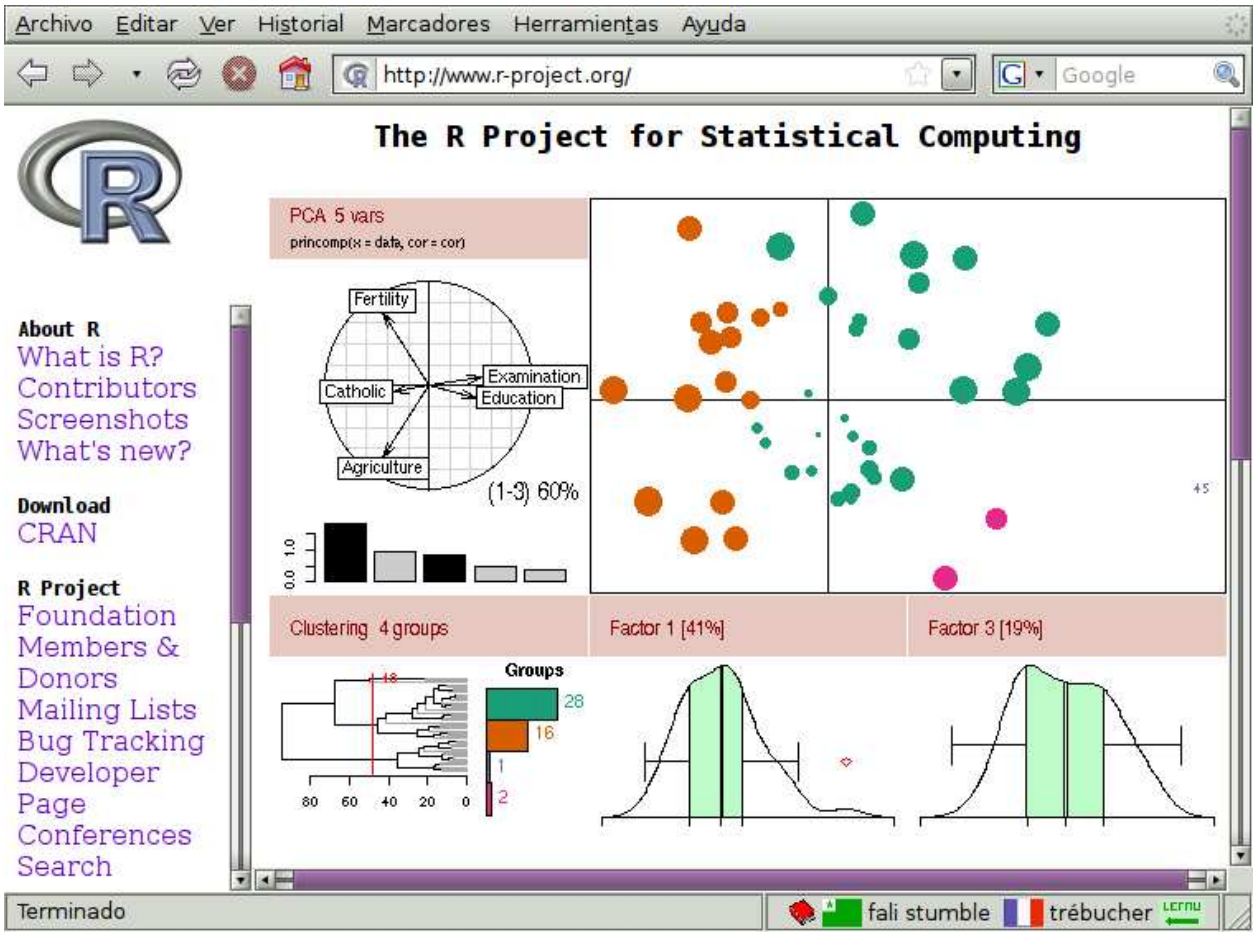
Estructura de R
y
Análisis descriptivo
de una variable

Estadística EUITIO

El paquete R[1] es una colección de programas libres[2] diseñada para el análisis estadístico de datos, que permite desde los análisis descriptivos más sencillos (como tablas de frecuencias simples) a procedimientos inferenciales más complejos (como el análisis de varianza o el análisis de componentes principales).

R realiza tres funciones esenciales: (1) leer datos, (2) especificar el tipo de análisis que se quiere realizar con esos datos y (3) mostrar los resultados obtenidos tras los análisis. La interpretación de los resultados es tarea del investigador.



[1]Sitio de referencia: `http://www.r-project.org`.

[2]En el sentido GNU: `http://gnu.org/philosophy/free-sw.es.html`.

Instalación de R

Debian GNU/Linux

Mediante algún gestor de paquetes, solicite la instalación de los paquetes `r-cran-rcmdr` y `r-cran-fbasics`. Por ejemplo, en la línea de órdenes:

```
# aptitude install r-cran-rcmdr r-cran-fbasics
```

Emacs dispone de una interfaz a R llamada ESS (*Emacs speaks statistics*). Puede instalarla mediante

```
# aptitude install ess
```

y ejecutarla desde Emacs mediante `M-x R`.

OpenBSD

Instale el paquete `R`:

```
$ sudo pkg_add R
```

y luego los paquetes de R llamados `Rcmdr` y `fBasics`:

```
$ R
> install.packages(c("Rcmdr","fBasics"))
```

FreeBSD

Instale el porte `R`:

```
# cd /usr/ports/math/R
# make install clean
```

y luego los paquetes de R llamados `Rcmdr` y `fBasics`:

```
$ R
> install.packages(c("Rcmdr","fBasics"))
```

ReactOS y Microsoft Windows

Si dispone de acceso a Internet:

1. Descargue el fichero ejecutable

   `http://cran.es.r-project.org/bin/windows/base/R-2.8.0-win32.exe`

2. Ejecute el fichero descargado, teniendo en cuenta:

   a) Cuando pregunta si deseamos establecer opciones de instalación, escoja *Sí*.

   b) Para el modo de presentación (MDI o SDI), escoja *SDI* (es conveniente por la implementación Tcl/Tk de Rcommander).

   c) Para la conexión a la red, escoja *Internet2* (al menos, dentro de la Universidad Oviedo, para obtener adecuadamente la información sobre el proxi).

3. Ejecute el programa R, ya instalado.

4. En el menú *Paquetes*, pinche en *Seleccionar espejo CRAN*.

5. En el cuadro de diálogo, escoja *Spain (Madrid)* o algún otro cercano, y pulse OK.

6. En el menú *Paquetes*, pinche en *Instalar paquete(s)*

7. Escoja *fBasics* y *Rcmdr* y acepte.

Si no tiene acceso a Internet, consiga los archivos necesarios en

`ftp://carleos.epv.uniovi.es/euitio/R`

Para instalar un paquete, en el menú *Paquetes*, escoja *Instalar desde ficheros ZIP locales* y seleccione el fichero correspondiente.

Configuración para la asignatura

Ciertas definiciones usadas en la parte teórica de la asignatura no se corresponden exactamente con las definiciones usadas por omisión en R (p.ej. varianza, cuantil). Para que el entorno utilice las definiciones acordes a la teoría, cargue al principio de la sesión el fichero de código fuente siguiente si su sistema usa codificación de signos UTF-8:

`ftp://carleos.epv.uniovi.es/euitio/oviedo-u8.R`

o, si todavía usa Latin-1, cargue en su lugar

`ftp://carleos.epv.uniovi.es/euitio/oviedo-l1.R`

Para cargar código, utilice la función `source` o en el menú *Archivo* escoja *Cargar código R*. Para no tener que repetir esta operación en cada sesión, incluya el código en `~/.Rprofile`.

Lenguaje R: tipos de objetos

En esta sección se realiza una descripción somera de los tipos de objetos disponibles en el lenguaje de programación R, que puede ser útil para interpretar la información que presentan las salidas de Rcommander.[3]

Tipos básicos

Los tipos de valores básicos en R son: lógicos (booleanos), numéricos y cadenas de caracteres.[4] Las constantes lógicas son `TRUE` (verdadero) y `FALSE` (falso)

```
> 1 == 1
[1] TRUE
> 1 == 2
[1] FALSE
> T
[1] TRUE
> F
[1] FALSE
```

Téngase en cuenta que `T` y `F` son objetos (variables del lenguaje) que valen inicialmente `TRUE` y `FALSE` respectivamente, aunque pueden ser redefinidos.

Las constantes numéricas adoptan la notación habitual en informática; por ejemplo, la notación exponencial o científica:

```
> 1e4
[1] 10000
```

El resultado de una operación matemática puede ser `Inf` (infinito) o `NaN` (*not a number*: no un número):

```
> 3/0
[1] Inf
> log(0)
```

[3] Para profundizar más, consúltese por ejemplo `http://cran.r-project.org/doc/contrib/rdebuts_es.pdf` o `http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf`.

[4] Detallando más, hay tres tipos numéricos (enteros, reales [llamados también dobles o de coma deslizante] y complejos) y un tipo adicional *crudo* para almacenar octetos.

```
[1] -Inf
> log(-1)
[1] NaN
Warning message:
In log(-1) : NaNs produced
> sqrt(-1)
[1] NaN
Warning message:
In sqrt(-1) : NaNs produced
> sqrt(-1+0i)
[1] 0+1i
> log(-1+0i)
[1] 0+3.141593i
```

Nótese cómo en el ejemplo algunos `NaN`es pueden evitarse usando números complejos.

Las cadenas pueden delimitarse por comillas o por apóstrofos, lo cual puede aprovecharse para incluir cómodamente comillas en la cadena[5]:

```
> "nombre"
[1] "nombre"
> 'Lo llamaban "Trinidad"'
[1] "Lo llamaban \"Trinidad\""
```

Queda mencionar la constante `NA` (*not available*: no disponible) que indica un valor *ausente*. Aparece, por ejemplo, al leer campos numéricos vacíos durante la importación de un fichero. Al igual que `NaN`, tiene comportamiento *viral* (contagia):

```
> 3 + NA
[1] NA
```

La asignación de valores a objetos se realiza mediante el operador `<-`[6]:

```
> a <- 8.5
```

Pero ¡ojo!

```
> a < -3
[1] FALSE
> a<-3
> a
[1] 3
```

Ten cuidado dónde pones los espacios.

Hasta ahora los valores considerados son simples; podrían llamarse *átomos* o *escalares*. En R hay también estructuras compuestas, entre las que veremos los vectores, las matrices, las listas y los dataframes.

[5] Sin necesidad de *escaparlas* mediante una retrobarra.

[6] Fuera de los argumentos de funciones puede sustituirse por `=`, así: `a = 8.5`.

Vectores

Podemos crear un vector mediante la función `c`, que concatena:

```
> b <- c(1,2,3,10)
```

Los escalares en R son siempre vectores de longitud 1, así que la asignación de `a` en la sección anterior es equivalente a:

```
> a <- c(8.5)
```

Toda la aritmética en R (así como muchas de sus funciones) es vectorial. Por ejemplo, se pueden sumar vectores:

```
> a+b
[1]  9.5 10.5 11.5 18.5
```

El `[1]` que aparece al comienzo del renglón es una ayuda para localizar posiciones en vectores largos:

```
> 10:50
 [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
[26] 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
```

Así, el 35 ocupa la posición vigésimo-sexta.

