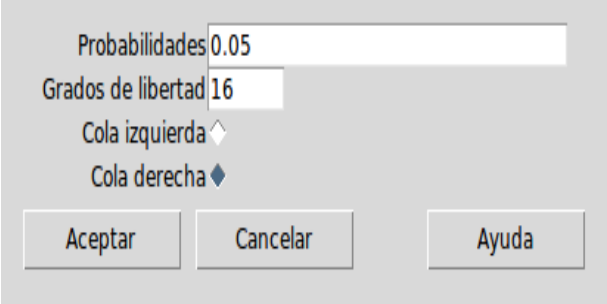


Figure 3.22: Cuantiles de la distribución t con 16 grados de libertad.

3.5.2 Variable Chi^2 (χ^2)

■ **Ejemplo 3.5** Hallar el valor de la v.a. χ^2 con $n = 13$ grados de libertad que deje a su izquierda bajo la función de densidad un área de 0.05. ■

Solución

Seleccionamos **Distribuciones** → **Distribuciones continuas** → **Distribución Chi-cuadrado** → **cuantiles Chi-cuadrado** y marcamos el valor 0.05 y `cola izquierda`. Se obtiene

```
> qchisq(c(0.05), df=13, lower.tail=TRUE)
 [1] 5.891864
```

3.6 Ejercicios propuestos

Ejercicio 3.5 Un agente de seguros vende pólizas a 5 individuos, todos de la misma edad. La probabilidad de que un individuo viva 30 años más es de 3/5. Determinar la probabilidad de que dentro de 30 años vivan:

(a) 4 indivividuos;
(b) como mucho 2;
(c) al menos tres individuos;
(d) más de un individuo y a lo sumo 4.

■

Ejercicio 3.6 La probabilidad de que un paciente se recupere de una cierta enfermedad es 0.4. Si 15 personas contraen la enfermedad, >cuál es la probabilidad de que

(a) al menos 10 sobrevivan?
(b) sobrevivan entre 3 y 8?
(c) sobrevivan exactamente 5 personas?

■

Ejercicio 3.7 Un estudio revela que el 50% de los alumnos que estudian el último curso de bachillerato continúan sus estudios en la universidad. Suponiendo que esta afirmación es cierta, calcular la probabilidad de que entre 18 estudiantes del último año de bachillerato elegidos al azar:

(a) exactamente 10 irán a la universidad;
(b) al menos 10 irán a la universidad;
(c) como mucho 8 irán a la universidad;

■

Ejercicio 3.8 El número de llamadas telefónicas que se reciben por minuto en una centralita se distribuye según una variable aleatoria Poisson de media 1.8. Calcular la probabilidad de que en un minuto se reciban:

(a) Al menos tres llamadas.
(b) Exactamente dos llamadas.

■

Ejercicio 3.9 El número promedio de partículas radiactivas que pasan a través de un contador durante 1 ms en un experimento de laboratorio es 4.

(a) >Cuál es la probabilidad de que entren 6 partículas al contador en un milisegundo determinado?
(b) >Cuál es la probabilidad de que entren 5 o 6 partículas?

■

Ejercicio 3.10 La demanda semanal de un artículo sigue una distribución normal de media 100 y desviación típica 20.

(a) Calcular la probabilidad de que la demanda semanal del artículo esté comprendida entre 80 y 110 unidades.
(b) >De cuántas unidades se debe disponer al principio de la semana para satisfacer la demanda con una probabilidad de 0.99?

■

Ejercicio 3.11 En un proceso fotográfico, el tiempo de revelado de impresiones se puede considerar como una variable aleatoria normal de parámetros $\mu = 15'40$ segundos y $\sigma = 0'48$ segundos. Calcular la probabilidad de que el tiempo que tarda en revelar una de las impresiones sea:

(a) al menos 16 segundos;
(b) como mucho 14.20 segundos;
(c) cualquier valor comprendido entre 15 y 15.80 segundos.

■

Ejercicio 3.12 Representar gráficamente las funciones de densidad y de distribución de una variable aleatoria uniforme continua en el intervalo $[0, 3]$. ■

Ejercicio 3.13 Sabiendo que la demanda de gasolina durante un cierto período de tiempo se comporta con arreglo a la ley normal de media 150000 litros y desviación típica 10000 litros, determinar la cantidad que hay que tener dispuesta a la venta en dicho período para poder satisfacer la demanda con una probabilidad de 0.95. ■

Ejercicio 3.14 El llenado de los paquetes de 1 Kg. de café de una determinada marca se realiza de forma mecánica, siendo el peso real de cada paquete una variable aleatoria con distribución $N(1000, 25)$. Desde el punto de vista comercial, un paquete de café se considera aceptable si su peso es superior a 990 grs.

(a) >Cuál es la probabilidad de que un paquete elegido al azar cumpla las condiciones del mercado?
(b) Los paquetes se embalan en cajas de 100. >Cuál es la probabilidad de que en una caja elegida al azar haya a lo sumo tres paquetes que pesen menos de la cantidad exigida?
(c) El proveedor, para establecer el control sobre el buen funcionamiento de la máquina, elige 25 paquetes al azar y si el peso medio de estos es superior a 1002 grs. detiene el proceso de empaquetado. >Cuál es la probabilidad de que ocurra esto?

■

Ejercicio 3.15 Una empresa eléctrica fabrica focos que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Encuéntrese la probabilidad de que una muestra aleatoria de 16 focos tenga una vida promedio de menos de 775 horas. ■

Ejercicio 3.16 Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura. Se pintan 18 especímenes con el tipo A y en cada uno se registra el tiempo de secado en horas. Lo mismo se hace con el tipo B. Se sabe que las desviaciones estándar de la población son ambas 10.

Supongamos que el tiempo medio de secado es igual para los dos tipos de pintura. Encontrar la probabilidad de que el tiempo promedio de las pinturas de tipo A exceda en más de una hora al tiempo promedio de secado de las pinturas B, suponiendo tamaños muestrales $n_A = n_B = 18$. ■

Ejercicio 3.17 Sabiendo que el 30% de los enfermos con infarto de miocardio que ingresan en el hospital fallecen en el mismo, y que en un año ingresan 2000, determina la probabilidad de que fallezcan en el hospital 550 a lo sumo. ■

Ejercicio 3.18 La probabilidad de que una determinada máquina fabrique una pieza defectuosa es 0.0001. En un año se fabrican 2000 piezas. >Cuál es la probabilidad de que el número de piezas defectuosas producidas en un año sea mayor que 2? ■

4. Interv. de Confianza. Contrastes de hipótesis

4.1 Introducción

Los Intervalos de Confianza y los Contrastes de Hipótesis son dos herramientas estadísticas íntimamente relacionadas, hasta el punto de que R Commander las muestra conjuntamente: es decir, cuando solicitamos la realización de una prueba de hipótesis, R Commander incluye en los resultados el intervalo de confianza correspondiente a la prueba.

El objetivo de esta práctica es obtener e interpretar los resultados de algunas de las pruebas de hipótesis que centran sus afirmaciones en alguno de los parámetros de la población (pruebas paramétricas) como son

1. Realizar contrastes de hipótesis sobre la media de una y dos poblaciones.
2. Realizar contrastes de hipótesis sobre varianzas sobre una y dos poblaciones.
3. Obtener intervalos de confianza mediante R Commander.

Una vez realizados los contrastes, hemos de tomar una decisión. Dicha toma de decisión la haremos, por regla general, por las dos primeras de las siguientes tres opciones:

- **Mediante el p-valor**. Si el **p-valor** es **menor** que el nivel de significación α significa que el valor del estadístico usado pertenece a la región crítica y, en consecuencia, **se rechaza la hipótesis nula**. Si el **p-valor** es **mayor** que el nivel de significación α significa que el valor del estadístico usado pertenece a la región de aceptación y, en consecuencia, **no se rechaza la hipótesis nula**.
- **Mediante el valor del estadístico de contraste**. Una vez calculado el valor del estadístico de contraste (o T–experimental), tendremos que rechazar H_0 si el dicho valor del T–experimental cae en la región de rechazo.
  Advertencia: *Cuando utilicemos un estadístico T (que puede representar una normal, t de Student, chi–cuadrado, etc) la notación T_α significa $P[X \leq T_\alpha] = \alpha$.*
- **Mediante el intervalo de confianza**. Esta tercera forma la utilizaremos opcionalmente como regla práctica para saber si hemos de rechazar o no rechazar la hipótesis nula, fijándonos en el el intervalo de confianza para el estadístico utilizado. Si dicho intervalo contiene al valor de prueba, hemos de aceptar el dicho valor de prueba como cierto y *no rechazar la hipótesis nula*.

A continuación, vamos a realizar algunos contrastes y a calcular intervalos de confianza, de los que hemos estudiado en teoría.

4.2 Contrastes de hipótesis e intervalos de confianza sobre medias

4.2.1 Contraste de hipótesis e intervalo de confianza sobre la media de una población

El contraste sería de la forma

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \quad \text{(o unilateral)} \end{cases}$$

Si la varianza poblacional es desconocida el estadístico $\overline{X}$ se tipifica mediante

$$T = \frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$$

.

Para que el estadístico T se ajuste correctamente al modelo de distribución de probabilidad t de *Student*, es necesario que la población muestreada sea *normal*. No obstante, con tamaños muestrales grandes (a partir de 20 o 30 casos), el ajuste del estadístico T a la distribución t de *Student* es suficiente. En caso de muestras pequeñas, tendríamos que hacer un contraste de normalidad, que no hemos estudiado en teoría.

Veamos un ejemplo:

■ **Ejemplo 4.1** El archivo de datos `ozono.txt` contiene datos relativos a la medición del nivel de ozono en una determinada localidad onubense. Si el valor óptimo para la calidad del aire es 30 >Podremos asegurar, con un nivel de significación del 95%, que el nivel medio de dicha localidad es óptimo? ■

Los datos se encuentran en el fichero `ozono.txt`. Vamos a empezar planteando el problema y posteriormente veremos cómo resolverlo.

Fijémonos que nos piden claramente que confirmemos o desmintamos una afirmación: “el nivel medio de ozono en dicha localidad es óptimo". Esto nos obliga a plantear un contraste de hipótesis. Por lo tanto, si notamos μ al nivel medio de ozono, tenemos que contrastar

$$\begin{cases} H_0 : \mu = 30, \\ H_1 : \mu \neq 30 \end{cases}$$

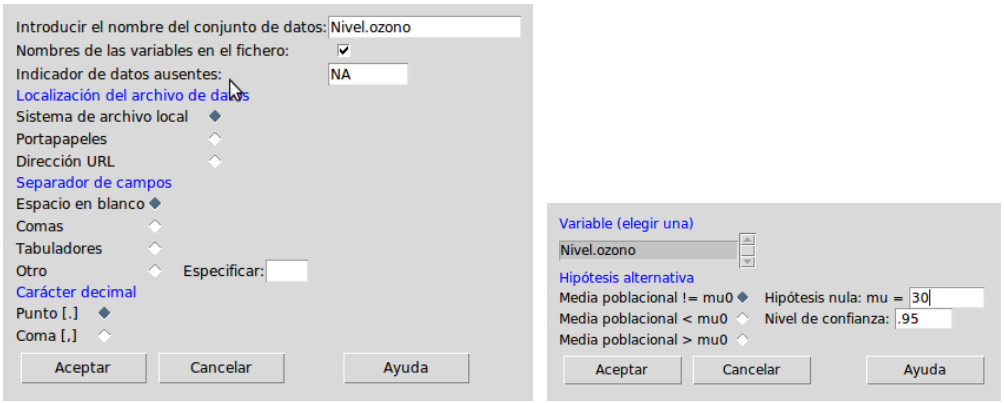
Dado que el tamaño muestral es generoso (191, muy superior a 30), no necesitamos la hipótesis de normalidad de los datos.

Resolución del problema mediante R Commander

El primer paso suele ser desplazarse al directorio de trabajo. Esto lo hacemos mediante el menú ‘‘`Fichero -> Cambiar el directorio de trabajo`’’.

A continuación debemos importar los datos desde el fichero `ozono.txt` para manejar la variable en cuestión, ‘‘`Datos -> Importar datos -> Desde archivo de texto`’’ y completamos las opciones según la Figura 4.1a.

A continuación elegimos la opción del menú ‘‘`Estadísticos -> Medias -> Test t para una muestra`’’. Esta opción abrirá la ventana que aparece en la Figura 4.1b. Fijémonos con detalle en ella:



(a) Importar datos (b) t–Test para una muestra

Figure 4.1: Contraste de hipótesis sobre la media de una población.