En el ejemplo anterior se ha utilizado el operador `:` para crear una secuencia. La función `seq` generaliza la creación de secuencias:

```
> seq(0,3,0.5)
[1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0
```

Es útil también la función `rep` para crear vectores con un elemento repetido:

```
> rep("A",5)
[1] "A" "A" "A" "A" "A"
```

Los vectores se pueden concatenar:

```
> bab <- c(b,a,b)
> bab
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0
```

pero tenga en cuenta que en un vector todos los elementos han de ser del mismo tipo:

```
> c(bab,"cadena")
 [1] "1"      "2"      "3"      "10"     "8.5"    "1"      "2"      "3"
 [9] "10"     "cadena"
```

Se puede acceder a los elementos de un vector indicando a su vez un vector de subíndices (enteros positivos):

```
> bab[4]
[1] 10
> bab[4:6]
[1] 10.0  8.5  1.0
> bab[c(4,6)]
[1] 10 1
```

Si los subíndices son negativos, indican exclusión:

```
> bab[-4]
[1]  1.0  2.0  3.0  8.5  1.0  2.0  3.0 10.0
> bab[-(4:6)]
[1]  1  2  3  2  3 10
> bab[-c(4,6)]
[1]  1.0  2.0  3.0  8.5  2.0  3.0 10.0
```

Matrices

Una matriz es la versión bidimensional de un vector:

```
> m <- matrix(1:6,nrow=2)
> m
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

Para obtener una submatriz, hay que indicar subíndices para filas y columnas; en blanco, para seleccionar toda la dimensión:

```
> m[1,2]
[1] 3
> m[1,]
[1] 1 3 5
> m[,2]
[1] 3 4
>  m[,2:3]
     [,1] [,2]
[1,]    3    5
[2,]    4    6
> m[,-1]
     [,1] [,2]
[1,]    3    5
[2,]    4    6
```

Se pueden pegar vectores o matrices por columnas mediante la función `cbind` para formar una nueva matriz:

```
> cbind(1:3,c(8,9,10))
     [,1] [,2]
[1,]    1    8
[2,]    2    9
[3,]    3   10
> cbind(m,c(100,200))
     [,1] [,2] [,3] [,4]
[1,]    1    3    5  100
[2,]    2    4    6  200
```

o por filas mediante la función `rbind`:

```
> rbind(1:3,c(8,9,10))
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    8    9   10
```

Se puede trasponer una matriz con la función `t`, hallar su determinante mediante `det`, y un largo etcétera.

Al igual que los vectores, todos los elementos de una matriz han de ser del mismo tipo (homogéneos):

```
> m[1,1] <- "nombre"
> m
     [,1]     [,2] [,3]
[1,] "nombre" "3"  "5"
[2,] "2"      "4"  "6"
```

Listas

Para almacenar elementos heterogéneos hay que recurrir a una lista:

```
> l <- list(bab,"cadena")
> l
[[1]]
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0

[[2]]
[1] "cadena"
```

Al subindicar con corchetes simples se obtiene una sublista:

```
> l[1]
[[1]]
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0
```

Al subindicar con corchetes dobles se puede extraer UN elemento de la lista:

```
> l[[1]]
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0
> l[[1]][4:6]
[1] 10.0  8.5  1.0
```

Los elementos de una lista pueden tener etiquetas[7]:

```
> n <- list(bab=bab,cadena="Cadena")
> n[[1]]
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0
> n$bab
[1]  1.0  2.0  3.0 10.0  8.5  1.0  2.0  3.0 10.0
> n$cadena
[1] "Cadena"
```

Dataframes

Un dataframe es una lista cuyos elementos son vectores, todos de igual longitud. Por tanto, se le pueden aplicar los operadores de listas y los de matrices. Es la estructura esencial para guardar valores de variables estadísticas:

```
> d <- data.frame(v1=c(2,4,8,9),v2=c("v","v","m","m"))
> d
  v1 v2
1  2  v
2  4  v
3  8  m
4  9  m
> d[1,2]
[1] v
Levels: m v
> d[1,]
  v1 v2
1  1  v
> d[,2]
[1] v v m m
Levels: m v
> d$v1
[1] 2 4 8 9
> d$v2
[1] v v m m
Levels: m v
```

Nótese que, al construir el dataframe, el vector de cadenas se ha convertido en "factor", es decir, en un vector acompañado de una lista de "niveles" (modalidades de la variable estadística).

[7] También los de los vectores y matrices, pero no nos ocuparemos de eso.

Funciones

Un objeto de R también puede almacenar una función:

```
> f <- function (x) x^2
```

que puede usarse como argumento de otras funciones

```
> sapply (5:10, sqrt)
[1] 2.236068 2.449490 2.645751 2.828427 3.000000 3.162278
> sapply (5:10, f)
[1]  25  36  49  64  81 100
```

Como se ve, la función `sapply` aplica una función a cada elemento de un vector, y devuelve un vector de resultados. La función `apply` hace lo mismo con matrices, bien por filas:

```
> m <- matrix(1:6,nrow=2)
> m
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> apply(m,1,sum)
[1]  9 12
```

bien por columnas:

```
> apply(m,2,sum)
[1]  3  7 11
```

Téngase en cuenta que `sum` es una función que calcula la suma de un vector.

Tema 1

Estructura de R

El objetivo de este primer tema es que el alumno aprenda a manejar el programa R. Para ello, primero hablaremos de su estructura: sus ventanas y los elementos que las constituyen (barras de menús, de elementos activos, etcétera).

1.1. Comienzo de sesión

Tras arrancar el programa, aparece una ventana titulada *Consola R* que indica la versión de R y cómo obtener información de la licencia de uso.

```
R version 2.6.2 (2008-02-08)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener mas informacion y
'citation()' para saber como citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

>
```

1.2. Ventana de órdenes (consola)

Por debajo del título, esta ventana contiene una barra con los siguientes menús, cuyas opciones principales destacamos:

Archivo Operaciones básicas con los ficheros. Sólo usaremos:

Salir Para salir del programa.

Editar Típico menú con opciones de edición (copiar, pegar...).

Misc Opciones avanzadas.

Paquetes Permite gestionar los paquetes adicionales de R. Nos interesará la opción:

Cargar paquete Para activar un paquete en concreto.

Ayuda Información abundante sobre R.

La consola permite una interacción con el intérprete de lenguaje R. A grandes rasgos, se trata de un lenguaje a alto nivel, al estilo de Octave/Matlab, pero orientado a la computación estadística[1].

1.3. Interfaz gráfica (Rcommander)

Desde el menú *Paquetes*, pinche en *Cargar paquete* y escoja *Rcmdr*. Aparece una interfaz gráfica[2] que permite acceder a muchas de las órdenes de gestión y análisis de datos del lenguaje R. La pantalla presenta el aspecto de la figura 1.1.

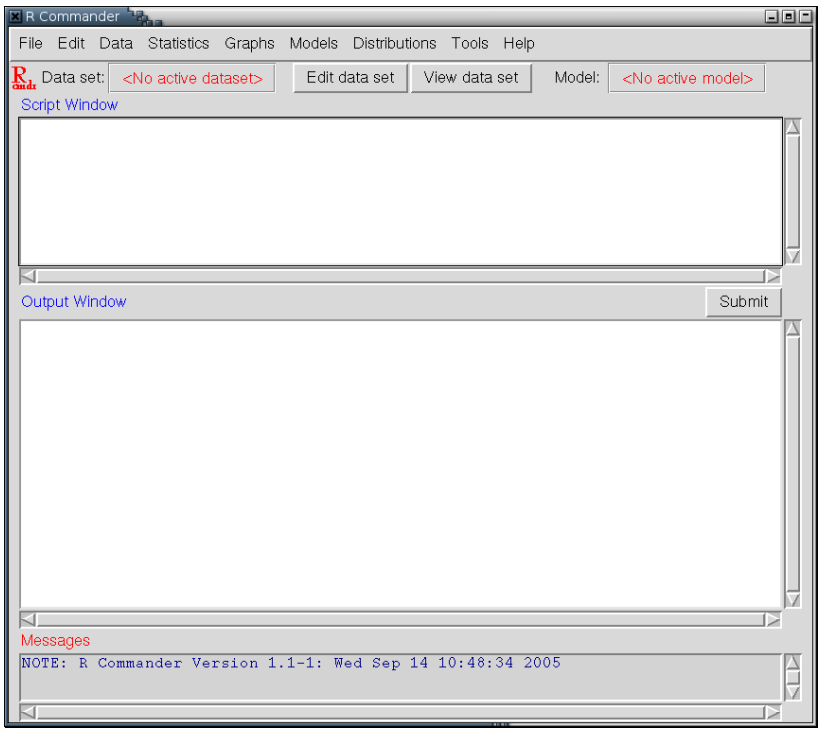


Figura 1.1: Aspecto inicial de la interfaz gráfica

En la parte superior puede observarse una barra que consta de una serie de menús (*Archivo*, *Editar*, *Datos*, etc.). Si se selecciona con el ratón cada una de ellas aparece un menú desplegable

[1] En realidad, es un lenguaje con la semántica de Lisp/Scheme (con clausuras), lo que lo hace mucho más elegante que Octave y Splus.

[2] Basada en Tcl/Tk.

donde se ofrecen otros submenús, cada uno de los cuales tiene a su vez un cuadro de diálogo que es el lugar donde se especifican los detalles de cada procedimiento.

Inmediatamente debajo aparece otra barra que indica el conjunto de datos activo (*Datos:*) y el modelo activo. Hay botones para ver los datos (*Visualizar datos*) y modificarlos (*Editar datos*).

A continuación viene la *ventana de instrucciones*. Muestra las órdenes de R correspondientes a las opciones de los menús escogidas por el usuario. Además, tales instrucciones se pueden modificar, y ejecutar mediante el botón *Ejecutar*.

La *ventana de resultados* contiene aquellas salidas de las órdenes ejecutadas que se muestran en formato de texto.

Finalmente, la ventana de *mensajes* recoge la información adicional que R nos quiere hacer llegar (por ejemplo, advertencias o mensajes administrativos).

Si por cualquier motivo salimos de la interfaz gráfica, podemos volver a acceder a ella escribiendo en la consola la orden `Commander()`.

A continuación, repasamos con más detalle cada una de las partes.

1.3.1. Barra de menús

En cada menú describimos solamente las opciones de interés para este curso:

Archivo Hay opciones para cargar o grabar instrucciones, resultados o el entorno de trabajo. También para salir de la interfaz gráfica, o también de R.

Editar Típico menú de edición. Permite seleccionar, cortar, copiar, pegar y buscar.

Datos Permite la gestión de los datos por analizar. R mantiene distintos conjuntos de datos dentro del entorno de trabajo. Uno (y sólo uno) de ellos se considera *activo*.

Un conjunto de datos es una matriz con variables como columnas y casos como filas. Lo comentaremos en la sección 2.

Estadísticas Recoge los diferentes métodos de análisis que se pueden aplicar al conjunto de datos activo. Se comentará a partir de la sección 3.

Gráficas Recoge los diferentes tipos de gráfico que se pueden obtener. Se comentará a partir de la sección 3.

Modelos Un conjunto de datos puede tener asociados varios modelos estadísticos. Este menú sirve para la gestión de los mismos.

Distribuciones Para trabajar con funciones de distribución de probabilidad: cuantiles, probabilidades y gráficas asociadas a las distribuciones normal, t, χ^2, F, binomial... Se empleará en el segundo trimestre.

Herramientas Este menú de utilidades contiene:

Cargar paquet(e) Para cargar paquetes de R adicionales. No será necesario para los contenidos del curso.

Opciones Para ajustar diferentes características de la interfaz, por ejemplo, el tamaño tipográfico.

Ayuda La ayuda de la interfaz gráfica es una extensión de la ofrecida por la consola.

Ayuda de R Commander Uso de la interfaz gráfica.

Introducción a R commander Artículo introductorio con imágenes.

Ayuda sobre los datos activos (si es posible) Misma opción que bajo el menú *Datos*.

Información sobre Rcmdr Versión y autores de la interfaz gráfica.

Además, la mayoría de los cuadros de diálogo dispone de un botón *Ayuda* que ofrece información sobre las órdenes de R asociadas a la acción correspondiente.

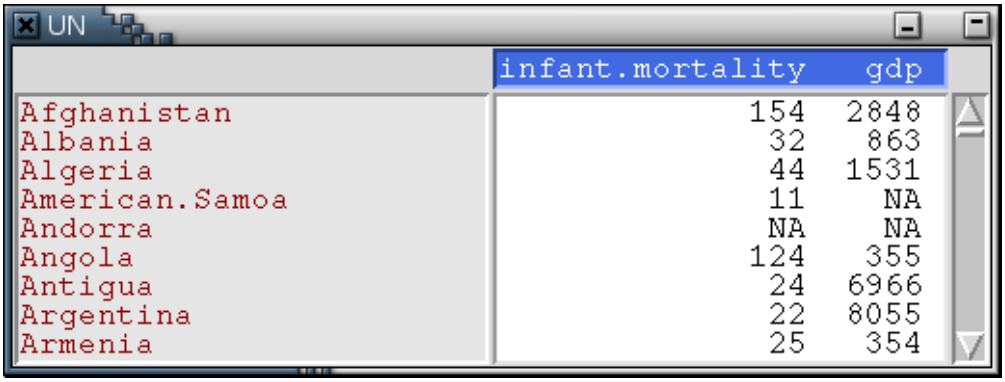
1.3.2. Barra de elementos activos

Consta de:

Datos Nombre del conjunto de datos activo, es decir, el que se toma por omisión a la hora de ejecutar una orden.

Editar datos Hace aparecer una cuadrícula donde es posible modificar el contenido del conjunto actual de datos.

Visualizar datos Muestra el contenido del conjunto actual de datos.



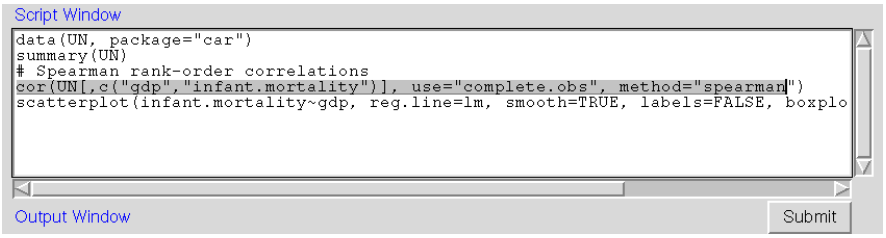
	infant.mortality	gdp
Afghanistan	154	2848
Albania	32	863
Algeria	44	1531
American.Samoa	11	NA
Andorra	NA	NA
Angola	124	355
Antigua	24	6966
Argentina	22	8055
Armenia	25	354

Modelo Para un mismo conjunto de datos se pueden crear diferentes modelos de análisis (de regresión lineal, de componentes principales...). Este menú permite escoger el modelo activo, es decir, aquél considerado por omisión cuando se ejecuta una orden.

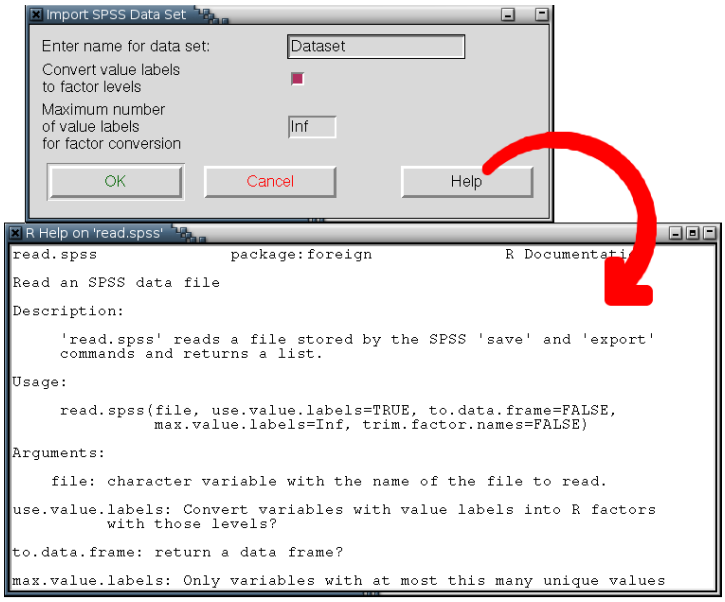
1.3.3. Ventana de instrucciones

Se puede acceder a muchas órdenes desde los menús y los cuadros de diálogo. No obstante, algunas órdenes y opciones sólo están disponibles mediante el uso del lenguaje R. Además, se pueden grabar los guiones en un fichero de texto (habitualmente con extensión `.R`) con lo que podrá repetir los análisis en otro momento o ejecutarlos en un trabajo automatizado.

Un fichero `.R` es simplemente un fichero de texto que contiene órdenes. Es posible escribir órdenes directamente en la ventana de guiones[3]. Sin embargo, es más sencillo permitir que el programa le ayude a construir un guión aprovechando que la realización de una acción desde un cuadro de diálogo añade la orden a la ventana de guiones. En ésta puede ser modificada para su posterior ejecución. Para ello, ha de seleccionar con el ratón la orden u órdenes y despúes ha de pinchar en el botón *Submit* o *Ejecutar*.



En el cuadro de diálogo de un procedimiento determinado, pulse en el botón *Help* o *Ayuda* para saber qué opciones del lenguaje R están disponibles (si hay alguna) para ese procedimiento. Si desea información completa sobre el lenguaje de órdenes, consulte el manual de referencia incluido con la documentación de R.



1.3.4. Ventana de resultados

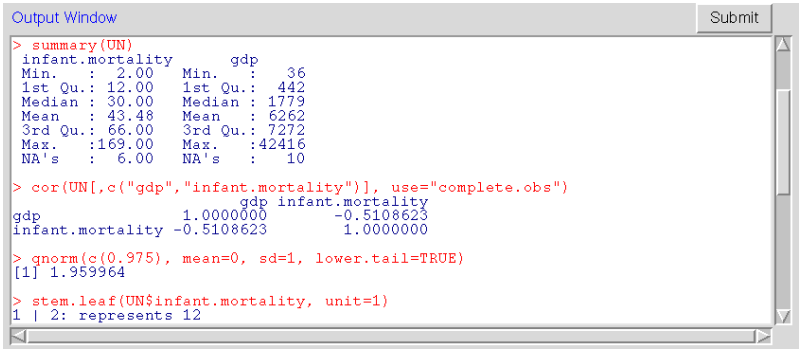
Una vez que se solicita un análisis con los datos, los resultados obtenidos se muestran en la ventana inferior, mostrada en la página siguiente.

El texto en rojo son las órdenes correspondientes que aparecen en la ventana de instrucciones. El texto en azul es el resultado de cada orden.

Los contenidos de la ventana de resultados son texto puro, que puede ser copiado a cualquier editor de texto para su procesamiento.

Podemos usar las ventanas de instrucciones y resultados a manera de calculadora. El lenguaje R permite las operaciones aritméticas básicas, la definición de variables y funciones y dispone de

[3]También en la consola.



una colección muy extensa de funciones ya definidas. En el siguiente ejemplo[4], definimos la función `varianza`.