1. Hay que especificar la variable cuya media queramos analizar. En nuestro caso `Nivel.ozono`, puede comprobarse solicitando un resumen de variables ‘‘`Estadísticos -> Resúmenes -> Conjunto de datos activo`’’.
2. Nos pide que indiquemos qué contraste deseamos realizar. En nuestro caso hemos elegido la opción de un test bilateral.
3. Hay que introducir el valor hipotético con el que estamos comparando la media, en nuestro caso, 30.
4. Nos pide, por último, que especifiquemos un nivel de confianza. *En realidad este nivel de confianza no lo es para el contraste, que se resolverá a través del p-valor, sino para el intervalo de confianza asociado al problema.*

Análisis de los resultados

El resultado es el siguiente:

```
------------------------------------------------------------
       One Sample t-test

data:  Nivel.ozono$Nivel.ozono

t = 2.5421, df = 190, p-value = 0.01182

alternative hypothesis: true mean is not equal to 30

95 percent confidence interval:

 30.43765 33.46916

sample estimates:

mean of x
  31.9534
------------------------------------------------------------
```

Analicemos el resultado con detalle:

- En primer lugar, nos recuerda la variable que estamos analizando: `Nivel.ozono$Nivel.ozono` .
- A continuación nos informa del valor del estadístico de contraste (`t = 2.5421`), de los grados de libertad (`df = 190`) y del p-valor (`p-value = 0.01182`).
  Ya podemos, por tanto, concluir que dado que el p-valor es inferior al 5% (aprox. 1%), tenemos suficientes evidencias en los datos para rechazar la hipótesis nula ($\mu = 30$) en favor de la alternativa ($\mu \neq 30$), es decir, con los datos de la muestra podemos asegurar al 95% de confianza que el nivel medio de ozono en la localidad no es óptimo, incluso podríamos decir que es un poco alto.
- Nos recuerda cuál era la hipótesis nula que habíamos planteado: `alternative hypothesis: true mean is not equal to 30`.
- A continuación proporciona un intervalo de confianza bilateral, con un nivel de confianza del 95%, para la media de la distribución que se le supone a los datos: `95 percent confidence interval: 30.43765 33.46916`. Lo que quiere decir el resultado es que

$$P[\mu \in (30.43765, 33.46916)] = 0.95.$$

  La relación que guarda el intervalo de confianza con el contraste de hipótesis es la siguiente: fijémonos que el valor hipotético que hemos considerado para la media, 30, no está dentro de este intervalo, luego éste no es un valor de confianza para la μ. Esta es otra forma de concluir que los datos que avalan que la media de la variable es significativamente distinta de 30, ya que éste no es un valor suficientemente plausible para esta media. Si los datos fueran tales que el intervalo de confianza para μ incluyeran al valor 30, no tendríamos suficientes razones para pensar que el valor de μ fuera distinto de 30, pero no es el caso.
- Finalmente, proporciona los estadísticos muestrales utilizados, en este caso, la media muestral: `sample estimates: mean of x 31.9534`.

Conclusiones

Por lo tanto podemos decir que existen evidencias suficientes de que el nivel de ozono en la localidad es distinto y superior a 30.

Quizá podríamos recomendar a los expertos que deseaban contrastar la hipótesis que aumenten el tamaño de la muestra para tener una mayor certeza aunque la muestra ya es suficientemente significativa.

4.2.2 Contraste para la diferencia de medias de poblaciones independientes

Disponemos de dos muestras: una proviene de una distribución $N(\mu_1, \sigma_1^2)$ y otra de una distribución $N(\mu_2, \sigma_2^2)$.

Para el contraste $\begin{cases} H_0 : \mu_1 = \mu_2, \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$ (con σ_1, σ_2 desconocidas) se obtiene la siguiente región crítica $R = \{|T| > t_{m,1-\frac{\alpha}{2}}\} = \left\{ \left| \frac{\overline{X} - \overline{Y}}{S^*} \right| \right\}$.

Dependiendo de si las desviaciones típicas son iguales o diferentes se tienen diferentes valores para m y S^* como se ha visto en teoría. Para distinguir entre ambos casos se hace un contraste previo sobre la igualdad de varianzas (se verá más adelante en esta práctica). Dependiendo que de este test se asuma o no igualdad de varianzas, se obtiene un estadístico u otro para la diferencia de medias.

■ **Ejemplo 4.2** Dividimos aleatoriamente un conjunto de pacientes afectados por cierto virus en dos grupos a los que se administra respectivamente placebo y un tratamiento en estudio para combatir dicha afección vírica. Después de tratar a los pacientes durante 2 meses se mide la concentración de virus de cada uno de ellos. Parte de los resultados del fichero `pacientes.txt` se muestran en la siguiente tabla:

Sujeto	*Administración*	*Nivel*	*Sujeto*	*Administración*	*Nivel*
1	*placebo*	23.00	2	*placebo*	25.00
3	*placebo*	26.00	4	*placebo*	24.00
5	*placebo*	*NA*	6	*placebo*	*NA*
7	*placebo*	22.00	8	*placebo*	25.00
9	*placebo*	27.00	10	*placebo*	25.00
...	...	...	...	...	...
111	*tratamiento*	22.00	112	*tratamiento*	23.00
113	*tratamiento*	22.00	114	*tratamiento*	*NA*
115	*tratamiento*	*NA*	116	*tratamiento*	20.00
117	*tratamiento*	*NA*	118	*tratamiento*	23.00
119	*tratamiento*	22.00	120	*tratamiento*	23.00
...	...	...	...	...	...

>Puede decirse, con un 95% de confianza, que el tratamiento reduce la infección? ■

Los datos se encuentran en el fichero `pacientes.txt`. Comenzaremos planteando el problema para resolverlo posteriormente.

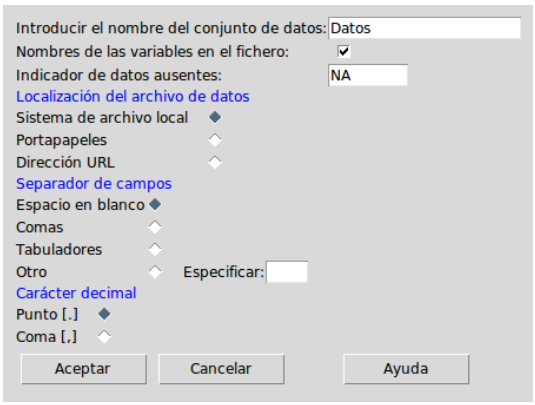
Al igual que en el caso anterior nos piden que contrastemos una afirmación: “el tratamiento reduce la infección". Esto, de nuevo, nos obliga a plantear el contraste de hipótesis $\begin{cases} H_0 : \mu_p = \mu_t, \\ H_1 : \mu_p > \mu_t \end{cases}$ al 5% de significación. Donde μ_p es la media de la concentración en pacientes a los que se les suministró placebo y μ_t es la media de la concentración en pacientes a los que se les administró el tratamiento real.

- En primer lugar, podemos suponer que las muestras son independientes. Nada hace pensar que los datos de una muestra hayan tenido nada que ver con la otra muestra al haber sido distribuidos los pacientes aleatoriamente entre los dos grupos.
- Con respecto al tamaño muestral, podría preocuparnos la hipótesis de normalidad: recordemos que si el tamaño de la muestras es pequeño, éstas deberían proceder de una distribución normal, pero no es el caso: ambas muestras tienen tamaños superiores a 30.
- Finalmente, deberemos plantearnos si podemos suponer o no que las varianzas son iguales. Aún no sabemos realizar contraste de varianzas, cuando sepamos habría que hacerlo antes del contraste de medias. *Para este ejemplo supondremos que son distintas*.

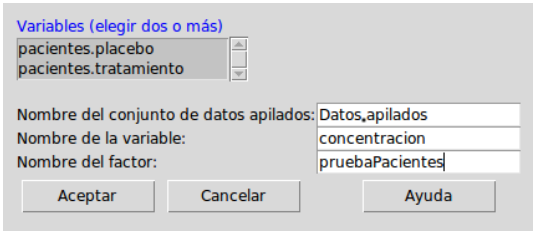
Resolución del problema mediante R Commander

El primer paso suele ser desplazarse al directorio de trabajo. Esto lo hacemos mediante el menú ‘`Fichero -> Cambiar el directorio de trabajo`’.

Después para importar datos que se encuentran en un fichero de tipo texto, tenemos que ver cómo están almacenados (si lo abrimos, por ejemplo, con un editor, vemos que están separados por tabulaciones y que los nombres de las variables están en la primera fila). A continuación debemos importar los datos desde el fichero `pacientes.txt` para manejar la variable en cuestión, ‘`Datos -> Importar datos -> Desde archivo de texto`’ y completamos las opciones según la Figura 4.2a.



(a) Importar datos



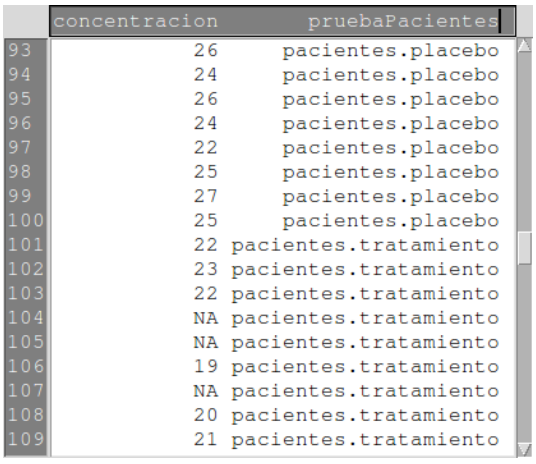
(b) Apilar datos

Figure 4.2: Contraste de hipótesis para la media de poblaciones independientes (1 de 2).

Fijémonos que los datos de las dos muestras aparecen en dos columnas paralelas. Esa es una forma no demasiado correcta de especificarlas, ya que parece que cada dato de una de las muestras está relacionado con otro dato de la otra muestra y, en realidad, las muestras son independientes (de hecho al haber datos perdidos tiene distinto tamaño muestral real).

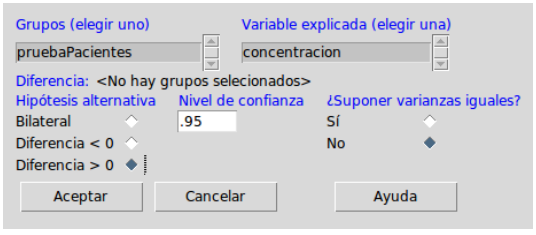
Por este motivo, tenemos que preparar los datos para que R Commander entienda que se trata de dos muestras independientes. Lo que tenemos que hacer es juntar o apilar las dos muestras en una sola columna, indicando en una segunda columna si el dato es de una muestra u otra.

Esta operación se realiza mediante la opción ‘`Datos -> Conjunto de datos activo -> Apilar variables del conjunto de datos activo`’. Lo que tenemos que indicar en esta ventana Figura 4.2b es qué variables son las que queremos apilar, el nombre del nuevo conjunto de datos, el nombre de la nueva variable y el nombre del factor que separa los datos de las dos muestras. En la parte de la derecha de la Figura 4.3a aparecen los datos tal y como quedan tras apilarlos.



	concentracion	pruebaPacientes
93	26	pacientes.placebo
94	24	pacientes.placebo
95	26	pacientes.placebo
96	24	pacientes.placebo
97	22	pacientes.placebo
98	25	pacientes.placebo
99	27	pacientes.placebo
100	25	pacientes.placebo
101	22	pacientes.tratamiento
102	23	pacientes.tratamiento
103	22	pacientes.tratamiento
104	NA	pacientes.tratamiento
105	NA	pacientes.tratamiento
106	19	pacientes.tratamiento
107	NA	pacientes.tratamiento
108	20	pacientes.tratamiento
109	21	pacientes.tratamiento

(a) Datos apilados



(b) t–Test para muestras independientes

Figure 4.3: Contraste de hipótesis para la media de poblaciones independientes (2 de 2).

Observemos que, en efecto, ha creado una primera columna con todos los datos y una segunda columna con un factor que especifica cuáles de ellos son del placebo y cuáles del tratamiento.

Hay un último aspecto muy importante: los datos son ordenados según el factor, y éste ordena sus categorías por orden alfabético. Eso en nuestro ejemplo hace que aparezcan primero los datos según el placebo y después según el tratamiento.