```
> muestra <- c(1,3,5,7,33)
> mean(muestra)
[1] 9.8
> varianza <- function(x) mean(x^2) - mean(x)^2
> varianza(muestra)
[1] 138.56
```

1.3.5. Ventana de mensajes

Recoge las indicaciones y advertencias de R.

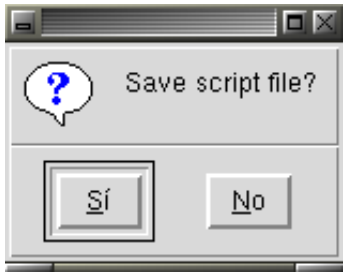
1.4. Fin de sesión

En el menú *Archivo*, óptese por (*Salir*)

De Commander Se abandona la interfaz gráfica, pero no la consola de R. Recuerde que para volver a la interfaz gráfica puede escribir `Commander()`.

De Commander y R Abandona el entorno R completamente.

En ambos casos se pide confirmación del abandono, y se pregunta si se quiere guardar el contenido de las ventanas de instrucciones y de resultados.



[4] R dispone de una función `var`, pero calcula la cuasivarianza.

Tema 2

Manejo de datos

En este tema aprenderemos a manejar los conjuntos de datos y a leer y almacenar en un fichero los datos necesarios para realizar un análisis. Estas tareas se realizan a través del menú *Datos*, cuyas opciones mostramos someramente a continuación:

Nuevos datos Para introducir nuevos datos por el teclado. Requiere dar un nombre a los datos nuevos, que no puede contener espacios ni caracteres especiales.

Importar datos Para leer datos contenidos en un fichero. Soporta varios formatos: texto puro, SPSS, Minitab...

Datos en paquetes R contiene una colección de datos de ejemplo, por si queremos ejercitarnos con el programa pero no disponemos de datos propios adecuados.

Datos activos Aquí se gestiona el conjunto de datos activo.

- **Seleccionar los datos activos** Elegir el conjunto de datos activo entre los que hay disponibles en ese momento en la sesión.
- **Ayuda sobre los datos activos (si es posible)** Algunos conjuntos de datos (como los de ejemplo) contienen una descripción.
- **Variable de los datos activos** Lista los nombres de las variables del conjunto de datos.
- **Establecer nombre de casos** A veces una variable no es tal, sino que contiene los nombres de los casos. Esta opción permite indicárselo a R.
- **Filtrar los datos activos** Si queremos que los análisis subsiguientes se realicen sobre una subconjunto de los casos, aquí podemos indicar una expresión de filtro. El filtro construye un nuevo conjunto de datos, cuyo nombre conviene indicar; en caso contrario, la selección se hace permanente (se eliminan los casos que no pasan el filtro).
- **Eliminar los casos sin datos** En algunas variables, puede que se desconozca el valor para cierto caso: se trata de un dato ausente (*missing*). Esta opción elimina los casos con algún dato ausente.
- **Exportar los datos activos** Para guardar una tabla con el conjunto de datos activo en un fichero de texto.

Modificar variables de los datos activos Para realizar trasformaciones en los datos.

- **Recodificar variable** Crea una nueva variable a partir de una ya existente. Sirve para agrupar datos cuantitativos en intervalos.

Calcular una nueva variable Crea una nueva variable a partir de una fórmula, la cual puede involucrar al resto de las variables.

Tipificar variables Para tifipicar variables cuantitativas.

Convertir variable numérica en factor Indica al programa que los números no representan cantidades, sino categorías.

Segmentar variable numérica Simplifica la agrupación de datos cuantitativos en intervalos, aunque por ello es menos flexible que Recode variable.

Renombrar variables Cambia el nombre de la variable.

Eliminar variables de los datos Elimina la variable.

2.1. Obtención de datos

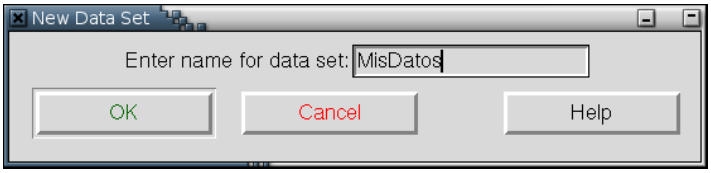
Podemos introducir datos directamente, leerlos de un fichero ya existente, o bien utilizar datos que R trae de ejemplo.

2.1.1. Creación de un conjunto de datos nuevo

Esta opción es conveniente cuando el conjunto de datos es pequeño. Para conjuntos de datos mayores, es más cómodo crear un fichero de datos por otros medios (por ejemplo, desde una hoja de cálculo o una base de datos) y luego importarlo.

Lo primero que hay que tener en cuenta y no olvidar es que los conjuntos de datos (*data set*, *data frame*) están organizados de forma matricial, donde las filas se refieren a los casos (individuos, unidades u observaciones) de la muestra y las columnas a las variables.

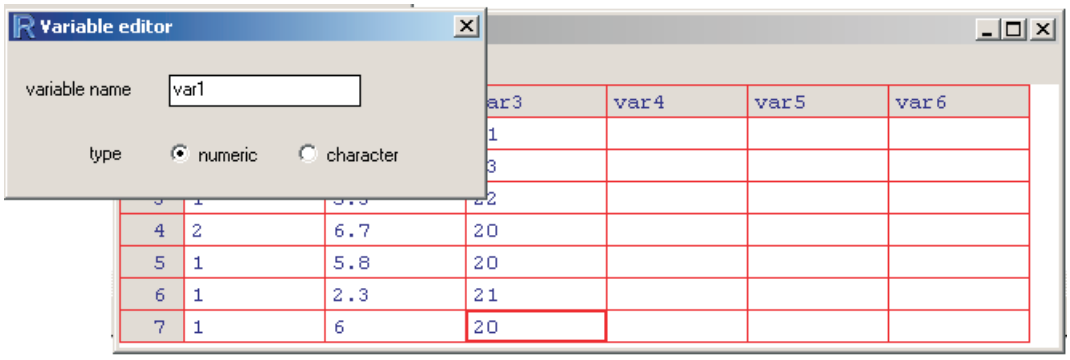
Para introducir nuevos datos ha de escogerse al opción *Nuevos datos* del menú *Datos*. Se nos pide entonces un nombre para el conjunto de datos (pues pueden manejarse varios simultáneamente).



Para introducir los datos simplemente se coloca el cursor en la celda correspondiente a cada individuo y variable. Para moverse de una celda a otra se puede utilizar el ratón, o las teclas del cursor y retorno para el desplazamiento vertical, o las teclas del cursor y tabulador para el desplazamiento horizontal.

Al introducir los datos, se observa que R da por omisión nombre a las variables (`var1`, `var2`, ...) y define sus características. En principio, una variable puede ser numérica (*numeric*) o de caracteres (*character*). Si se desea cambiar el nombre o definir el tipo de variable hay que pulsar en la cabecera de la columna correspondiente.

Téngase en cuenta que para representar variables cualitativas, es decir, aquéllas cuyos valores toman un número finito de modalidades, utilizaremos *factores*. Un factor se crea a partir de una variable de tipo numérico, a cuyas categorías se asignan etiquetas, como veremos más adelante. NO USE VARIABLES DE TIPO CARÁCTER.

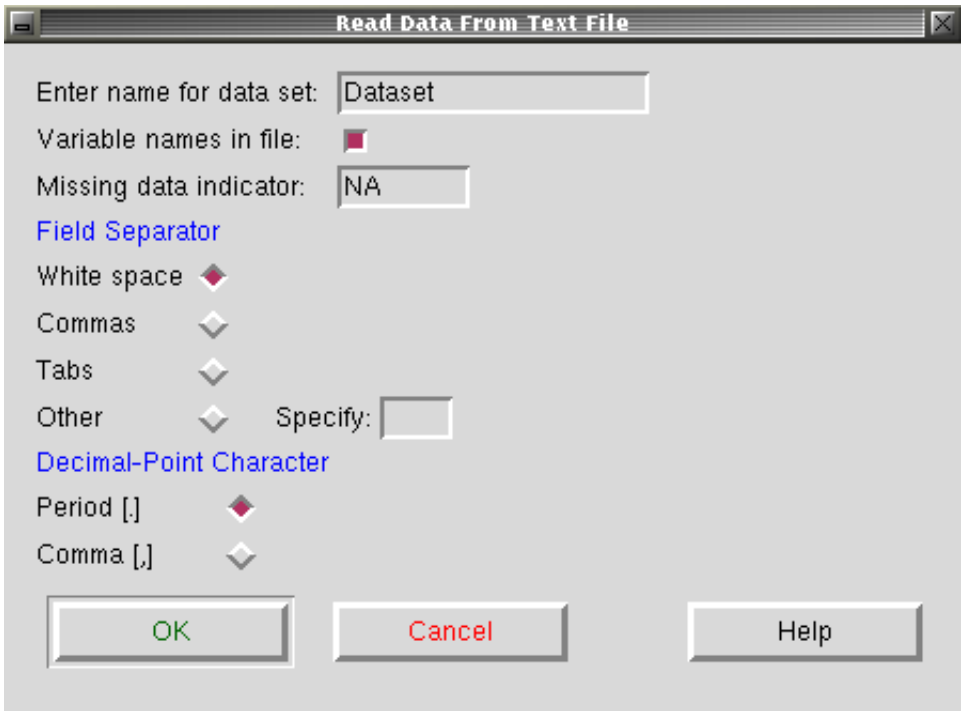


2.1.2. Importar datos de un fichero externo

El fichero externo puede contener datos en el formato "binario" (código objeto) nativo de R (ficheros con extensión `Rdata` o `Rda`), en texto puro (ASCII), o en alguno de los formatos binarios nativos de otros programas (SPSS, Excel...).

En el primer caso (formato nativo de R) úsese la opción *Cargar datos...* del Menú *Datos*. En los otros casos ha de recurrirse al menú *Datos / Importar datos*.

Los ficheros de texto (columnas de números) representan la forma más universal para intercambio de datos. Para importar datos de texto se elegirá la opción *desde un fichero de texto*, que abrirá el cuadro de diálogo *Leer datos de archivo de texto*.



Es necesario indicar:

Introducir nombre de datos: Para el conjunto construido a partir de los datos del fichero.

Nombres de las variables en el fichero: Si el fichero contiene los nombres de las variables en la primera fila.

Indicador de datos ausentes: Cómo se indica si un campo no contiene valores, esto es, que se considera un valor ausente. Por omisión, el indicador es `NA` (*not available*, no disponible). Puede dejarse así a menudo, pues si un campo de una variable numérica está vacío, también se considera ausente.

Separador de campos: Indique el carácter que separa los campos, bien espacio en blanco, comas, tabuladores, o cualquier otro carácter que se puede especificar.

Carácter decimal: Si se utiliza punto o coma para separar los decimales de la parte entera.

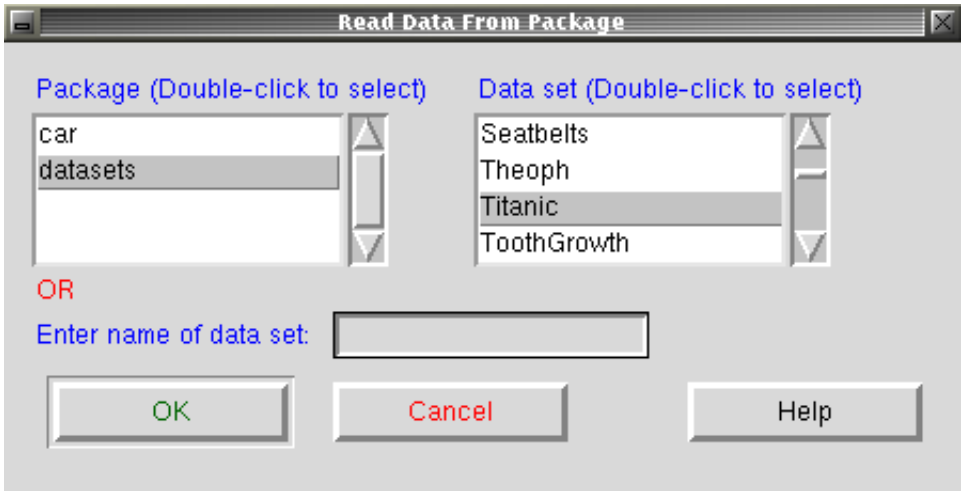
En el caso de ficheros binarios, se pueden abrir ficheros guardados desde otros programas estadísticos, como SPSS y Minitab. Así, para abrir un fichero SPSS elegimos *desde datos SPSS* en el menú *Datos / Importar datos*.

2.1.3. Utilizar datos incluidos en R

R incluye en su distribución una colección importante de datos de todo tipo. Para ver una descripción sucinta de los datos disponibles, elija la opción *Listar datos en paquetes* del menú *Datos en paquetes*.