Ahora ya tenemos los datos preparados para ser analizados. Elegimos la opción ‘‘`Estadísticos -> Medias -> Test t para muestras independientes`’’. Esta opción abre la ventana de entradas que aparece en la Figura 4.3b. En ella podéis ver que tenemos que especificar el factor que separa las dos muestras, la variable que estamos analizando, la hipótesis alternativa, el nivel de confianza requerido y si podemos suponer varianzas iguales. En nuestro caso:

- El factor es el único que aparece, que hemos llamado `pruebaPacientes`.
- La variable es la única numérica del conjunto de datos, que hemos llamado `concentracion`.
- Es un test unilateral: >cómo sabemos cuál es la muestra 1 y cuál la muestra 2? Lo sabemos porque el factor se ordena alfabéticamente, luego la concentración con placebo es la muestra 1 y con el tratamiento la muestra 2. Como queremos contrastar $H_1 : \mu_1 > \mu_2$, señalamos `Diferencia > 0`.
- El nivel de confianza exigido es del 95%, que va a ser el que consideremos para el intervalo de confianza asociado.
- No tenemos (aún) razones que avalen que las varianzas deban ser consideradas iguales.

Análisis de los resultados

El resultado es el siguiente:

```
------------------------------------------------------------------------
   Welch Two Sample t-test

data:  concentracion by pruebaPacientes

t = 13.2892, df = 189.532, p-value < 2.2e-16

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

 2.414698      Inf

sample estimates:

mean in group pacientes.placebo   mean in group pacientes.tratamiento
              24.69388                             21.93617
------------------------------------------------------------------------
```

Analicemos el resultado con detalle:

- Especifica que se trata de un test t para la variable `concentracion` separada por el factor `pruebaPacientes`.
- Proporciona el valor del estadístico de contraste (`t = 13.2892`), los grados de libertad (`df = 189.532`), y el p-valor (`p-value < 2.2e-16`). Dado que es inferior a 0.05, ya podemos concluir que tenemos evidencias en los datos para afirmar con un 95% de confianza que la media de la concentración en pacientes tratados con placebo es superior a la de los pacientes a los que se les ha suministrado el tratamiento correcto.
- Explicita cuál es nuestra hipótesis alternativa.
- Proporciona un intervalo de confianza para la diferencia de las medias. En este caso, la probabilidad de que el intervalo $(2.414698, +\infty)$ contenga a la diferencia de las medias es del 95%. El cero no es, por tanto, un valor bastante plausible y por ello hemos rechazado la hipótesis nula en favor de la alternativa.
- Proporciona las dos medias muestrales.

Conclusiones

Podemos por tanto afirmar, con una confianza del 95% que el tratamiento mejora a los pacientes y reduce la concentración de virus.

4.2.3 Contraste para la diferencia de medias para muestras relacionadas

Este problema es básicamente como el anterior, la diferencia es que se tienen dos distribuciones normales X e Y relacionadas entre sí.

Construimos una variable $D = X - Y$ que sigue una distribución $N(\mu_D, \sigma_D^2)$.

El contraste a efectuar es

$$\begin{cases} H_0 : \mu_D = 0, \\ H_1 : \mu_D \neq 0 \end{cases}$$

con la siguiente región crítica $R = \left\{ |T| > t_{n-1,1-\frac{\alpha}{2}} \right\} = \left\{ \left| \frac{\overline{X} - \overline{Y}}{S_{c_d}/\sqrt{n}} \right| > t_{n-1,1-\frac{\alpha}{2}} \right\}$.

Veamos un ejemplo.

Ejercicio 4.1 Se quiere comprobar la rapidez de dos programas informáticos A y B para la resolución de cierta clase de problemas de ingeniería. Para ello se analizan 9 casos utilizando tanto el programa A como el B. Los tiempos obtenidos para resolver estos casos fueron (cargar fichero `programas.txt`):

Caso	1	2	3	4	5	6	7	8	9
Programa A	11.5	13.2	15.7	9.8	12.6	10.5	11.3	12.6	14.1
Programa B	12.3	12.9	13.1	10.9	11.2	12.1	9.9	11.8	12.3

En este caso, se tienen dos distribuciones normales relacionadas ya que se miden dos variables (cada programa de ordenador) sobre una misma muestra de individuos (los problemas de ingeniería). Al no especificarse, tomamos una significación del 5%.

- En primer lugar, las muestras no son independientes ya que son dos programas distintos ejecutados sobre el mismo conjunto de problemas.
- Con respecto al tamaño muestral es muy pequeño pero tenemos la hipótesis de normalidad.
- Finalmente, deberemos plantearnos si podemos suponer o no que las varianzas son iguales. Aún no sabemos realizar contraste de varianzas, cuando sepamos habría que hacerlo antes del contraste de medias. *Para variar, este ejemplo, supondremos que son iguales (aunque no haya ninguna razón objetiva para ello).*

Resolución del problema mediante R Commander

En primer lugar debemos cargar los datos. Debemos realizar un contraste de igualdad de medias en poblaciones relacionadas. En este caso el conjunto de datos tiene exactamente la forma que R Commander necesita para ello, ya que tenemos dos variables, correspondientes a los tiempos de ambos programas. Elegimos, por tanto, la opción ‘‘`Estadísticos -> Medias -> Test t para datos relacionados`’’, lo cual abrirá una ventana de entradas como la de la Figura 4.4. En ella ya se han marcado las opciones requeridas por nuestro análisis. El resultado es el siguiente:

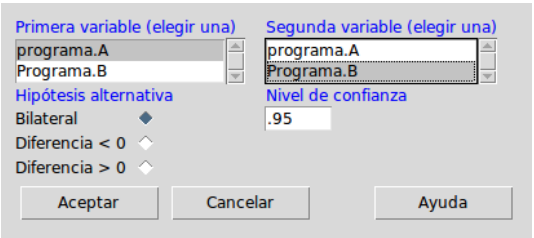


Figure 4.4: Contraste de hipótesis para la media de poblaciones dependientes

Análisis de los resultados

El resultado es el siguiente:

```
------------------------------------------------------------------------
Paired t-test

data:  Datos$programa.A and Datos$Programa.B

t = 1.1141, df = 8, p-value = 0.2976

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -0.5705821  1.6372488

sample estimates:

mean of the differences
              0.5333333
------------------------------------------------------------------------
```

Analicemos el resultado con detalle:

- En las dos primeras líneas se nos informa que estamos realizando un test t para muestras relacionadas sobre los datos relativos a las variables `Programa.A` y `Programa.A` del conjunto de datos `Datos`. El test lo podemos realizar gracias a que la diferencia media entre las dos variables normales sigue una distribución normal.
- En la siguiente línea aparece el valor del estadístico de contraste (`t = 1.1141`), de los grados de libertad (`df = 8`) y el p-valor (`p-value = 0.2976`). Visto el valor de este último podemos concluir que no existen evidencias en los datos de la muestra de que el la diferencia de tiempos entre programas sean significativas. Observad que el p-valor es muy alto, luego las diferencias detectadas son mínimas, en absoluto significativas.
- A continuación aparece el tipo de hipótesis alternativa que hemos elegido.
- Posteriormente aparece un intervalo de confianza al 95% para la diferencia de las medias. El hecho de que el cero sea un valor posible de dicho intervalo es otra forma de ver que no podemos rechazar la hipótesis nula en favor de la alternativa.
- Finalmente aparece el valor muestral de la diferencia de las medias.

Conclusiones

En resumen, los datos no muestran el más mínimo indicio de que ninguno de los programas tenga un comportamiento mejor que el otro.

4.3 Contrastes de hipótesis e intervalos de confianza sobre varianzas

4.3.1 Contraste de hipótesis e intervalo de confianza para el cociente de varianzas de dos poblaciones.

Para contrastar la igualdad de varianzas, tendríamos que realizar el siguiente contraste

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2, \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases},$$

aunque en realidad el que se hace es

$$\begin{cases} H_0 : \sigma_1^2/\sigma_2^2 = 1, \\ H_1 : \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$$

Con la siguiente región crítica

$$R = \{F \notin (f_{n_1-1,n_2-1,\frac{\alpha}{2}}, f_{n_1-1,n_2-1,1-\frac{\alpha}{2}})\}$$

siendo $F = \dfrac{S_1^2}{S_2^2}$ que sigue una distribución F de Fisher–Snedecor con $n_1 - 1$ y $n_2 - 1$ grados de libertad. (n_1 y n_2 son los tamaños muestrales correspondientes y S_1^2 y S_2^2 las respectivas varianzas muestrales).

Veamos un ejemplo:

■ **Ejemplo 4.3** En una finca ganadera se distribuyen cargas de pienso por el campo a razón de 24 kilos de media, como ocurre en cualquier proceso industrial, el número de kilos no es exactamente de 24, sino que se producen variaciones aleatorias en el peso real. El capataz responsable de la explotación es consciente del problema que supone esta variabilidad. Para intentar reducirla, la empresa se plantea cambiar la máquina distribuidora por una nueva, pero previamente quiere probarla para garantizarse que la inversión merece la pena. Para ello calibra la máquina y recopila los datos de 50 cargas. También realiza un muestreo de otras 50 cargas con la vieja máquina, en las mismas condiciones que con la nueva máquina. A la luz de esos datos (ver fichero `granja.txt`), >puede concluir el capataz que la nueva máquina disminuye la variabilidad del peso de las cargas?

Utilícese un nivel de significación del 2.5% y supóngase que la distribución de ambas muestras es normal.

■

Lo que vamos a plantear para resolver el problema es un contraste de igualdad de varianzas (o de desviaciones típicas). Si denotamos por σ_V la desviación típica de las cargas de la máquina vieja y σ_N a la desviación típica de las cargas de la máquina nueva, se trata de contrastar

$$\begin{cases} H_0 : \sigma_V^2 = \sigma_N^2, \\ H_1 : \sigma_V^2 > \sigma_N^2 \end{cases}$$

- De nuevo podemos a asumir desde el principio que los datos proceden de distribuciones normales.
- Utilizaremos un nivel de significación del 2.5%.

Resolución del problema mediante R Commander

Como siempre, el primer paso suele ser desplazarse al directorio de trabajo y después importar los datos que se encuentran en un fichero de tipo texto denominado `granja.txt` y completamos las opciones como en los casos anteriores.

Los datos de las dos muestras aparecen en dos columnas paralelas pero realmente las muestras son independientes. Por este motivo, tenemos que preparar los datos para que R Commander entienda que se trata de dos muestras independientes. Lo que tenemos que hacer es juntar o apilar las dos muestras en una sola columna, indicando en una segunda columna si el dato es de una muestra u otra.

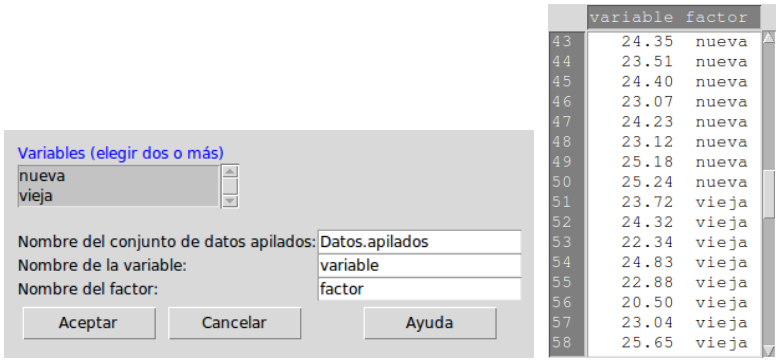
Esta operación se realiza mediante la opción ‘‘`Datos -> Conjunto de datos activo -> Apilar variables del conjunto de datos activo`’’. Lo que tenemos que indicar en esta ventana Figura 4.5a es qué variables son las que queremos apilar, el nombre del nuevo conjunto de datos, el nombre de la nueva variable y el nombre del factor que separa los datos de las dos muestras. En la Figura 4.5b podemos ver cómo quedan los datos una vez apilados.

Observemos que, en efecto, ha creado una primera columna con todos los datos y una segunda columna con un factor que especifica cuáles de ellos son de la nueva máquina y cuáles de la vieja.

Como antes un último aspecto muy importante: los datos son ordenados según el factor, y éste ordena sus categorías por orden alfabético. Eso en nuestro ejemplo hace que aparezcan primero los datos de la nueva y después de la vieja.

Ahora ya tenemos los datos preparados para ser analizados. Elegimos la opción ‘‘`Estadísticos -> Varianzas -> Test F para dos varianzas`’’. Esta opción abre la ventana de entradas que aparece en la Figura 4.6. En ella tenemos que especificar el factor que separa las dos muestras, la variable que estamos analizando, la hipótesis alternativa, el nivel de confianza requerido. En nuestro caso:

- El factor es el único que aparece, que hemos llamado `factor`.
- La variable es la única numérica del conjunto de datos, que hemos llamado `variable`.
- Es un test unilateral donde, al ordenarse alfabéticamente, la primera corresponde a la máquina nueva y la segunda a la vieja. Como queremos contrastar $H_1 : \mu_N < \mu_V$, señalamos `Diferencia < 0`.



(a) Apilar datos (b) Datos apilados

Figure 4.5: Contraste de hipótesis de dos varianzas (1 de 2).

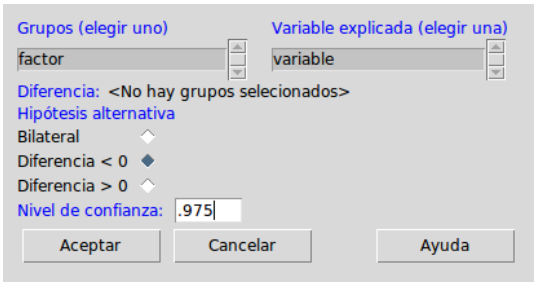


Figure 4.6: Contraste de hipótesis para dos varianzas (2 de 2)

- El nivel de confianza exigido es del 97.5%, que va a ser el que consideremos para el intervalo de confianza asociado.

→ **Importar Datos**

Análisis de los resultados

El resultado es el siguiente:

```
------------------------------------------------------------------------------
F test to compare two variances

data:  variable by factor

F = 0.8969, num df = 49, denom df = 49, p-value = 0.3525

alternative hypothesis: true ratio of variances is less than 1

97.5 percent confidence interval:

0.000000 1.580564

sample estimates:

ratio of variances

         0.8969326
------------------------------------------------------------------------------
```

Analicemos el resultado con detalle:

- En las dos primeras líneas se nos informa que estamos realizando un test F para la variable `variable` y y el factor `factor`. El test lo podemos realizar gracias a que ambas variables siguen una distribución normal.
- En la siguiente línea aparece el valor del estadístico de contraste (`F = 0.8969`), de los grados de libertad del numerados y del denominador (num `df = 49`, denom `df = 49`) y el p-valor (`p-value = 0.3525`). Ese `p-value = 0.3525` indica que no hay suficientes evidencias en los datos para concluir que la nueva máquina disminuya significativamente la varianza en el reparto de la carga. Observad que el p-valor es muy alto, luego las diferencias detectadas son mínimas, en absoluto significativas.
- El cociente de varianzas `ratio of variances = 0.8969326` indican que, realizándole la raiz cuadrada que el cociente de desviaciones típicas es `ratio of variances = 0.947065256`, esto es, aún siendo menor la de la nueva máquina varían tan sólo en aproximadamente un 5%.

Conclusiones

En resumen los datos no muestran suficiente indicio de que la nueva máquina tenga un comportamiento mejor que el otro.

4.4 Contraste de hipótesis e intervalo de confianza para la varianza de una población

El contraste sería de la forma

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2, \\ H_1 : \sigma^2 \neq \sigma_0^2 \quad \text{(o unilateral)} \end{cases}$$

Admitiendo la hipótesis de normalidad en la población de partida, se puede calcular el intervalo de confianza para la varianza. De las fórmulas de teoría dicho intervalo es:

$$\left(\frac{(n-1)*s^2}{\chi^2(n-1,1-\alpha/2)}, \frac{(n-1)*s^2}{\chi^2(n-1,\alpha/2)} \right),$$

Veamos un ejemplo:

■ **Ejemplo 4.4** El archivo de datos `ozono.txt` contiene datos relativos a la medición del nivel de ozono en una determinada localidad onubense. Podemos considerar que la contaminación por ozono se mantiene estable si la desviación típica de las muestras es menor de 9 unidades. A la vista de las muestras >Podremos asegurar, con un nivel de significación del 95%, que el nivel medio se mantiene estable? ■

Los datos se encuentran en el fichero `ozono.txt`. Vamos a plantear el problema y después veremos cómo resolverlo.

Hemos de confirmar o desmentir una afirmación: "*la contaminación por ozono se mantiene estable*". Esto nos obliga a plantear un contraste de hipótesis. Denotando por σ^2 la varianza del nivel de ozono, hemos de contrastar:

$$\begin{cases} H_0 : \sigma^2 = 9^2, \\ H_1 : \sigma^2 > 9^2 \end{cases}$$

suponiendo normalidad en los datos.

Resolución del problema mediante R Commander

Carguemos los datos y veamos cómo resolverlo. Importamos los datos:

Datos → Importar Datos → Desde archivo de texto

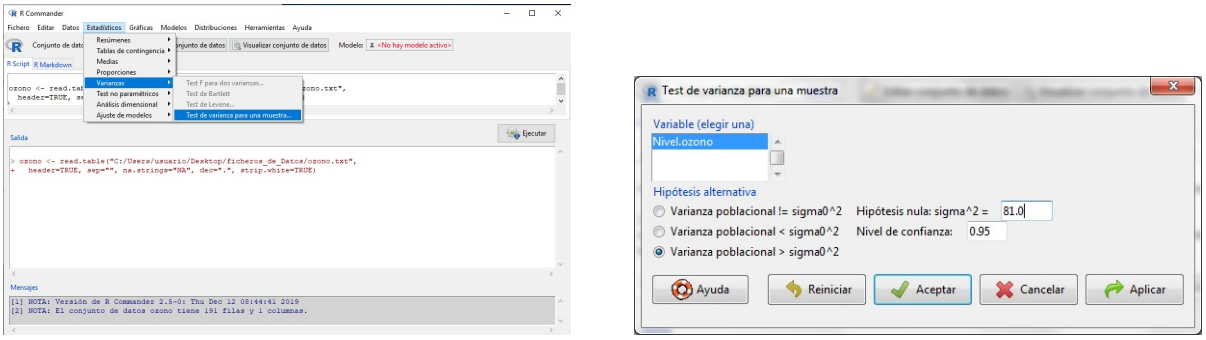
y completamos las opicones como en la Figura 4.1a.

A continuación elegimos la opción

Estadísticos → Varianzas → Test de varianza para una muestra

, tal como se muestra en la Figura 4.7(a). Esta opción abre la ventana que aparece en la Figura 4.7(b).

1. Hay que especificar la variable cuya varianza queremos analizar (en nuestro caso `Nivel.ozono`).
2. Hemos de indicar el tipo de contraste (en nuestro caso unilateral a la derecha).
3. Hay que introducir el valor de la hipótesis nula (en nuestro caso $\sigma_0^2 = 81$).
4. Por último, hemos de especificar un nivel de confianza. *En realizad este nivel de confianza no lo es para el contraste, que se resolverá a través del p–valor, sino para el intervalo de confianza asociado al problema*.



(a) (b)

Figure 4.7: Contraste de hipótesis para la varianza de una muestra.

Análisis de resultados

El resultado es el siguiente:

```
  One sample Chi-squared test for variance

data:  Nivel.ozono[!is.na(Nivel.ozono)]
X-squared = 264.55, df = 190, p-value = 0.0002826
alternative hypothesis: true variance is greater than 81
95 percent confidence interval:
 96.02425      Inf
sample estimates:
var of Nivel.ozono[!is.na(Nivel.ozono)]
                              112.7831
```

Nota 4.4.1 En aquellas versiones de `RCommander` que no dispongan en su menú de la opción `Test de varianza para una muestra`, optaremos por utilizar la función **sigma.test** del paquete **TeachingDemos**, como sigue:

```
library(TeachingDemos)
sigma.test(ozono$Nivel.ozono, alternative='greater', sigma=9.0, conf.level=0.95)
```

cuyo resultado es:

```
One sample Chi-squared test for variance

data:  ozono$Nivel.ozono
X-squared = 264.55, df = 190, p-value = 0.0002826
alternative hypothesis: true variance is greater than 81
95 percent confidence interval:
 96.02425      Inf
sample estimates:
var of ozono$Nivel.ozono
                112.7831
```

es decir, el mismo que obtuvimos anteriormente.

Analizemos los resultados:

1. Nos dice el test que estamos ejecutando "Test Chi-cuadrado para la varianza".
2. A continuación nos recuerda la variable que estamos analizando: Datos$Nivel.ozono.
3. Seguidamente nos informa del valor del estadístico de contraste (`X-squared = 264.553`), de los grados de libertad (`df = 190`) y del p-valor (`p-value = 0.0002826`).

4. Ya podemos, por tanto, concluir que dado que el p-valor es inferior al 5% (aprox. 0.05%), tenemos suficientes evidencias en los datos para rechazar la hipótesis nula ($\sigma^2 = 81$) en favor de la alternativa ($\sigma^2 > 81$), es decir, con los datos de la muestra podemos asegurar al 95% de confianza que la varianza del nivel de ozono en la localidad supera el nivel de estabilidad.
5. A continuación proporciona un intervalo de confianza unilateral, con un nivel de confianza del 95%, para la varianza de la distribución que se le supone a los datos: `95 percent confidence interval: [96.02425 Inf) (es decir,` $[96.02425, +\infty)$). Lo que quiere decir el resultado es que $P[\sigma^2 \in [96.02425, +\infty)] = 0.95$.
   La relación que guarda el intervalo de confianza con el contraste de hipótesis es la siguiente: fijémonos que el valor hipotético que hemos considerado para la varianza, 81, no está dentro de este intervalo, luego éste no es un valor de confianza para σ^2. Esta es otra forma de concluir que los datos que avalan que la varianza de la variable es significativamente distinta de 81, ya que éste no es un valor suficientemente plausible para esta variable. Si los datos fueran tales que el intervalo de confianza para la varianza incluyeran al valor 81, no tendríamos suficientes razones para pensar que el valor de σ^2 estiviese fuera de los valores de estabilidad, pero no es el caso.
6. Finalmente, proporciona los estadísticos muestrales utilizados, en este caso, la varianza: `sample estimates: var of Datos$Nivel.ozono 112.7831`.

4.5 Ejercicios Propuestos

Ejercicio 4.2 Los ingresos semanales de un grupo de 9 personas seleccionadas al azar entre un gran número de individuos, han resultado ser

$$1.570,\ 1.550,\ 1.530,\ 1.520,\ 1.560,\ 1.500,\ 1.510,\ 1.540,\ 1.580$$

> Debe rechazarse la hipótesis de que la muestra procede de una población cuyos ingresos semanales medios son de 1.557 pts. al nivel de significación del 5%. ■

Ejercicio 4.3 De los datos experimentales se deduce que la duración de una pila sigue una distribución normal. El fabricante asegura que el tiempo medio de vida es de 15 horas de duración. Temiendo que la duración pueda ser menor se toma una muestra de 10 pilas y los datos que se obtuvieron fueron los siguientes:

$$14 \quad 14'5 \quad 15 \quad 15'5 \quad 13 \quad 16'5 \quad 10 \quad 15'5 \quad 14 \quad 14$$

Decidir si en base a los datos muestrales debemos rechazar o no la afirmación del comerciante con un nivel de significación del 0'05. ■

Ejercicio 4.4 Durante los últimos 5 años el número de ordenadores que vende por semana cierta empresa informática es aproximadamente normal $N(\mu, \sigma^2)$. En una muestra aleatoria simple de 10 semanas de los últimos 5 años, dicha empresa vendió: 175, 168, 171, 169, 183, 165, 188, 177, 167 y 180 ordenadores. Contrastar si el valor medio de las ventas es 200 con un nivel de significación $\alpha = 0'1$. ■

Ejercicio 4.5 Contrastar si el peso medio de un cierto tipo de piezas mecánicas difiere según sean fabricados en dos lugares diferentes A y B con un nivel de significación igual $\alpha = 0'01$. Los datos obtenidos son los siguientes:

A	80	97	85	73	92	97	100	94
B	99	92	85	79	91	96	105	89

■

Ejercicio 4.6 Estudiar si un cierto tipo de componente mecánico es simétrico o no en base a la medida de sus lados izquierdo y derecho con un nivel de significación $\alpha = 0'05$. Los datos que se han obtenido en las medidas son:

Pieza	1	2	3	4	5
Izquierdo	11.5	13.2	15.7	9.8	12.6
Derecho	12.3	12.9	13.1	10.9	11.2

■

5. Regresión Lineal

5.1 Introducción

La regresión lineal es probablemente una de las técnicas estadísticas más utilizadas y útiles para resolver problemas. Los modelos de regresión lineal son extremadamente poderosos, y tienen la capacidad de extraer complicadas relaciones entre variables. Se suele utilizar tanto para Descripción, para Control como para Predicción.

En términos generales la técnica es útil, entre otras aplicaciones, para ayudar a explicar las observaciones de una variable dependiente, por lo general denotada por y (`variable respuesta o dependiente`), con los valores observados de una o más variables, por lo general denotada por $x_1, x_2, \ldots$ (`variables explicativas o independientes`).

El objetivo de esta práctica es realizar un análisis de regresión básico sobre un conjunto de variables. Para ello trabajaremos de forma secuencial los siguientes aspectos.