```
Data sets in package 'car':

Adler                   Experimenter Expectations
Angell                  Moral Integration of American Cities
Anscombe                U. S. State Public-School Expenditures
Baumann                 Methods of Teaching Reading Comprehension
Bfox                    Canadian Women's Labour-Force Participation
Blackmoor               Exercise Histories of Eating-Disordered and
                        Control Subjects
Burt                    Fraudulent Data on IQs of Twins Raised Apart
Can.pop                 Canadian Population Data
Chile                   Voting Intentions in the 1988 Chilean Plebiscite
Chirot                  The 1907 Romanian Peasant Rebellion
Cowles                  Cowles and Davis's Data on Volunteering
Davis                   Self-Reports of Height and Weight
DavisThin               Davis's Data on Drive for Thinness
Duncan                  Duncan's Occupational Prestige Data
Ericksen                The 1980 U.S. Census Undercount
Florida                 Florida County Voting
Freedman                Crowding and Crime in U. S. Metropolitan Areas
Friendly                Format Effects on Recall
Ginzberg                Data on Depression
Greene                  Refugee Appeals
Guyer                   Anonymity and Cooperation
```

Si alguno resulta de interés, escoja, en el mismo menú, la opción *Leer datos de paquete adjunto*. Indique el paquete y el conjunto de datos buscado, que se convertirá en el conjunto de datos activo.



2.2. Trasformaciones de las variables

Vamos a utilizar diferentes opciones del menú *Datos / Modificar variables de los datos activos*.

Considérese el siguiente conjunto de datos `servidor.http`:

servidor.http

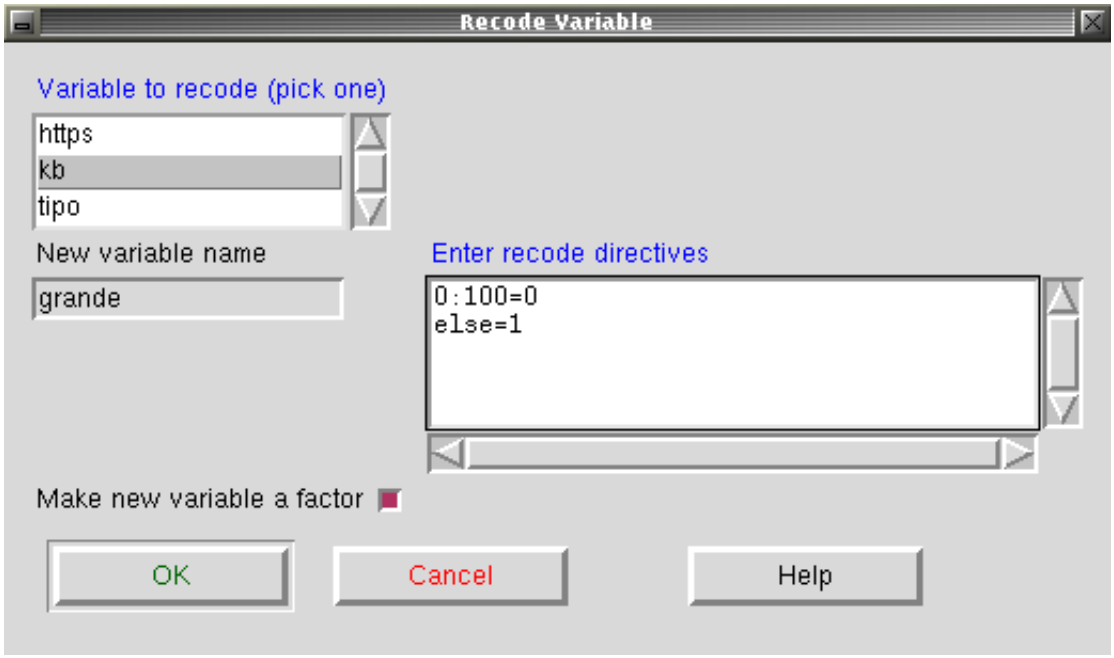
	tipo	kb	https
1	HTML	38	0
2	PNG	15	0
3	PNG	72	1
4	HTML	113	1
5	SVG	35	1
6	PNG	221	0
7	PNG	98	1
8	HTML	284	0
9	SVG	320	0
10	HTML	52	0
11	OGG	875	1

La variable `tipo` es un factor, y las variables `kb` y `https` son numéricas. Sin embargo, los valores de `https` son binarios, donde 1 indica **sí** y 0 indica **no** (conexión segura o insegura, respectivamente).

Vamos a considerar las siguientes trasformaciones:

2.2.1. Recodificar

Sirve para trasformar una variable cualquiera en una variable de tipo factor (de caracteres). Supongamos que queremos crear una variable binaria que tome el valor 1 si un fichero es mayor de 100 `kb` (se considera *grande*) y 0 si no. El cuadro de diálogo para *Recodificar variable* sería:



En la ventana de mensajes aparece `NOTE: The dataset servidor.http has 11 rows and 4 columns.` lo que indica que se ha añadido una columna.

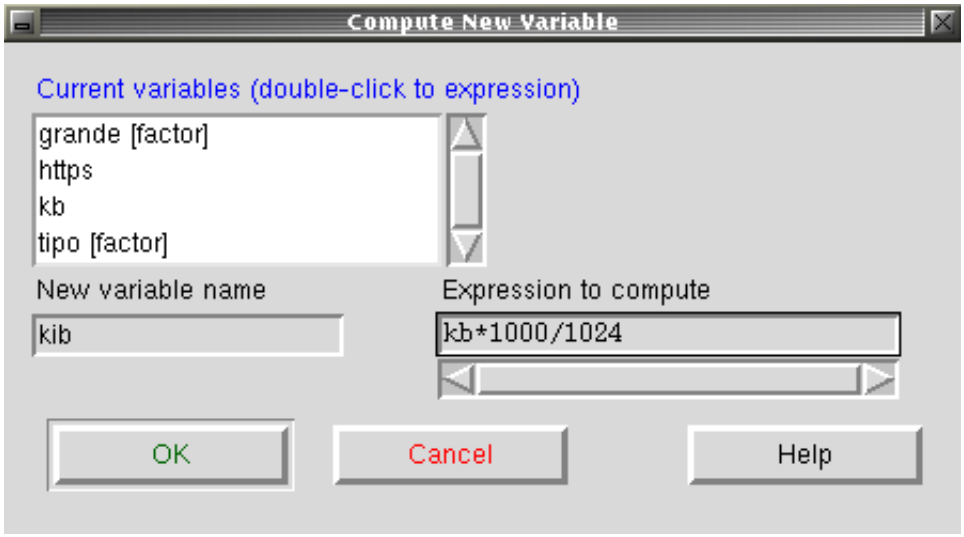
servidor.http

	tipo	kb	https	grande
1	HTML	38	0	0
2	PNG	15	0	0
3	PNG	72	1	0
4	HTML	113	1	1
5	SVG	35	1	0
6	PNG	221	0	1
7	PNG	98	1	0
8	HTML	284	0	1
9	SVG	320	0	1
10	HTML	52	0	0
11	OGG	875	1	1

2.2.2. Calcular

Aquí podemos definir una nueva variable (o sobrescribir una antigua) mediante una expresión arbitraria. Supongamos que queremos obtener la variable `kib` a partir de `kb`[1]. La opción correspondiente es *Calcular una nueva variable*.

[1]El prefijo *k* indica *kilo*, es decir, múltiplo 1000; el prefijo *ki* indica *kibi*, es decir, múltiplo $2^{10} = 1024$. Véase `http://es.wikipedia.org/wiki/Prefijos_binarios`.



Se añade una nueva columna a `servidor.http`:

servidor.http

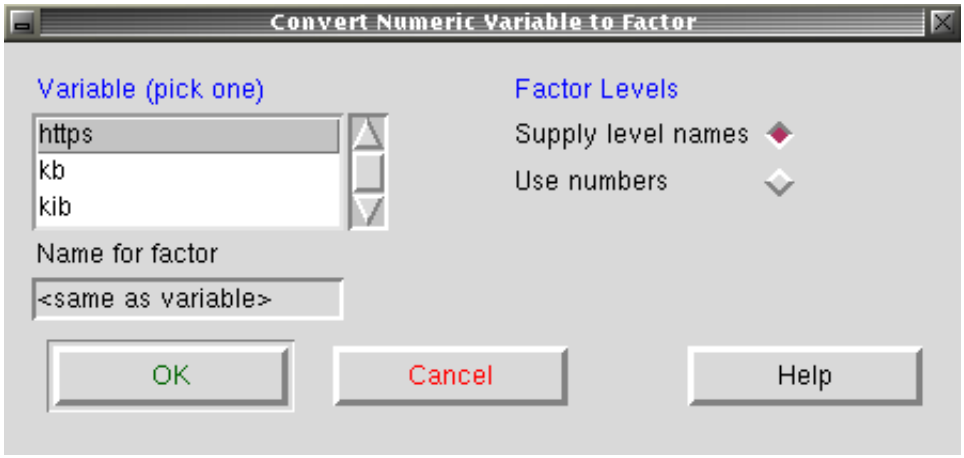
	tipo	kb	https	grande	kib
1	HTML	38	0	0	37.10938
2	PNG	15	0	0	14.64844
3	PNG	72	1	0	70.31250
4	HTML	113	1	1	110.35156
5	SVG	35	1	0	34.17969
6	PNG	221	0	1	215.82031
7	PNG	98	1	0	95.70312
8	HTML	284	0	1	277.34375
9	SVG	320	0	1	312.50000
10	HTML	52	0	0	50.78125
11	OGG	875	1	1	854.49219

2.2.3. Convertir a factor

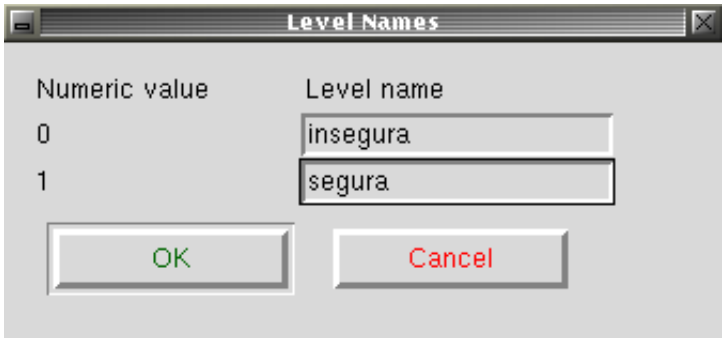
Cuando una variable numérica representa, en realidad, a una variable cualitativa en que los números son códigos correspondientes a las modalidades, es necesario indicarlo al programa (ya que éste no puede discernir si los números son cantidades o códigos).

Como se puede comprobar en las capturas de la sección anterior, `grande` se considera factor[2] mientras que `https` no. Vamos a utilizar la opción *Convertir variable numérica en factor*.

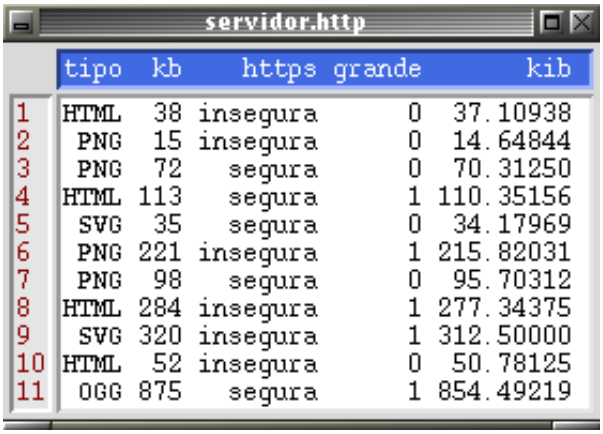
[2]Ya que fue creada a partir de una recodificación, y por omisión el resultado es un factor.



Se ofrece la posibilidad de dar nombre a las modalidades (*Asignar nombres a los niveles*).



En este caso hemos sustituido la definición anterior de `https` por la nueva. Podíamos haber elegido crear una nueva variable.



	tipo	kb	https	grande	kib
1	HTML	38	insegura	0	37.10938
2	PNG	15	insegura	0	14.64844
3	PNG	72	segura	0	70.31250
4	HTML	113	segura	1	110.35156
5	SVG	35	segura	0	34.17969
6	PNG	221	insegura	1	215.82031
7	PNG	98	segura	0	95.70312
8	HTML	284	insegura	1	277.34375
9	SVG	320	insegura	1	312.50000
10	HTML	52	insegura	0	50.78125
11	OGG	875	segura	1	854.49219

2.2.4. Agrupar

Mediante la recodificación habíamos visto cómo agrupar una variable numérica en intervalos. La opción `Segmentar variable numérica` permite agrupar en intervalos de forma cómoda, si nos conformamos con obtener intervalos de alguna de las tres formas siguientes:

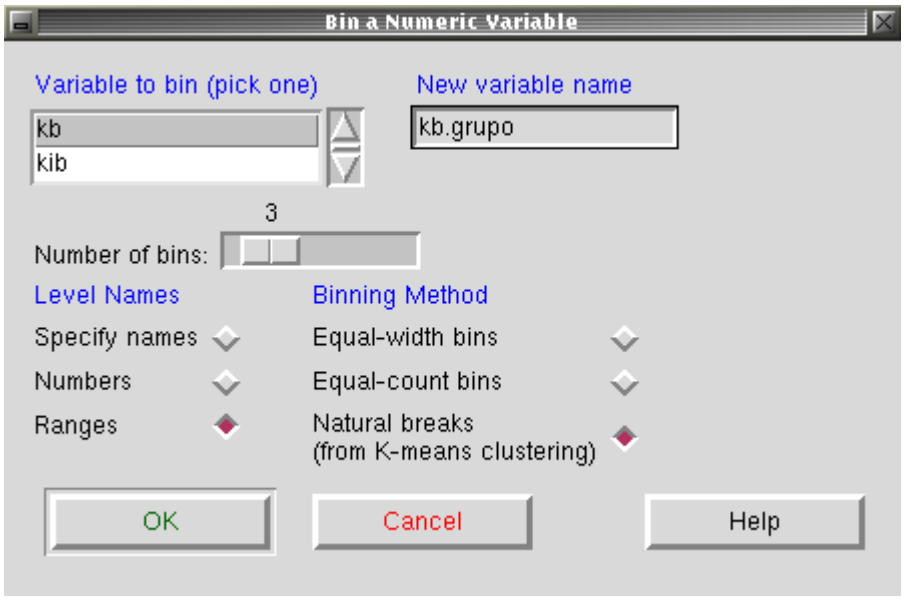
Segmentos equidistantes: Intervalos de igual amplitud.

Segmentos de igual cantidad: Intervalos de igual frecuencia.

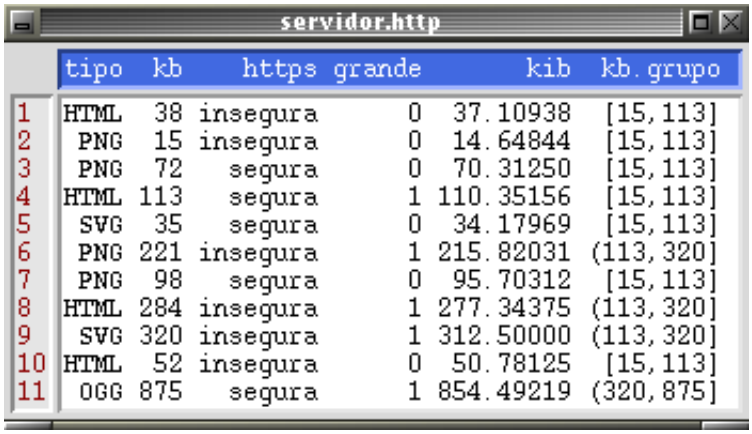
Segmentos naturales: Se aplica un algoritmo de agrupación automática (k medias) para obtener los intervalos.

Supongamos que queremos obtener 3 grupos *naturales* de casos según la variable `kb`.

Los nombres de los grupos se pueden especificar, o crearse automáticamente como números o como rangos.



Obtenemos un nuevo factor.



	tipo	kb	https	grande	kib	kb.grupo
1	HTML	38	insegura	0	37.10938	[15,113]
2	PNG	15	insegura	0	14.64844	[15,113]
3	PNG	72	segura	0	70.31250	[15,113]
4	HTML	113	segura	1	110.35156	[15,113]
5	SVG	35	segura	0	34.17969	[15,113]
6	PNG	221	insegura	1	215.82031	(113,320]
7	PNG	98	segura	0	95.70312	[15,113]
8	HTML	284	insegura	1	277.34375	(113,320]
9	SVG	320	insegura	1	312.50000	(113,320]
10	HTML	52	insegura	0	50.78125	[15,113]
11	OGG	875	segura	1	854.49219	(320,875]

2.2.5. Renombrar

La opción *Renombrar variables* permite cambiar el nombre a una o varias de las variables del conjunto de datos activo.

2.2.6. Eliminar

La opción *Eliminar variables de los datos* permite eliminar una o varias variables del conjunto de datos activo.

Tema 3

Análisis descriptivo con una variable

Se pueden obtener resultados numéricos o representaciones gráficas.

3.1. Análisis numéricos

Los diferentes tipos de análisis numéricos para una variable se albergan bajo la opción *Resúmenes* del menú *Estadísticas*. Las posibilidades son:

3.1.1. Resumen rápido

Pulsando *Datos activos*, se muestra una descripción de todas las variables contenidas en un conjunto de datos (figura 3.1). Para las variables cuantitativas, se indica: máximo, mínimo, cuartiles y media. Para las variables cuantitativas, se da la frecuencia absoluta de las modalidades más frecuentes (y la de los valores ausentes, si hay alguno).

Si hay más de diez variables en el conjunto de datos, R pide confirmación, pues la abundancia de información puede resultar incómoda.

3.1.2. Resúmenes numéricos

En *Resúmenes numéricos* podemos obtener los valores de la media (*mean*), desviación típica (*standard deviation*) y cuantiles (*quantiles*) arbitrarios para una variable cuantitativa (figura 3.2).

Conviene resaltar que R utiliza la cuasivarianza, es decir, cuando se le pide que calcule la varianza y la desviación típica, lo que da exactamente es el resultado de las fórmulas:

$$\hat{s}^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \qquad \hat{s} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

Para calcular otros descriptivos ha de recurrirse a la ventana de instrucciones (figura 3.3). En primer lugar, fíjese en que al efectuar el cálculo de la media aparecía en la ventana de instrucciones la orden

```
mean(servidor.http$kb, na.rm=TRUE)
```

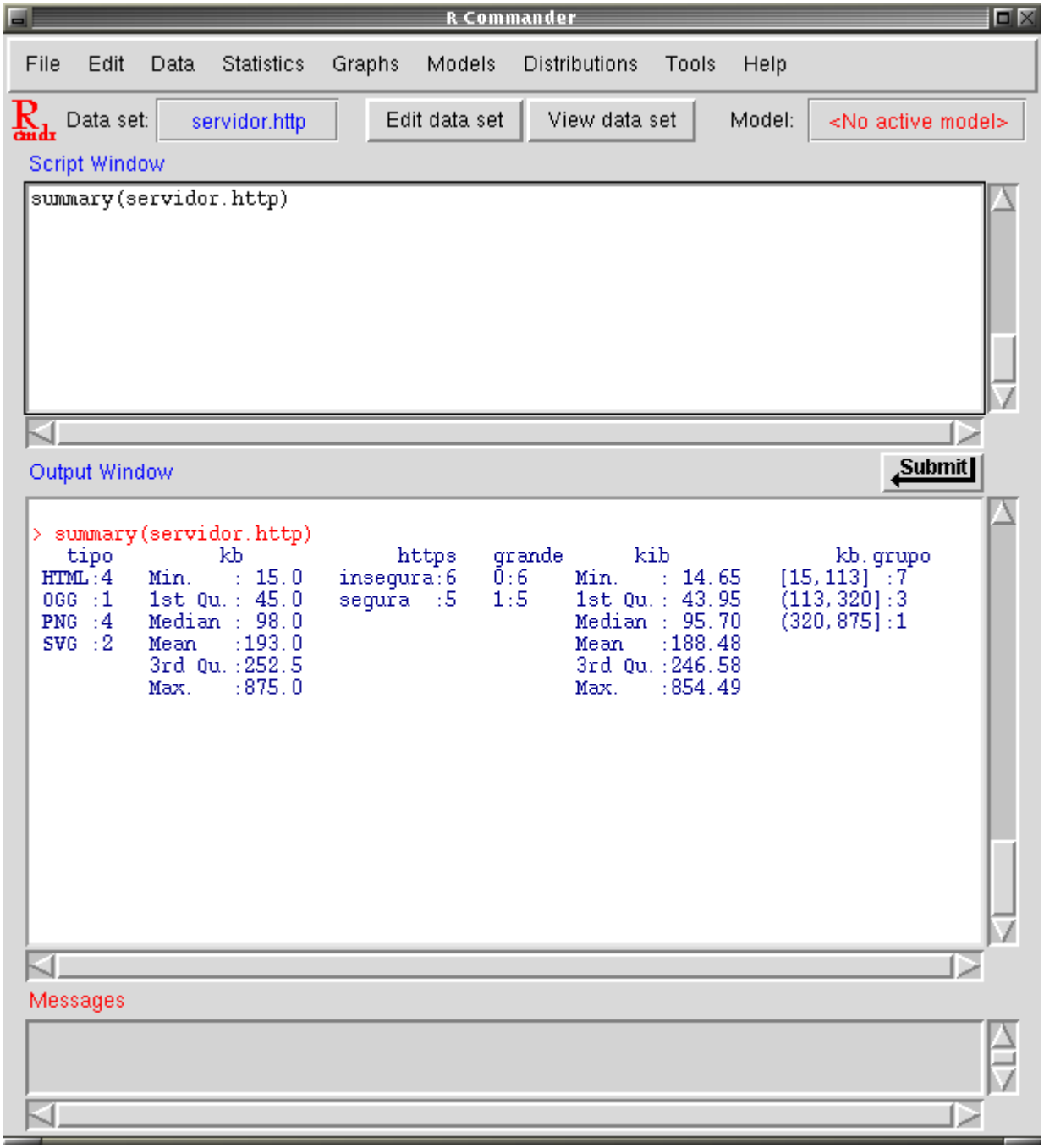


Figura 3.1: Resumen de variables cuantitativas y cualitativas.

Veamos cómo obtener el descriptivo deseado sustituyendo la orden `mean` por la que corresponda en su lugar[1]. Para ejecutar la orden, ha de seleccionar la línea completa y pulsar *Submit*.

Mediana: Utilice un cuantil de orden 0,5, como se vio arriba, o bien la orden `median`:

```
median (servidor.http$kb)
```

[1]La parte "`, na.rm=TRUE`" es opcional y se utiliza para obtener un valor numérico (en lugar de un valor ausente, NA) cuando los datos contienen valores ausentes.

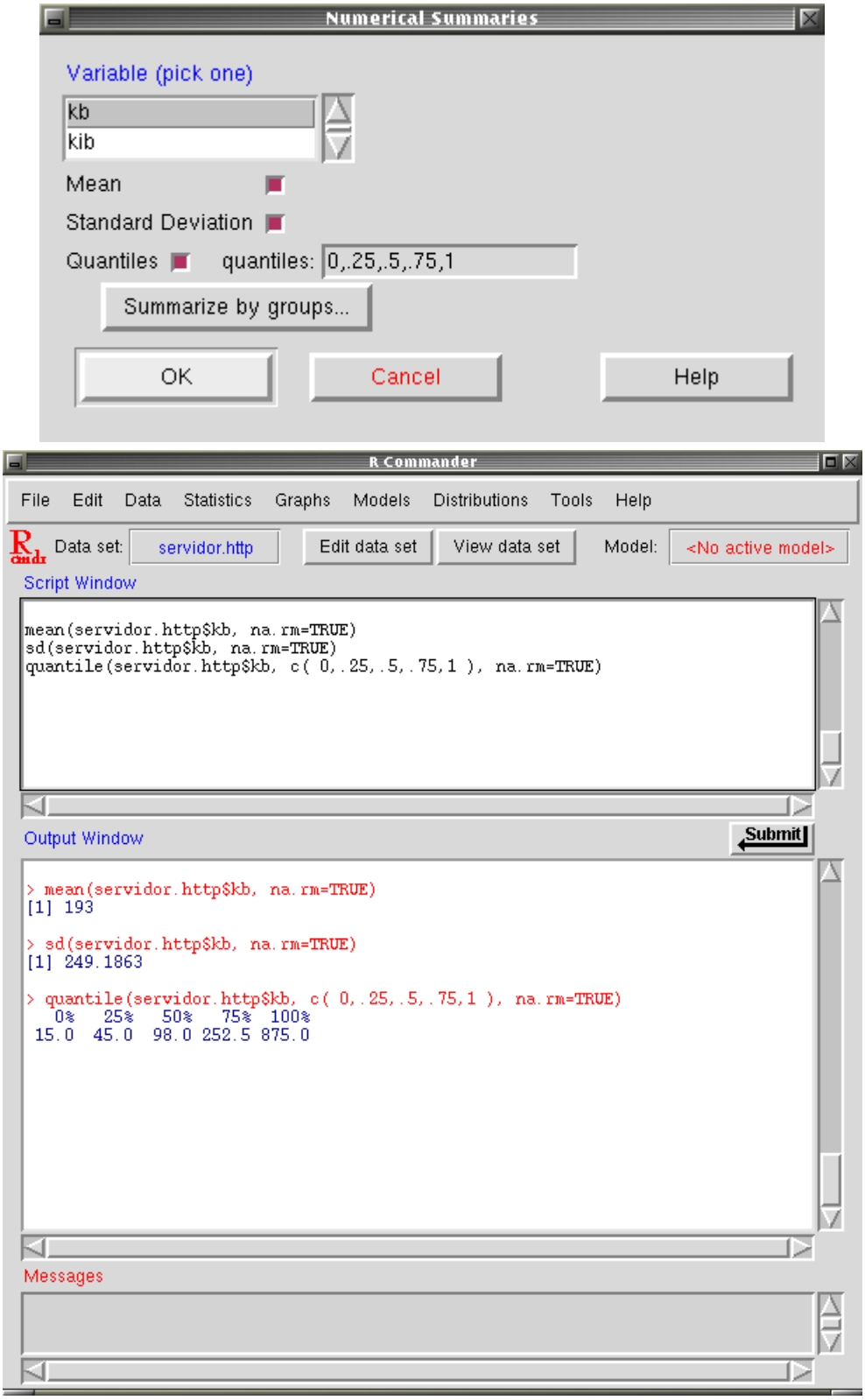


Figura 3.2: Resúmenes numéricos

Amplitud o recorrido: Aquí es necesario combinar dos órdenes:

```
diff (range (servidor.http$kb))
```

Si la va a utilizar varias veces, es mejor definir una función `amplitud`:

```
amplitud <- function (x) diff (range (x))
amplitud (servidor.http$kb)
```

Recorrido intercuartílico: Utilice la orden `IQR`:

```
IQR (servidor.http$kb)
```

Coeficiente de variación: Definamos la función `CV`, bien en la forma más simple,

```
CV <- function (x) sd (x) / mean (x)
```

o, para obtener un valor numérico incluso en datos con valores ausentes,

```
CV <- function (x) sd (x, na.rm=TRUE) / mean (x, na.rm=TRUE)
```

Simetría: Utilice `skewness` así[2]:

```
skewness (servidor.http$kb)
```

Curtosis: La orden correspondiente es `kurtosis`[3]:

```
kurtosis (servidor.http$kb)
```

[2]Puede ser necesario que antes cargue el paquete `fBasics`, eligiéndolo en el menu *Herramientas / Cargar paquete(s)*.
[3]Puede ser necesario que antes cargue el paquete `fBasics`.

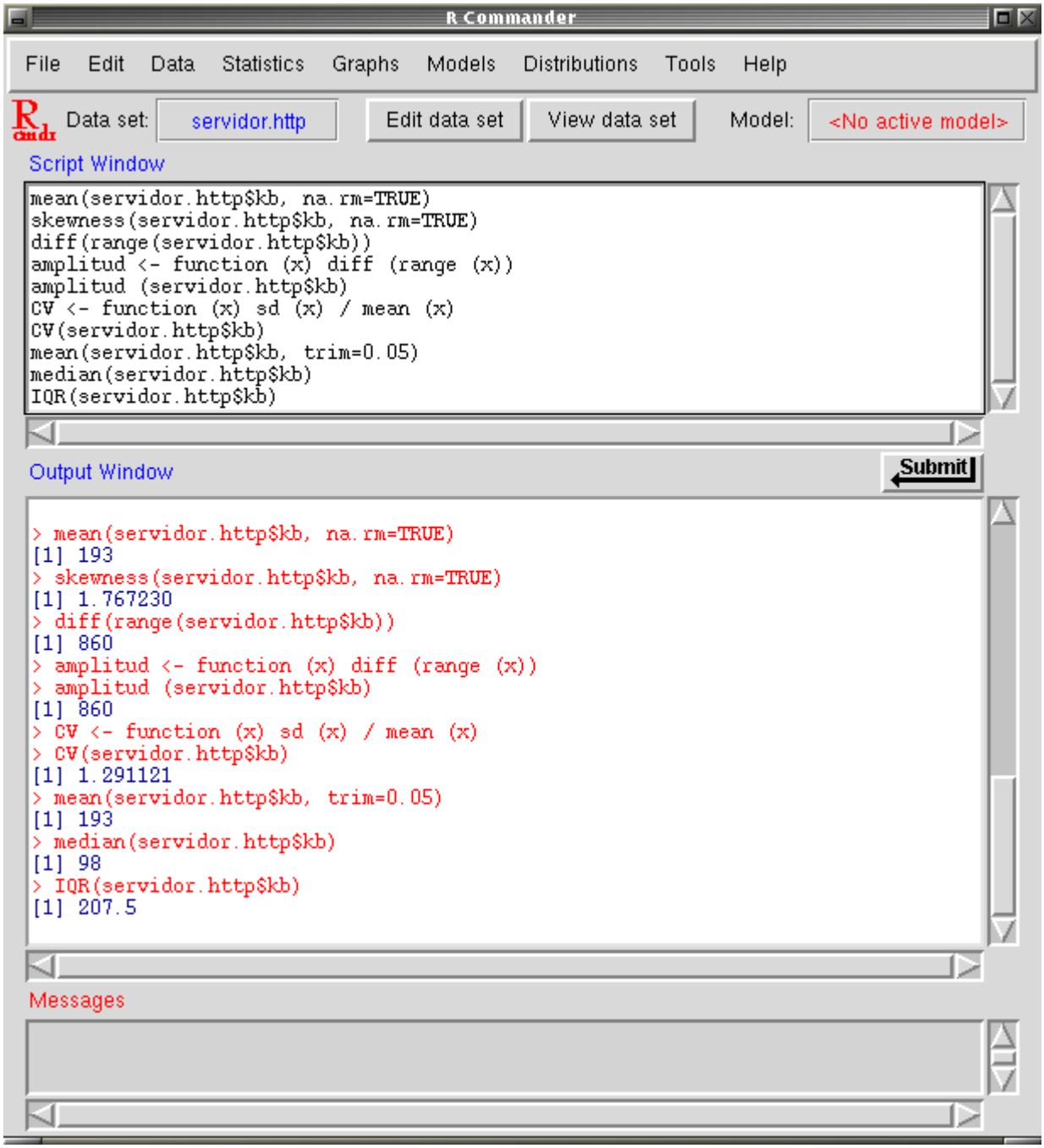
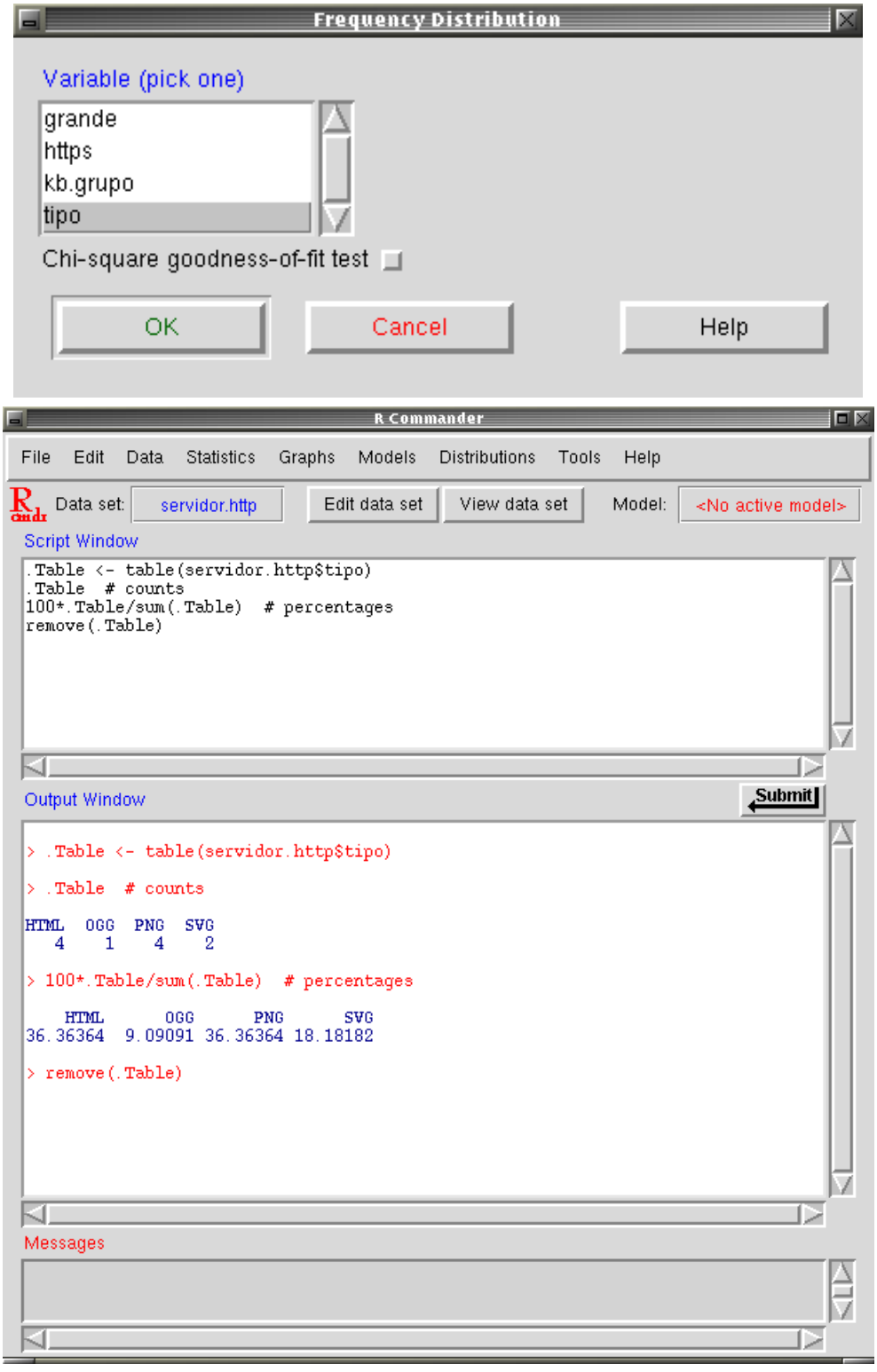


Figura 3.3: Cálculo de descriptivos no incluidos en los menús.

3.1.3. Distribuciones de frecuencias

Para las variables cualitativas, puede confeccionarse con el procedimiento *Estadísticas / Resúmenes / Distribución de frecuencias* una tabla donde aparezcan los valores de la variable, sus frecuencias absolutas y las frecuencias relativas en forma de porcentajes.



3.2. Representaciones gráficas

Las representaciones gráficas permiten captar rápidamente y sin gran esfuerzo las principales características de una distribución de frecuencias. Son un medio complementario, aunque muy

importante, para realizar un análisis estadístico de los datos.

Están recogidas bajo el menú *Gráficas*. Describimos sólo las opciones de interés en nuestro curso.

Si la orden ejecutada proporciona una salida gráfica, R abre una nueva ventana (*device*) que contiene el gráfico. Éste puede ser grabado en un fichero mediante la opción *Guardar gráfica* del menú *Gráficas*.

3.2.1. Gráfico de sectores

Bajo la opción *Gráfica de sectores*. Para representar variables cualitativas. Véase la ilustración 3.4.

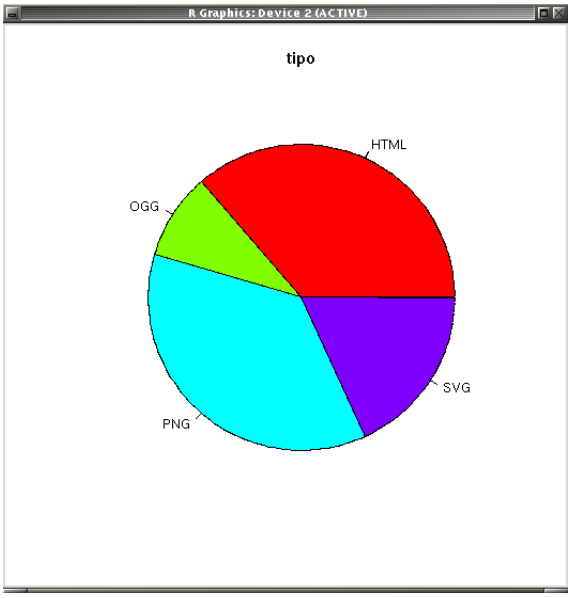


Figura 3.4: Gráfica de sectores.

3.2.2. Gráfico de barras

Bajo la opción *Gráfica de barras*. Para representar variables cualitativas. Véase la ilustración 3.5.

3.2.3. Histograma

Para representar la distribución de una variable cuantitativa, se puede recurrir a la opción *Histograma*. Es posible pedir el número aproximado de barras o dejar la elección a un algoritmo automático. Véase la ilustración 3.7 en la página 37.

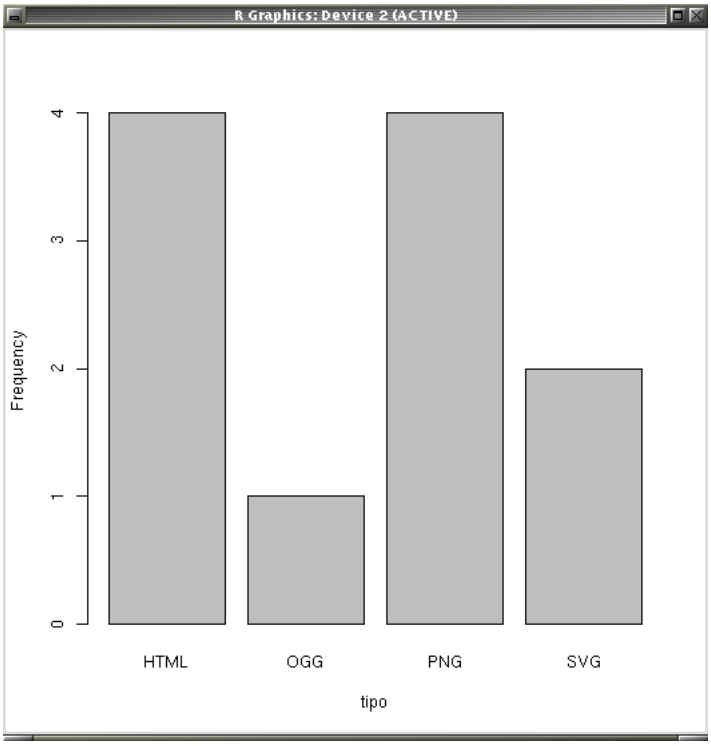


Figura 3.5: Gráfica de barras.

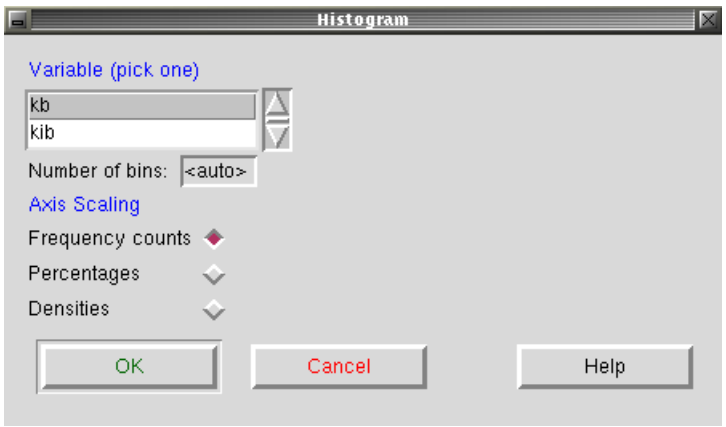


Figura 3.6: Cuadro de diálogo para construir un Histograma.

3.2.4. Tallo y hojas

Se obtienen con la opción *Gráfica de tallos y hojas*. Estos gráficos se representan mediante caracteres, y se utilizan para describir variables cuantitativas y permite visualizar globalmente la distribución manteniendo la individualidad de los datos. Tienen una gran similitud con los histogramas pero representan directamente los dígitos de los valores observados en vez de barras o rectángulos, por lo que ofrecen mayor cantidad de información.

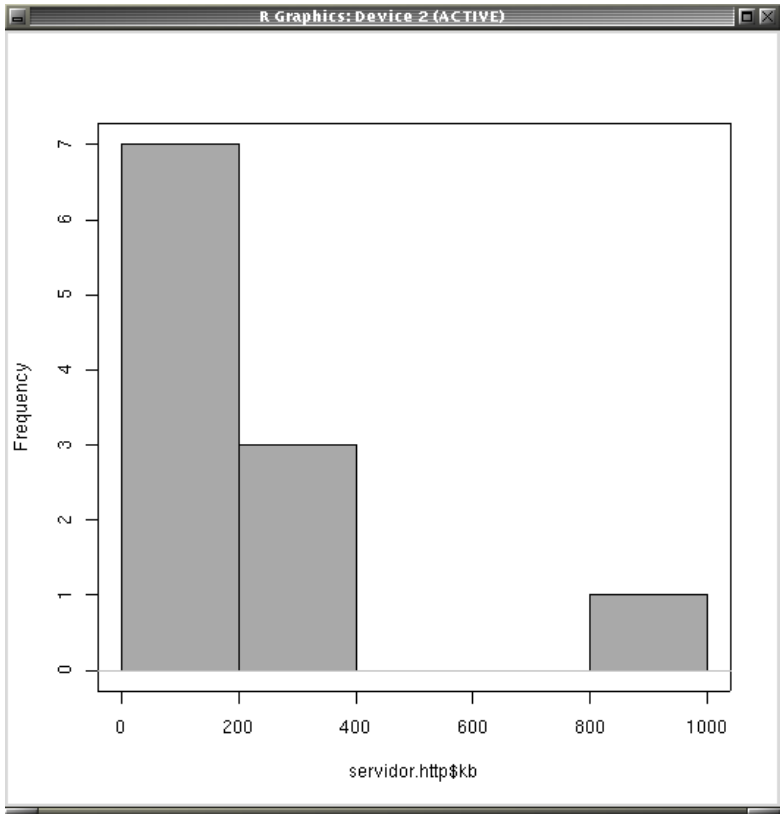


Figura 3.7: Histograma.

Expondremos a continuación un par de ejemplo. El primero corresponde a una muestra pequeña (once observaciones):