1. Contrastar si la relación entre pares de variables es estadísticamente significativa.
2. Ajustar la recta de regresión de una variable dependiente dada una variable independiente.
3. Realizar predicciones y estimaciones a partir de la recta de regresión.
4. Obtener intervalos de confianza para dichas predicciones y estimaciones.

5.2 Correlación

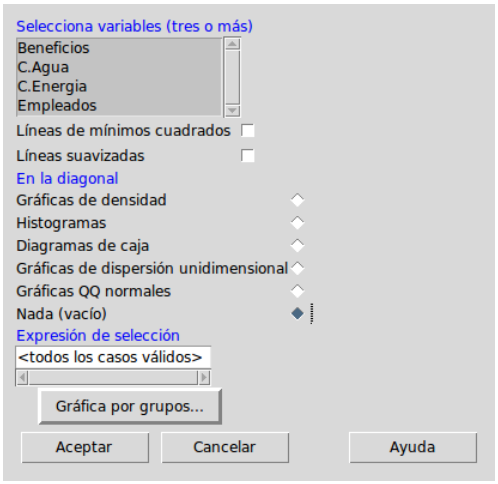
A lo largo de esta sección vamos a utilizar un fichero `empresas.txt` en el que se ha almacenado los beneficios, el gasto en electricidad, en agua y el número de empleados de una serie de empresas andaluzas.

Antes de empezar con el estudio de la correlación es conveniente explorar la relación entre las variables de forma gráfica. Para ello mostraremos el diagrama de dispersión o nube de puntos para obtener una primera impresión del tipo de relaciones que mantienen las variables. En R Commander esto podemos hacerlo mediante la opción "`Gráficas -> Matriz de diagramas de dispersión`".

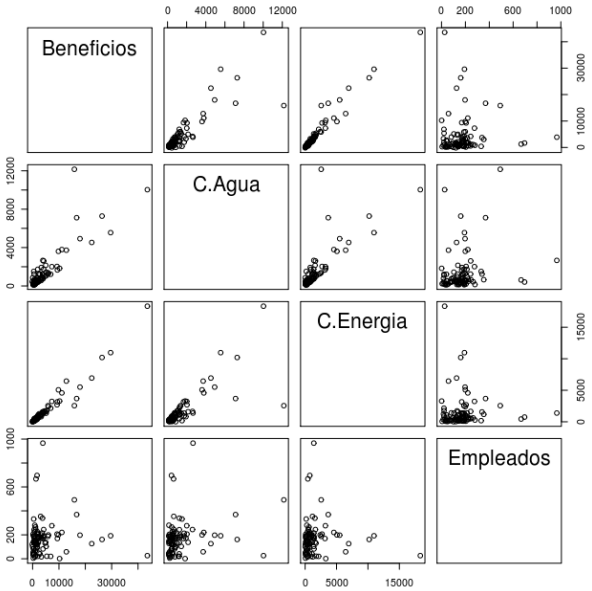
Este análisis es bastante rico, pero nosotros nos conformaremos con las opciones más simples (ver Figura 5.1a). El resultado aparece en la Figura 5.1b.

Como vemos, aunque en general se observa una tendencia positiva, es muy evidente la relación entre los beneficios, el consumo de agua y el consumo de energía y al mismo tiempo es francamente deficiente entre el número de empleados y cualquiera de las otras tres variables.

Una vez observadas gráficamente las posibles relaciones lineales comenzamos a preguntarnos por la fortaleza real de las mismas entre las distintas variables y en particular entre las de consumo de agua,



(a) Entradas



(b) Diagramas

Figure 5.1: Entradas y resultado diagramas de dispersión.

consumo de electricidad y empleados con la producción de la empresa. Para ello necesitamos los coeficientes de correlación lineal entre todas.

En el menú de R Commander elegimos la opción ‘‘`Estadísticos -> Resúmenes -> Matriz de correlaciones`’’. En la ventana emergente debemos seleccionar las variables que queremos estudiar, `C.agua, C.electrico, Empleados` y `Beneficios` (ver Figura **??**), dejando el resto de las opciones con sus valores predeterminados.

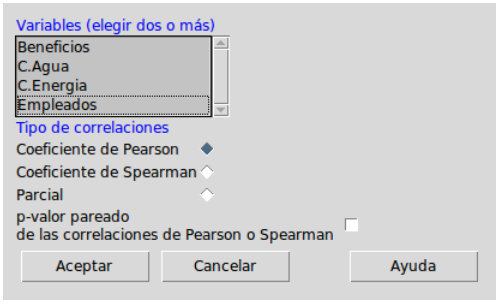


Figure 5.2: Entrada Matriz de correlaciones.

El resultado es la siguiente matriz.

	Beneficios	*C.Agua*	*C.Energia*	*Empleados*
Beneficios	1.00000000	0.8651403	0.97313596	0.01644869
C.Agua	0.86514028	1.0000000	0.78128262	0.20837191
C.Energia	0.97313596	0.7812826	1.00000000	−0.03099196
Empleados	0.01644869	0.2083719	−0.03099196	1.00000000

Vamos a comentar con detalle los resultados:

- En la diagonal, obviamente, aparecen unos, ya que el coeficiente de correlación lineal de una variable consigo misma siempre es uno.

- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es, es decir, el coeficiente de la variable X con la variable Y es idéntico al de la variable Y con la variable X.
- El coeficiente de correlación lineal entre el número de empleados y el beneficio, aunque es ligeramente positivo, es casi cero. Esto que podría indicar un grado casi nulo de relación directa entre ambas variables, esto es, que el número de empleados no está relacionado con los beneficios.
- Algo mejor es la relación del número de empleados con el consumo de agua, aunque sigue siendo bajo. Estos datos vendrían a decir, si se confirman, que parece haber una ligera tendencia a que las empresas con mayor numero de empleados tienen un mayor consumo de agua.
- El hecho de que el número de empleados esté relacionado negativamente con el consumo de agua, aunque muy bajo, indicaría una relación indirecta (si se confirmase) pero lo que parece indicar es que en realidad no está relacionado.
- Por último se observa una relación muy evidente entre cualesquiera dos de las variables consumo de agua, consumo energético y beneficios.

De todas formas, hasta ahora sólo hemos analizado estos coeficientes de una forma descriptiva, cuando lo más interesante es responder con claridad a la pregunta de si existe o no una relación estadísticamente significativa entre esos indicadores. Dicho de otra forma, refiriéndonos, por ejemplo, al coeficiente 0.20837191, ¿es suficientemente grande como para que indique una relación estadísticamente significativa o su valor podría deberse al azar? Esa misma pregunta es válida para los otros dos coeficientes (0.01644869 y -0.03099196) que, en principio, nos ha parecido que son demasiado pequeños (en valor absoluto) para tenerlos en cuenta. ¿será lo que acabamos de inferir cierto?

Lo que implícitamente estamos haciendo con estos comentarios es plantear la necesidad de contrastar si estos valores muestrales de los coeficientes son capaces, o no, de demostrar que los coeficientes de correlación poblaciones son significativamente distintos de cero. Esto último ha de hacerse de dos en dos, no con las cuatro variables a la vez.

A modo de ejemplos estudiaremos los pares de variables {`Beneficios`, `C.agua`}, {`Beneficios`, `C.energia`} y {`Beneficios`, `empleados`} por ser la relación con los beneficios la que más nos interesa.

5.2.1 Beneficios vs. C.agua

Elegimos la opción "`Estadísticos -> Resúmenes -> Test de correlación`" y en la ventana emergente seleccionamos las variables `Beneficios` y `C.agua`. Tenemos, además, la opción de elegir si el test es bilateral o unilateral. En nuestro caso la hipótesis alternativa es que es distinto de cero, luego elegimos bilateral.

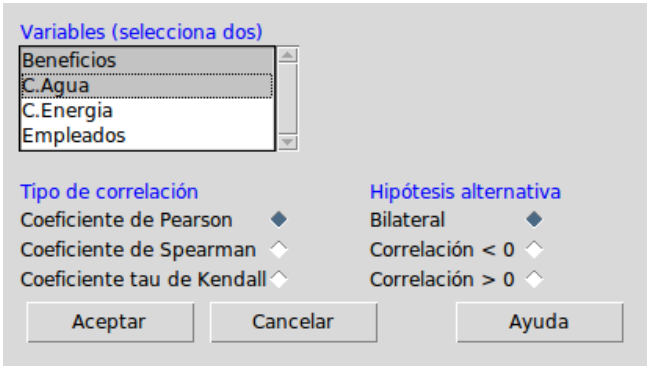


Figure 5.3: Entrada test de correlaciones.

Los resultados son los siguientes:

```
------------------------------------------------------------
        Pearson's product-moment correlation

data:  Datos$Beneficios and Datos$C.Agua

t = 16.3648, df = 90, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
 0.8025273 0.9089038

sample estimates:
      cor
0.8651403

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

Fijémonos que aparecen las variables que hemos elegido, el valor del estadístico `t= 16.3648`, los grados de libertad `df=90` y, lo que es más importante, el p-valor `p-value < 2.2e-16`. En este caso, es muy por debajo de 0.05, así que podemos rechazar la hipótesis nula (el coeficiente es cero) en favor de la alternativa (el coeficiente es distinto de cero) o, dicho de otra forma, **hemos encontrado indicios de que existe una relación estadísticamente significativa entre el beneficio de la empresa y el consumo de agua**.

Aparece además, un intervalo de confianza al 95% para el coeficiente de correlación lineal (que no contiene al cero, como cabía esperar) y el valor muestral del coeficiente.

5.2.2 Beneficios vs. C.energía

Repetimos el proceso con las variables `Beneficios` y `C.energia`.

Los resultados son los siguientes:

```
------------------------------------------------------------
        Pearson's product-moment correlation

data:  Datos$Beneficios and Datos$C.Energia

t = 40.0987, df = 90, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
 0.9595768 0.9821883

sample estimates:
     cor
0.973136
------------------------------------------------------------
```

Fijémonos que aparecen las variables que hemos elegido, el valor del estadístico `t= 40.0987`, los grados de libertad `df=90` y, lo que es más importante, el p-valor `p-value < 2.2e-16`. En este caso, es muy por debajo de 0.05, así que podemos rechazar la hipótesis nula (el coeficiente es cero) en favor de la alternativa (el coeficiente es distinto de cero) o, dicho de otra forma, **hemos encontrado indicios de que existe una relación estadísticamente significativa entre el beneficio de la empresa y el consumo de energía**.

Aparece además, un intervalo de confianza al 95% para el coeficiente de correlación lineal (que, de nuevo, no contiene al cero, como cabía esperar) y el valor muestral del coeficiente.

5.2.3 Beneficios vs. empleados

Lo volvemos a repetir con las variables `Beneficios` y `empleados`.

Los resultados son los siguientes:

```
------------------------------------------------------------
        Pearson's product-moment correlation
```

```
data:  Datos$Beneficios and Datos$Empleados

t = 0.1561, df = 90, p-value = 0.8763

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:
 -0.1890055  0.2205232

sample estimates:
       cor
0.01644869
------------------------------------------------------------
```

Fijémonos que aparecen las variables que hemos elegido, el valor del estadístico `t= 0.1561`, los grados de libertad `df=90` y, lo que es más importante, el p-valor `p-value = 0.8763`. En este caso, es muy superior a 0.05, así que no podemos rechazar la hipótesis nula (el coeficiente es cero) en favor de la alternativa (el coeficiente es distinto de cero) o, dicho de otra forma, **no hemos encontrado indicios de que exista una relación estadísticamente significativa entre el beneficio de la empresa y el número de empleados.**

Aparece además, un intervalo de confianza al 95% para el coeficiente de correlación lineal **(que contiene al cero, como cabía esperar)** y el valor muestral del coeficiente.

5.3 Ajuste de la recta de regresión

En el ejemplo que acabamos de ver no se detecta relación lineal entre las variables `Beneficios` y `Empleados` luego tiene poco sentido que nos planteemos un ajuste, es decir, un modelo, para dicha relación. Vamos a considerar, por tanto, otro ejemplo donde sí tenga sentido analizarlo como es la relación entre la variable `Beneficios` y `C.energia`.

Sabemos que el coeficiente de correlación lineal entre estas dos variables es 0.973136. El p-valor asociado a este valor muestral para contrastar si el coeficiente de correlación es distinto de cero es inferior a $2.2e-16$. Obviamente, esto indica que hay una clara relación entre las variables desde el punto de vista estadístico. Desde un punto de vista más práctico, podemos afirmar, por tanto, que existe una tendencia lineal bastante intensa.

Por otra parte, preguntémonos por si podemos considerar si una de las variables es la causa y otra el efecto en esta relación detectada. Hay ocasiones en la que esto no es posible o no están claras. En este caso, sin embargo, resulta difícil pensar que el consumo energético es consecuencia de los beneficios y sí es más fácil pensar que son los beneficios los que se ven afectados por el consumo energético. Así pues, desde el punto de vista de la regresión, la variable `Beneficios` sería la variable dependiente y `C.energetico`, la variable independiente.

¿Qué sentido puede tener ajustar la recta de regresión para estas variables?

- Podríamos plantearnos, por ejemplo, analizar en qué medida se vería afectado el beneficio ante un cambio en el consumo energético.
- Si se desconociera el dato de beneficio de una empresa, podríamos tratar de predecirlo utilizando la recta.
- Podremos analizar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias. Si, por ejemplo, una empresa tiene menos beneficios de lo que se espera según la recta, podríamos en primer lugar preguntarnos cuál es la causa y, en segundo lugar, iniciar una campaña de revisión del consumo para ajustarlo al margen de beneficios.

5.3.1 Cálculo de la recta de regresión

Para obtener el ajuste de la recta de regresión entre las dos variables, elegimos la opción ‘‘`Estadísticos -> Ajuste de modelos -> Regresión lineal`’’. La ventana emergente (Figura **??**) nos pide que especifiquemos la variable explicada (o variable dependiente) y la variable explicativa (o variable dependiente). Adicionalmente, nos permite ponerle un nombre al modelo resultante (por ejemplo `RegModel.1`).

Los resultados que aparecen en la ventana son los siguientes:

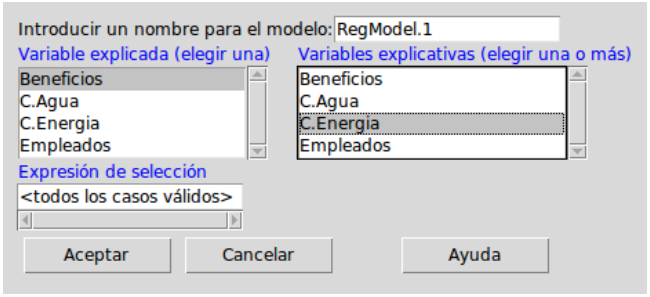


Figure 5.4: Entrada regresión lineal.

```
------------------------------------------------------------
Call:
lm(formula = Beneficios ~ C.Energia, data = Datos)

Residuals:
     Min       1Q   Median       3Q      Max
-4010.07  -441.83  -229.04    22.41  8983.68

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 334.07164  193.54386   1.726   0.0878 .
C.Energia     2.55396    0.06369  40.099   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1606 on 90 degrees of freedom

Multiple R-squared: 0.947,Adjusted R-squared: 0.9464

F-statistic:  1608 on 1 and 90 DF,  p-value: < 2.2e-16
------------------------------------------------------------
```

Analicemos los resultados:

1. La estimación del valor del parámetro β_0 (`Intercept`) es 334.07164. Hipotéticamente, se interpretaría como el beneficio estimado o el promedio de los beneficios con 0 empleados, lo que, claro está, no tiene sentido.
   A continuación lo que aparece es: el error estandar de esa estimación, 193.54386, el valor del estadístico de contraste (1.726) y el p-valor del contraste de $H_0 : \beta_0 = 0$ frente a $H_1 : \beta_0 \neq 0$, (observad que es un poco alto, $0.0878 > 0.05$,por lo que no podemos rechazar la hipótesis nula, esto es que $\beta_0 = 0$, con una significación $\alpha = 0.05$; aunque sí podríamos con $\alpha = 0.10$).
2. La estimación de β_1 es 2.55396, con un error estandar de la estimación de 0.06369. El contraste de $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$ arroja un valor del estadístico de 40.099 y un p-valor también inferior a $< 2e-16$.
   La recta ajustada aparece, por tanto, especificada a través de sus dos coeficientes: el término independiente (`Intercept`) y la pendiente de la recta:

$$\texttt{Beneficios} = 334.07164 + 2.55396 \times C.Energetico.$$

3. El error estandar del ajuste tiene un valor de 1606 con 90 grados de libertad.
4. El coeficiente R^2, aunque es obvio de calcular, aparece en la penúltima línea. Su valor, 0.947, indica que casi el 94.7% de toda la variabilidad que tiene el beneficio puede ser explicada por el consumo energético. Lo que desde luego resulta sorprendente.

Aparte del propio ajuste de los parámetros β_0 y β_1, podemos también obtener intervalos de confianza al nivel que deseemos de los mismos, sin más que clicar en ‘‘`Modelos -> Intervalos de confianza`’’. En la ventana emergente sólo tenemos que elegir el nivel de confianza deseado. En el caso de elegir el 90%, el resultado que aparece es el siguiente:

```
-----------------------------------------------------------
               Estimate       5 %         95 %
(Intercept)   334.071641   12.409270   655.734013
C.Energia       2.553964    2.448110     2.659817
-----------------------------------------------------------
```

5.4 Ejercicios Propuestos

Ejercicio 5.1 Un Broker quiere analizar, mediante un modelo de regresión lineal simple, la relación entre la posición que ocupa la empresa en el ranking (`rank`) y el valor de mercado de la misma (`marketvalue`). Para ello, ha utilizado el fichero de datos `Forbes` del paquete `HSAUR`. Se pide:

a Ajustar dicho modelo.

b ¿Puede concluir el broquer que existe relación estadísticamente significativa entre el valor de mercado y la posición en el ranking? Utilícese un 5% de significación para la respuesta.

c Calcular el porcentaje de bondad de ajuste (R^2) y decidir si el modelo proporcionaría predicciones muy fiables del valor de mercado en función de la posición.

Nota: Ver anexo ■

Ejercicio 5.2 En el fichero `notas.txt` están las horas de estudio, la nota de las actividades dirigidas y la calificación final de los alumnos de la asignatura de estadística del curso 2005-2006. Se pide:

a Ajustar un modelo de regresión lineal simple para explicar la nota final con respecto al tiempo dedicado al estudio.

b ¿Podemos concluir, con un 95% de confianza, que el tiempo de estudio es relevante para la nota final?

c Realizar una predicción de la nota que obtendrá uun alumno que ha estudiado 20 horas, mediante un valor puntual y mediante un intervalo de predicción al 95%.

d Realizar una predicción de la nota promedio que obtendrán los alumnos que hayan estudiado 18 horas, mediante un valor puntual y mediante un intervalo de predicción al 95%.

e Realizar un análisis gráfico de la relación lineal entre ambas variables.

■

Ejercicio 5.3 En el fichero `notas.txt` están las horas de estudio, la nota de las actividades dirigidas y la calificación final de los alumnos de la asignatura de estadística del curso 2005-2006. Se pide:

a Ajustar un modelo de regresión lineal simple para explicar la relación lineal entre la nota final y la nota de las actividades dirigidas.

b ¿Podemos concluir, con un 95% de confianza, que las notas de las actividades es un buen indicador lineal de la nota final?

c Realizar una predicción de la nota que obtendrá uun alumno que ha obtenido un 6 en las actividades mediante un valor puntual y mediante un intervalo de predicción al 95%.

d Realizar una predicción de la nota promedio que obtendrán los alumnos que hayan obtenido un 6 en las actividades, mediante un valor puntual y mediante un intervalo de predicción al 95%.

e Comparar y explicar los resultados de los dos apartados anteriores.

f Realizar un análisis gráfico de la relación lineal entre ambas variables.

g Comparar y explicar los resultados del ejercicio con los resultados del ejercicio anterior.

■

Anexo

Para realizar la actividad dirigida necesitamos los datos del fichero `Forbes2000` que se encuentran en el paquete `HSAUR`. Para ello primero cargaremos el paquete en `Herramientas->Cargar paquete(s)` y después cargaremos el fichero seleccionándo `HSAUR` en `Datos->Conjunto`

`de datos en paquetes -> Leer conjunto de datos desde paquete adjunto` y como segunda opción `Forbes2000`.

Al hacerlo obtenemos como conjunto de datos activo un conjunto con 2000 observaciones que contienen las siguientes 8 variables

- **rank** el ranking de la compañía.
- **name** el nombre de la compañía.
- **country** el país donde está situada la compañía.
- **category** lo que la compañía vende.
- **sales** las ventas de la empresa en billones de dólares.
- **profits** el beneficio de la empresa en billones de dólares.
- **assets** los activos de la empresa en billones de dólares.
- **marketvalue** el valor de mercado de la empresa en billones de dólares.

6. Introducción a la Programación Lineal

6.1 Introducción

La programación lineal estudia la optimización (en el sentido de maximizar o minimizar) de una función lineal sujeta a restricciones, también lineales, de desigualdad. Desde que George B. Dantzig desarrolló el método simplex en 1947 la programación lineal se ha utilizado ampliamente para la resolución de numerosos problemas entre los que podemos citar el transporte de bienes, la optimización de la producción industrial (refinerías de petróleo) o la asignación óptima de tareas.

De manera general un problema de programación lineal se escribe como:

$$\begin{array}{rccccccc} \max / \min & c_1x_1 & + & \cdots & + & c_nx_n & & \\ s.a: & a_{11}x_1 & + & \cdots & + & a_{1n}x_n & \leq & b_1 \\ & \vdots & & & & \vdots & \vdots & \vdots \\ & a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & \leq & b_m \end{array}$$

Los coeficientes $c_1, \ldots c_n$ son los **coeficientes de coste**, los a_{ij} se denominan **coeficientes tecnológicos**, y $b_1, \ldots, b_m$ forman el **vector de recursos**. Las **variables de decisión** son $x_1, \ldots, x_n$.

Nótese que el problema anterior se puede escribir en formato matricial como

$$\begin{array}{rccc} \max / \min & C^tX & & \\ s.a: & AX & \leq & B \end{array}$$

donde

$$C = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

A continuación veremos algunos ejemplos de problemas de programación lineal y cómo resolverlos utilizando para ello el software **R**.

6.2 Algunos ejemplos de aplicación

6.2.1 El problema del transporte

Consideremos n plantas de producción y m mercados de consumo. Cada planta puede producir como máximo a_i toneladas de producto y cada mercado demanda una cantidad de b_j toneladas (se supone que la oferta

total supera a la demanda, con objeto de que el problema tenga solución). El coste de transporte entre la fábrica i y el mercado j, de cada tonelada de producto, es de c_{ij} euros. El objetivo es satisfacer la demanda de todos y cada uno de los mercados con el mínimo coste.

Más concretamente, vamos a considerar la siguiente instancia del problema en la que se consideran 2 orígenes y tres destinos. La tabla refleja la distancia, en miles de kilómetros, desde cada origen a cada destino y el coste de transporte de una tonelada de producto es de 90 euros por cada 1000 kilómetros.

	Distancias (en miles de Km.)			Producción
	Mercados			
Plantas	Madrid	Badajoz	Santander	
Huelva	0.632	0.251	0.962	600
Cartagena	0.401	0.675	0.794	350
Demandas	325	300	275	

Las variables de decisión del problema serán las toneladas de producto transportadas desde la planta i hasta el mercado j y las denotaremos por x_{ij}, $1 \leq i \leq 2$, $1 \leq j \leq 3$. La formulación del problema sería entonces:

■ Ejemplo 6.1

$$
\begin{array}{lll}
\min & 90\cdot 0.632x_{11}+90\cdot 0.251x_{12}+90\cdot 0.962x_{13}+90\cdot 0.401x_{21}+90\cdot 0.675x_{22}+90\cdot 0.794x_{23} & \\
s.a: & x_{11}+x_{12}+x_{13}\leq 600 & \text{(Restricción 1)} \\
& x_{21}+x_{22}+x_{23}\leq 350 & \text{(Restricción 2)} \\
& x_{11}+x_{21}\geq 325 & \text{(Restricción 3)} \\
& x_{12}+x_{22}\geq 300 & \text{(Restricción 4)} \\
& x_{13}+x_{23}\geq 275 & \text{(Restricción 5)} \\
& x_{ij}\geq 0, i\in\{1,2\}, j\in\{1,2,3\} &
\end{array}
$$

■

esto es, minimizar el coste de enviar el producto solicitado desde las dos plantas, hasta los tres mercados, teniendo en cuenta que desde cada planta no se puede enviar una cantidad de producto superior a la oferta de cada una (restricciones 1 y 2) y que, a cada mercado, debe enviarse una cantidad total de producto que satisfaga la demanda correspondiente (restricciones 3, 4 y 5).

6.2.2 Problema de planificación de la producción

Una refinería tiene equipamiento para un proceso I, con capacidad de proceso de 2 Tm. de petróleo por día, y para un proceso II, con capacidad de proceso de 3 Tm. de petróleo por día. Como resultado, ambos procesos generan dos combustibles A y B. El proceso I requiere 20 horas de trabajo para convertir 1 Tm. de petróleo en 3/4 Tm. de A y 1/4 Tm de B, mientras que el proceso II requiere 10 horas de trabajo para convertir 1 Tm. de petróleo en 1/4 Tm. de A y 3/4 Tm. de B. El coste de la conversión es de 60 u.m. por tonelada de petróleo, para el proceso I y 30 u.m. por Tm. de petróleo para el proceso II. Se dispone de 40 horas de trabajo y 4 Tm. de petróleo por día, el precio de venta de A es de 285 u.m. y el de B de 105 u.m. >Cómo debe planificarse la producción diaria para maximizar el beneficio?

Usaremos las variables:

- x_1: Tm. de petróleo procesadas por el proceso I.
- x_2: Tm. de petróleo procesadas por el proceso II.

La cantidad obtenida de combustible de tipo A es $\frac{3}{4}x_1+\frac{1}{4}x_2$, la cantidad obtenida de combustible de tipo B es $\frac{1}{4}x_1+\frac{3}{4}x_2$ y la formulación del problema queda como

■ **Ejemplo 6.2**

$$\begin{array}{lll} \max & 285(\frac{3}{4}x_1+\frac{1}{4}x_2)+105(\frac{1}{4}x_1+\frac{3}{4}x_2)-60x_1-30x_2 & \text{(Beneficio obtenido)} \\ s.a: & x_1 \leq 2 & \text{(Capacidad máxima del proceso I)} \\ & x_2 \leq 3 & \text{(Capacidad máxima del proceso II)} \\ & 20x_1+10x_2 \leq 40 & \text{(Restricción relativa a las horas de trabajo)} \\ & x_1+x_2 \leq 4 & \text{(Restricción relativa al petróleo diario disponible)} \\ & x_1,x_2 \geq 0 & \end{array}$$

■

6.3 El método gráfico

Cuando el problema está formado por un máximo de dos variables, el método gráfico proporciona una herramienta sencilla pero eficaz para resolverlo.

■ **Ejemplo 6.3** Consideremos el problema

$$\begin{array}{lrcrcl} \max & f(x_1,x_2) = 2x_1+x_2 & & & & \\ s.a: & x_1 & + & x_2 & \leq & 5 \\ & x_1 & - & 2x_2 & \leq & 3 \\ & & & x_1,x_2 & \geq & 0 \end{array}$$

■

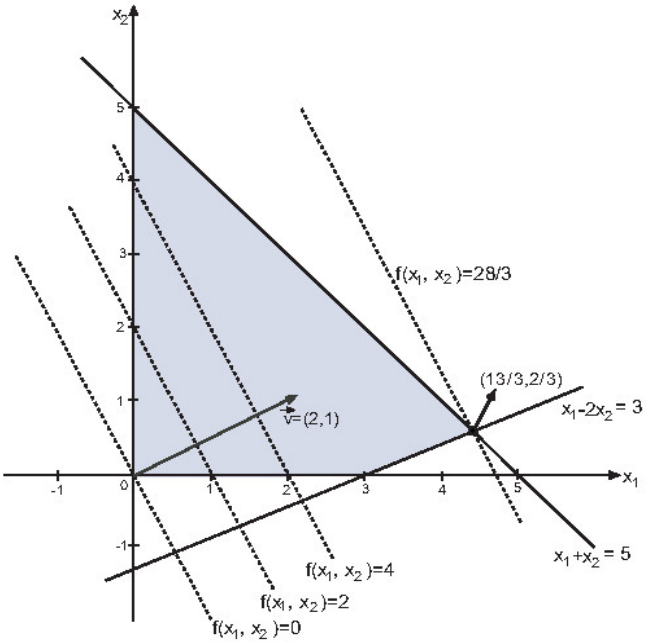


Figure 6.1: Región factible.

Cada una de las restricciones define un semiplano limitado por la recta cuya ecuación se obtiene al sustituir el símbolo '$\leq$' por '$=$'. De este modo, los puntos (x_1,x_2) que verifican la desigualdad $x_1+x_2 \leq 5$ definen un semiplano limitado por la recta $x_1+x_2=5$. Para definir si se trata del semiplano situado sobre dicha recta o bajo ella, tomaremos un punto cualquiera y veremos si sus coordenadas verifican la desigualdad o no. En caso afirmativo, el semiplano que estamos buscando será aquel al cual pertenece el punto considerado; en caso negativo, el semiplano buscado será aquél que no contiene al punto. Así pues vamos a considerar el punto $(x_1,x_2)=(0,0)$ que, efectivamente, verifica la desigualdad $x_1+x_2 \leq 5$ lo que nos indica que el semiplano buscado es el que contiene a dicho punto, esto es, el semiplano bajo la recta $x_1+x_2=5$.

De manera similar llegamos a la conclusión de que el conjunto de puntos que verifican la restricción $x_1 - 2x_2 \leq 3$ dan lugar al semiplano situado sobre la recta $x_1 - 2x_2 = 3$. Puesto que, adicionalmente, debe cumplirse $x_1 \geq 0$ y que $x_2 \geq 0$, la región factible corresponde a la región sombreada representada en la figura 6.1 y, como ya hemos dicho, está formada por el conjunto de puntos que verifican las restricciones.

En consecuencia, la solución del problema debe ser un punto de dicha región y ahora se trata de buscar dicho punto. En la figura, las rectas de trazo discontinuo representan las rectas $f(x_1,x_2) = k$ para los valores de k indicados en la figura. Podemos observar que si trazamos paralelas a la recta $f(x_1,x_2) = 0$ o, equivalentemente, movemos esa recta en la direccción del vector $(2,1)$ la función objetivo toma valor constante en los puntos de cada una de esas rectas y, además, dicho valor es mayor cuanto máyor sea el desplazamiento. Por lo tanto, para encontrar la solución del problema bastará desplazar la recta $f(x_1,x_2) = 0$, en la dirección del vector $(2,1)$ hasta que "*deje de tocar*" a la región factible. La solución del problema será el último punto de la región factible que toquemos al desplazar la recta paralela y que, en el ejemplo, es el punto $(x_1,x_2) = (2/3, 13/3)$. El valor óptimo de de la función objetivo será por tanto $f(2/3,13/3) = 28/3$.

6.4 Tipos de soluciones óptimas

Cuando resolvemos un problema de programación lineal se nos pueden presentar cuatro situaciones distintas, atendiendo al número de soluciones que presente el problema. Tomaremos como referencia un problema con dos variables, si bien los resultados se generalizan, de manera inmediata, a cualquier problema con tres o más variables.

- **El problema tiene una única solución óptima:** es la situación que se da en el ejemplo resuelto mediante el método gráfico.
- **El problema tiene solución múltiple.** Este caso se da cuando, al desplazar la recta definida por $f(x_1, x_2) = 0$, en la dirección de vector de coeficientes de la función objetivo (vector gradiente), la recta alcanza el óptimo tocando dos vértices de la región factible. En ese caso, todos los puntos del lado de la región factible, definido por ambos vértices, son soluciones óptimas del problema. Es lo que ocurre con el siguiente problema:

$$\begin{array}{lrcrcl} \max & f(x_1,x_2) = x_1 + x_2 & & & & \\ s.a: & x_1 & + & x_2 & \leq & 5 \\ & x_1 & - & 2x_2 & \leq & 3 \\ & & & x_1, x_2 & \geq & 0 \end{array}$$

- **El problema es infactible:** esto ocurre cuando no existe ningún punto que cumpla las restricciones del problema. Un ejemplo de este tipo de problema sería el siguiente:

$$\begin{array}{lrcrcl} \max & f(x_1,x_2) = 2x_1 + x_2 & & & & \\ s.a: & x_1 & + & x_2 & \leq & 5 \\ & x_1 & - & 2x_2 & \leq & 3 \\ & x_1 & & & \leq & 4 \\ & & & x_1, x_2 & \geq & 0 \end{array}$$

- **La solución del problema no está acotada.** En este caso la función objetivo puede crecer (o decrecer) de manera indefinida. El problema no tiene solución, no porque no existan puntos factibles, sino porque no existe un punto en el que se alcance la solución óptima. El siguiente problema ilustra esta situación:

$$\begin{array}{lrcrcl} \max & f(x_1,x_2) = 2x_1 + x_2 & & & & \\ s.a: & x_1 & + & x_2 & \leq & 5 \\ & x_1 & - & x_2 & \geq & 1 \\ & x_1 & - & 2x_2 & \leq & 3 \end{array}$$

Cuando se resuelve un problema de programación lineal con tres restricciones en lugar de dos, la región factible forma un poliedro en $\mathbb{R}^3$. Si tenemos n variables, ($n \leq 4$) la región factible será un politopo, que no es más que la generalización de un polígono bidimensional (o poliedro tridimensional) a un espacio de dimensión n. En todos estos casos, el tipo de solución óptima será uno de los indicados anteriormente, esto es, solución única, solución múltiple, solución no acotada o problema infactible.

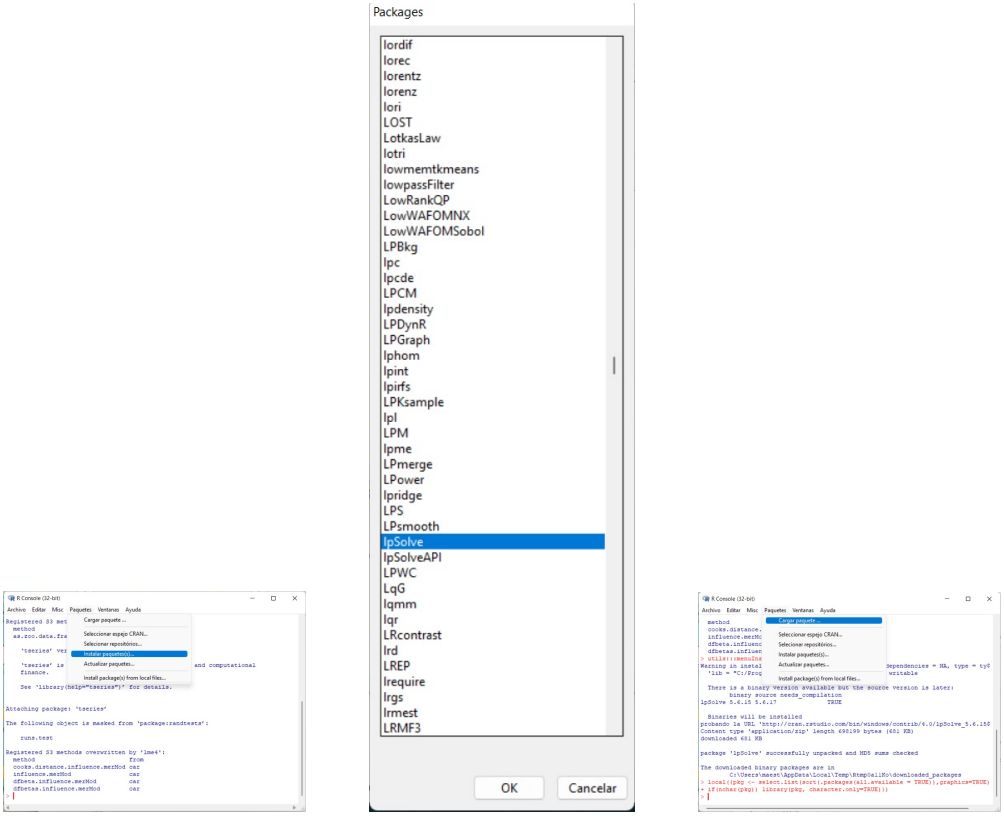
6.5 Resolución de un problema de optimización con R

R dispone de varios paquetes para la resolución de problemas de optimización. Usaremos el paquete **lpSolve** que permite resolver problemas de programación lineal. También admite variables enteras y variables binarias (programación 0-1).

El primer paso consiste en instalar el paquete **lpSolve**. Para ello, procedemos al igual que en la subsección 1.2.4 desde la consola de R,

Paquetes → **Instalar Paquete(s)**.

Con esto se nos presentará una lista de *mirrors* (no aparecerá si ya hemos instalado algún paquete previamente y R "*recordará esta elección*". Tras pulsar `Aceptar` , nos aparecerá una lista de paquetes. Seleccionamos **lpSolve** tal como se muestra en la Figura 6.2 (a) y (b). Posteriormente hemos de **cargar** el paquete instalado tal como se muestra en la Figura 6.2 (c) y (b).



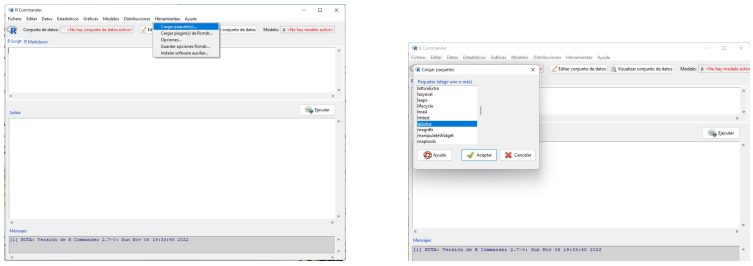
(a) (b) (c)

Figure 6.2: Instalación del paquete `lpSolve`.

Si queremos llamar a dicho paquete desde `R-Commander`. Para ello, desde la ventana gráfica de `R-Commander` seleccionamos

Herramientas → cargar paquete.

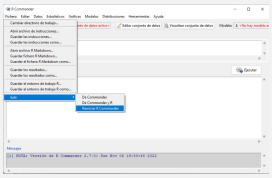
Tras pulsar `Aceptar` , nos aparecerá una lista de paquetes. Seleccionamos **lpSolve** tal como se muestra en la Figura 6.3 (a) y (b).



(a) (b)

Figure 6.3: Cargar el paquete `lpSolve` desde `R-Commander`.

Nota 6.5.1 Si al intentar cargar el paquete **lpSolve** desde `R-Commander`, éste no nos aprece en la ventana emergente, bast con `reiniciar R-Commander`



Reiniciar `R-Commander`.

Una vez cargado el paquete **lpSolve**, éste nos proporciona la función **lp()** que permite resolver problemas de programación lineal. También admite variables enteras y variables binarias (variables que sólo toman los valores 0 ó 1). La sintaxis de la función es la siguiente:

```
lp (direction, objective.in, const.mat, const.dir,
const.rhs, int.vec, binary.vec)
```

Describimos, a continuación, los argumentos de la misma.

- **direction:** variable de cadena de caracteres que indica el sentido de la optimización: **min:** minimizar, **max:** maximizar.
- **objective.in:** vector numérico con los coeficientes de la **función objetiva**.
- **const.mat:** matriz que contiene los coeficientes de las **restricciones**.
- **const.dir:** vector de cadenas de caracteres que indica el signo de las restricciones (`>=, =, <=`).
- **const.rhs:** vector de términos independientes de las restricciones.
- **int.vec:** vector que indica las posiciones de las variables que son enteras.

La salida de la función es una lista con componentes en la que encontramos (entre otras cosas):

- **objval:** valor de la función objetivo.
- **solution:** solución óptima.
- **status:** indicador numérico. **0:** éxito, **2:** no hay solución factible, **3:** solución no acotada.

La función **lp()** asume que las variables de decisión son mayores o iguales que 0. Si el problema tiene variables negativas o sin restricción, mediante un sencillo cambio de variable, podemos convertir estas variables en otras variables positivas.

6.5.1 Ejemplos

■ **Ejemplo 6.4** Resolver el problema formulado en la sección 6.3. ■ ■

Previamente a la resolución del problema es necesario instalar el paquete **lpSolve**, si no se ha hecho con anterioridad, y cargarlo para su uso como se detalló anterirormente. Recuérdese que el problema a resolver viene dado por las ecuaciones

$$\begin{array}{lrcrcl} \max & f(x_1,x_2) = 2x_1+x_2 & & & & \\ s.a: & x_1 & + & x_2 & \leq & 5 \\ & x_1 & - & 2x_2 & \leq & 3 \\ & & & x_1,x_2 & \geq & 0 \end{array}$$

Procedemos introducir los distintos vectores de datos en variables que después usaremos como argumentos de la función **lp()**. Nótese que también podríamos utilizar los datos, directamente, como argumentos de la función pero el almacenarlos previamente en variables permitirá su modificación, si fuera necesario, de manera más sencilla.

En la ventana *RScript* de **R-Commander** introducimos las siguientes instrucciones[1].

```
direction = "max"
objective.in = c(2 ,1)
const.mat = matrix( c(1 ,1 ,1,-2), nrow =2 , byrow = TRUE)\footnote{}
const.dir = c( "<=","<=")
const.rhs = c(5 ,3)
```

[2]

Finalmente, llamamos a la función **lp()** para resolver el problema:

```
sol = lp(direction , objective.in, const.mat, const.dir, const.rhs )
```

Una vez hecho esto seleccionamos todas las instrucciones en la ventana *R Script* y pulsamos en el botón ejecutar. De este modo se guarda la salida de la función **lp** en el objeto **sol**.

Para visualizar el valor óptimo de la función objetivo, tenemos que mostrar los elementos **objval** y **solution** del objeto **sol** que hemos creado previamente. Esto lo hacemos mediante las instrucciones

```
sol$objval
sol$solution
```

que debemos escribir y ejecutar desde la ventana *RScript*. Obtenemos entonces un valor óptimo de la función objetivo igual a 9.33333 para los valores de las variables $x_1 = 4.3333333$ y $x_2 = 0.6666667$.

Si deseamos imponer que, por ejemplo, la segunda variable sea entera debemos añadir esta información como una entrada de la función **lp()**.

```
int.vec <- c (2)
```

la llamada a la función sería:

```
sol = lp (direction , objective.in, const.mat, const.dir, const.rhs, int.vec)
```

■ **Ejemplo 6.5** Resolver el problema de transporte formulado en la sección 6.2.1. ■ ■

Solución:
$x_{11} = 0, x_{12} = 300, x_{13} = 250, x_{21} = 325, x_{22} = 0, x_{23} = 25$.
Valor de la función objetivo: 41913.75.

[1] Nótese que las variables `direction`, `objective.in`, etc. podrían llamarse de cualquier otra forma

[2] Para introducir los datos de la matriz de restricciones creamos un vector con dichos coeficientes ordenados por filas, de izquierda a derecha y de arriba a abajo. Mediante la instrucción **nrow**, le indicamos a R el número de filas de la matriz de restricciones

■ **Ejemplo 6.6** Una empresa que fabrica ordenadores debe planificar la producción semanal de los mismos. La compañía fabrica 3 tipos de ordenadores: de mesa (A), portátil básico (B) y portátil de alto rendimiento (C). Todos los ordenadores que se montan durante la semana se venden, proporcionando un beneficio neto de 350, 470 y 610 euros, respectivamente. Los ordenadores A y B pasan un control de calidad y la empresa dispone de 120 h. para realizar estos controles. Los ordenadores de tipo C pasan otro control distinto y la empresa dispone de 48 h. a la semana para realizarlos. Cada control requiere de 1 h. El resto de operaciones de montaje requieren 10, 15 y 20 h. para los ordenadores de tipo A, B y C, respectivamente. La empresa dispone de una capacidad de trabajo de 2000 horas/semana ¿Cuántos ordenadores de cada tipo debe producir para maximizar el beneficio? ■ ■

La formulación del problema es la siguiente:

Variables, x_A, x_B , x_C: número de ordenadores que se deben fabricar de tipo A, B y C, respectivamente.

$$\begin{array}{lrcrcrcl} \max & f(x_A,x_B,x_C) = 350x_A + 470x_B + 610x_C & & & & & & \\ s.a: & x_A & + & x_B & & & \leq & 120 \\ & & & & & x_C & \leq & 48 \\ & 10x_A & + & 15x_B & + & 20x_C & \leq & 2000 \\ & & & & & x_A,x_B,x_C & \geq & 0 \end{array}$$

Solución:
$x_A = 120$, $x_B = 0$, $x_C = 48$.
Valor de la función objetivo: 66400.

6.6 Ejercicios

Ejercicio 6.1 Resolver el problema de programación lineal

$$\begin{cases} \max & 15x_1 + 10x_2 \\ \text{s.a.} & x_2 \leq 50 \\ & 1.5x_1 - x_2 \geq 20 \\ & x_1, x_2 \geq 0 \end{cases}$$

■

Ejercicio 6.2 Resolver el problema de programación lineal

$$\begin{cases} \min & 50x_1 - 20x_2 \\ \text{s.a.} & 1.5x_2 \leq 15 \\ & x_1 + x_2 \geq 10 \\ & x_1, x_2 \geq 0 \end{cases}$$

■

Ejercicio 6.3 Resolver el problema de programación lineal

$$\begin{cases} \min & 3x_1 - x_2 + 2x_3 \\ \text{s.a.} & 2x_1 - x_2 + x_3 \geq -1 \\ & x_1 + 2x_3 \geq 2 \\ & -7x_1 + 4x_2 - 6x_3 \geq 1 \\ & x_1, x_2, x_3, \geq 0 \end{cases}$$

■

Ejercicio 6.4 Resolver el problema de programación lineal

$$\begin{cases} \max & -x_1 - x_2 + 2x_3 \\ \text{s.a.} & -3x_1 + 3x_2 + x_3 \leq 3 \\ & 2x_1 - x_2 - 2x_3 \leq 1 \\ & -x_1 + x_3 \leq 1 \\ & x_1, x_2, x_3, \geq 0 \end{cases}$$

■

Ejercicio 6.5 Resolver el problema de programación lineal

$$\begin{cases} \max & 5x_1 - 2x_2 - x_3 \\ \text{s.a.} & -2x_1 + 3x_3 \geq -1 \\ & 2x_1 - x_2 + x_3 \leq 1 \\ & 3x_1 + 2x_2 - x_3 \geq 0 \\ & x_1, x_2, x_3, \geq 0 \end{cases}$$

■

Ejercicio 6.6 Resolver el problema de programación lineal

$$\begin{cases} \min & -2x_2 + x_3 \\ \text{s.a.} & -x_1 - 2x_2 \geq -3 \\ & 4x_1 + x_2 + 7x_3 \leq -1 \\ & 2x_1 - 3x_2 + x_3 \geq -5 \\ & x_1, x_2, x_3, \geq 0 \end{cases}$$

■

Ejercicio 6.7 Resolver el problema de programación lineal

$$\begin{cases} \max & 3x_1+4x_2+5x_3 \\ \text{s.a.} & x_1+2x_2+2x_3 \leq 1 \\ & -3x_1+x_3 \leq -1 \\ & -2x_1-x_2 \leq -1 \\ & x_1, x_2, x_3, \geq 0 \end{cases}$$

■

Ejercicio 6.8 Resolver el problema de programación lineal

$$\begin{cases} \max & x_1-2x_2-3x_3-x_4 \\ \text{s.a.} & x_1-x_2-2x_3-x_4 \leq 4 \\ & 2x_1+x_3-4x_4 \leq 2 \\ & -2x_1+x_2+x_4 \leq 1 \\ & x_1, x_2, x_3, x_4 \geq 0 \end{cases}$$

■

Ejercicio 6.9 Un pastelero tiene 150 kg de harina, 44 kg de azúcar y 27,5 kg de mantequilla para hacer dos tipos de pasteles P y Q. Para hacer una docena de pasteles de tipo P necesita 3 kg de harina, 2 kg de azúcar y 1 de mantequilla y para hacer una docena de tipo Q necesita 6 kg de harina, 1 kg de azúcar y 1 kg de mantequilla. Se tiene que satisfacer una demanda mínima de 2 docenas de tipo P y no superar las 20 docenas

El beneficio que obtiene por una docena de tipo P es 20$ y por una docena de tipo Q es 30$. ¿Cuál es el número de docenas que tiene que hacer de cada clase para que el beneficio sea máximo? ■

Ejercicio 6.10 En una confitería se elaboran tartas de nata y de manzana. Cada tarta de nata requiere medio kilo de azúcar y 8 huevos; y una de manzana, 1 kg de azúcar y 6 huevos. En la despensa quedan 10 kg de azúcar y 120 huevos.

Sabiendo que los precios de venta son 12 euros las tartas de nata y 15 euros las tartas de manzana, calcula cuántas tartas de cada tipo se debe hacer para que los ingresos por su venta sean máximos. ■

Ejercicio 6.11 En una granja hay un total de 9000 conejos. La dieta mensual mínima que debe consumir cada conejo es de 48 unidades de hidratos de carbono y 60 unidades de proteínas. En el mercado hay dos tipos de productos, A y B, que aportan estas necesidades de consumo. Cada envase A contiene 2 unidades de hidratos de carbono y 4 unidades de proteínas y cada envase de B contiene 3 unidades de hidratos de carbono y 3 unidades de proteínas. Sabiendo que cada envase A cuesta 0.24 € y que cada envase de B cuesta 0.20 €, determina:

(a) El número de envases de cada tipo que deben adquirir los responsables de la granja con objeto de que el coste sea mínimo y se cubran las necesidades de consumo mensuales de todos los conejos.

(b) El valor de dicho coste mensual.

■