```
> stem.leaf(servidor.http$kb)
1 | 2: represents 120
 leaf unit: 10
        n: 11

   3    0* | 133
  (3)   0. | 579
   5    1* | 1
        1. |
   4    2* | 2
   3    2. | 8
   2    3* | 2
HI: 875
```

Para obtenerlo, se ha utilizado Rcommander, cuyo cuadro de diálogo para obtener un diagrama de tallo y hojas se muestra en la ilustración 3.8 en la página 38.

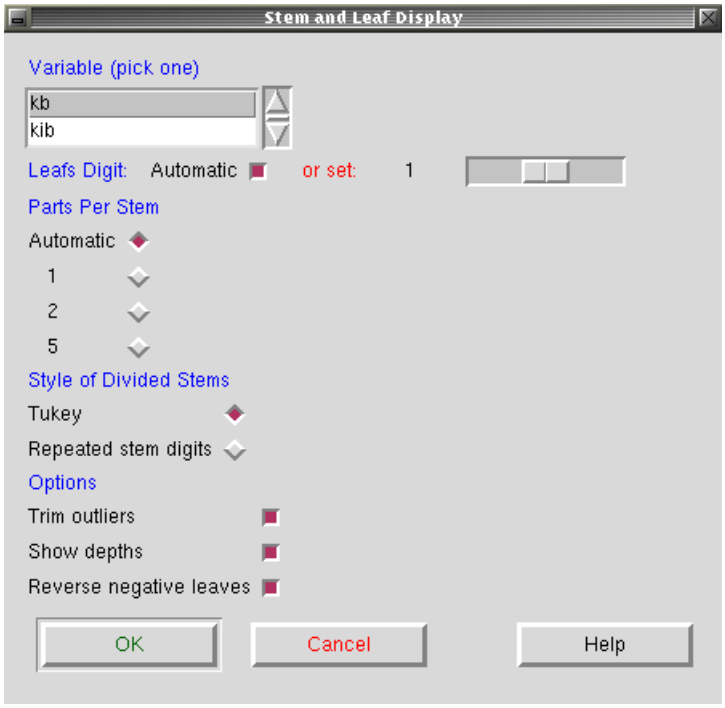


Figura 3.8: Diagrama de tallo y hojas.

Veamos otro ejemplo con una muestra más grande (tamaño 97):

```
> stem.leaf(MisDatos$MiVar, unit=1)
1 | 2: represents 12
 leaf unit: 1
         n: 97

    1     -3. | 7
    2     -3* | 0
    8     -2. | 877665
   22     -2* | 44443322222100
   27     -1. | 99976
   38     -1* | 33221110000
  (16)    -0. | 9998876666665555
   43     -0* | 444443322111
   31      0* | 00023444
   23      0. | 5566678999
   13      1* | 1234444
    6      1. | 59
    4      2* | 02
    2      2. | 5
HI: 42
```

Para interpretar las gráficas de tallo y hojas, téngase en cuenta que:

- El *tallo* representa la cifra más significativa del valor de cada valor de la muestra.
- Cada hoja representa, mediante la siguiente cifra significativa, el valor correspondiente a una sola observación.
- Se indica la "escala" mediante la indicación *leaf unit* (`1` indica unidades, `10` decenas...), así como mediante la explicación de en qué posición (magnitud) se encuentra la barra vertical: `1 | 2: represents 12`.
- Los valores atípicos se indican por `HI` (si son altos) o por `LO` (si son bajos).
- Los números de la columna a la izquierda indican frecuencias absolutas acumuladas hacia los extremos, salvo el número entre paréntesis, que indica la frecuencia absoluta de la barra en que está situada la mediana.

3.2.5. Gráfico de cajas

Se obtienen con la opción *Diagrama de caja*. Véase la ilustración 3.9 en la página 39.

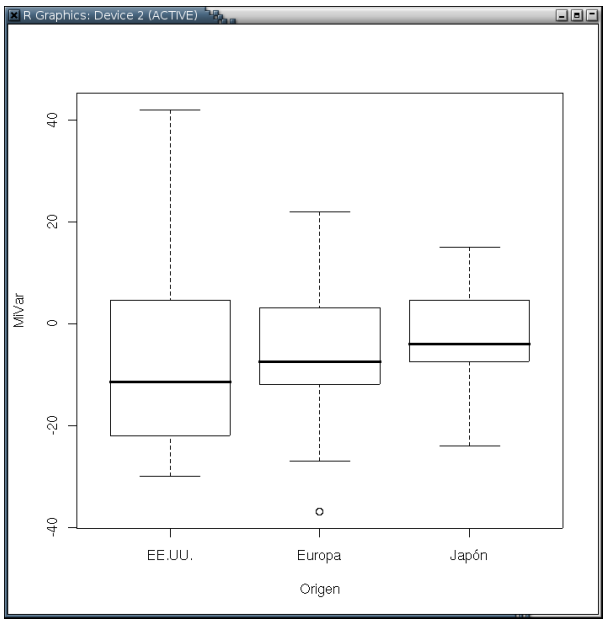


Figura 3.9: Diagrama de cajas.

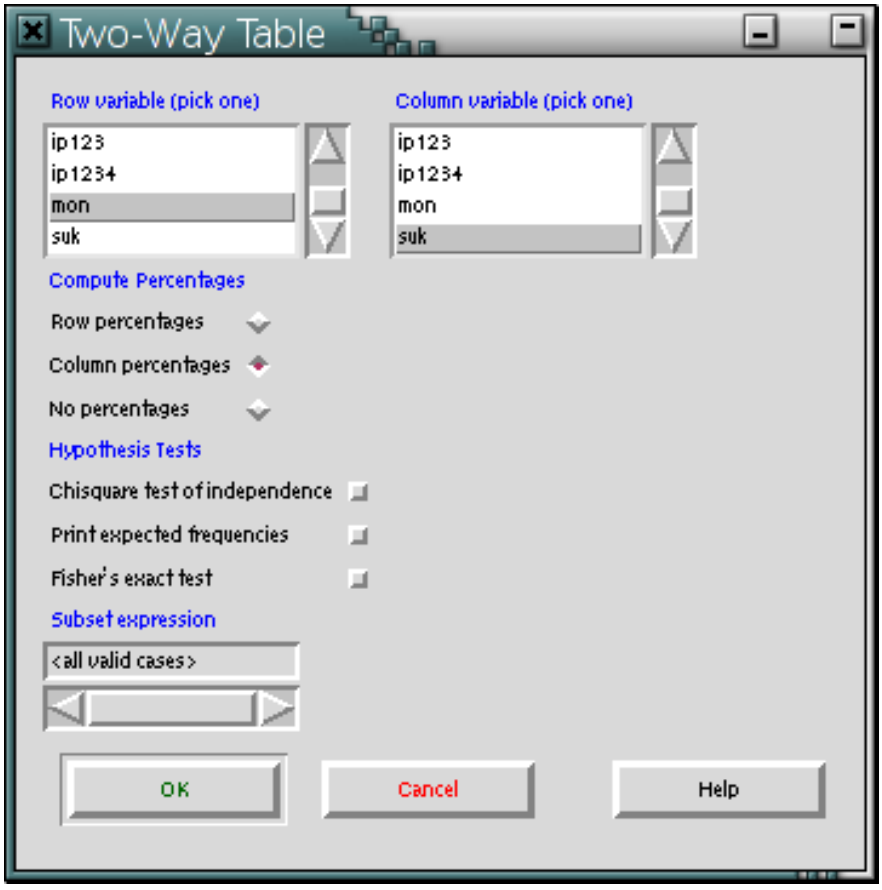
Tema 4

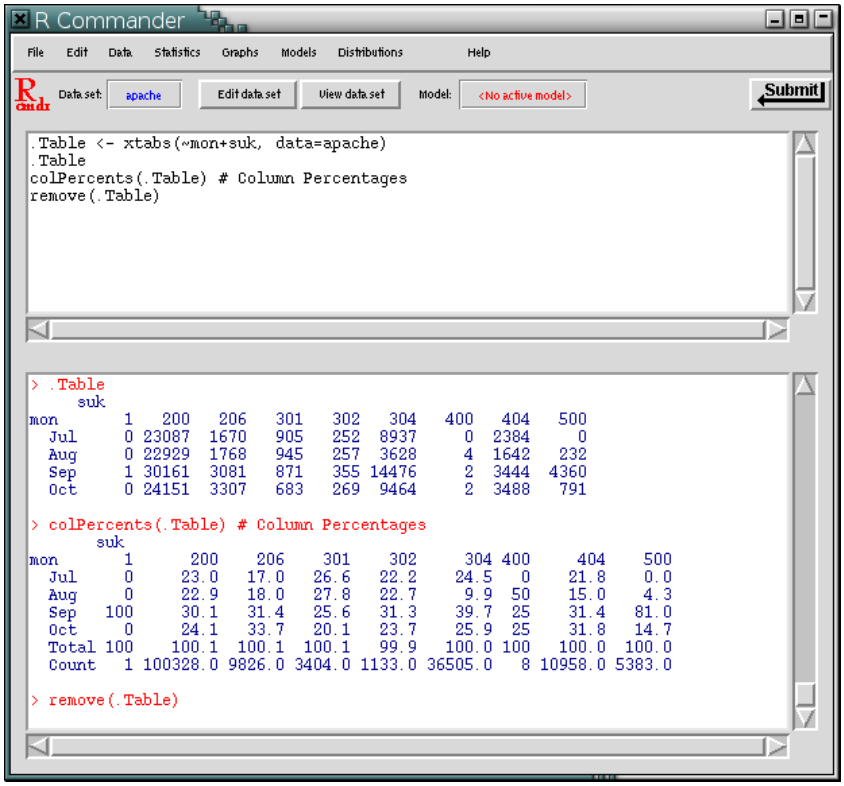
Análisis descriptivo con dos variables

4.1. Dos variables cualitativas

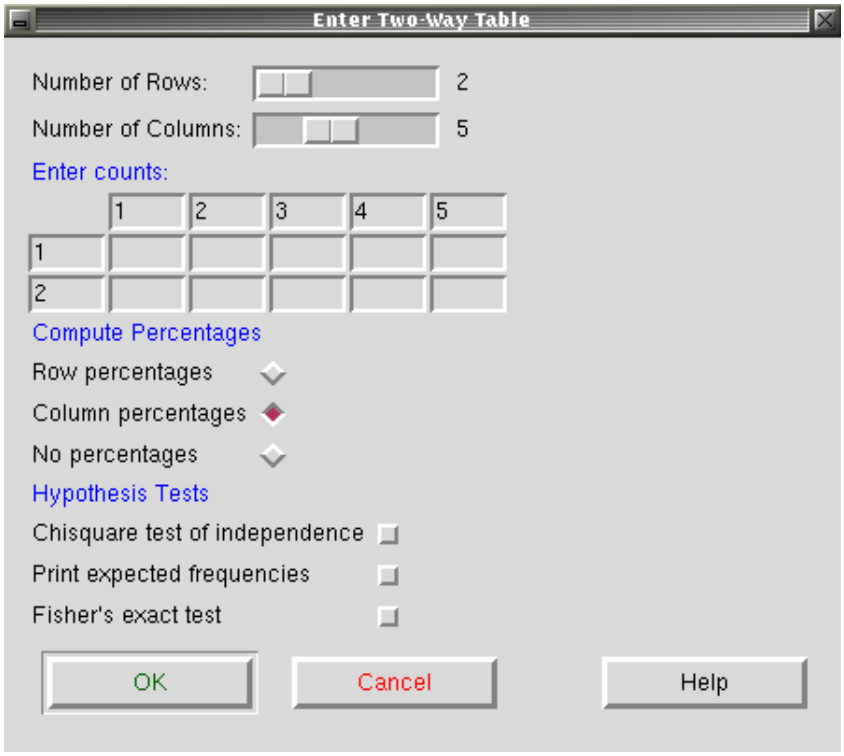
La relación entre dos variables cualitativas se estudia a través de la distribución conjunta de las mismas.

Dado un conjunto de datos que contenga más de una variable cualitativa, elija *Estadísticos / Tablas de contingencia / Tabla de doble entrada* para hallar la distribución conjunta de dos de ellas. El cuadro de diálogo permite calcular las frecuencias marginales por filas o columnas.

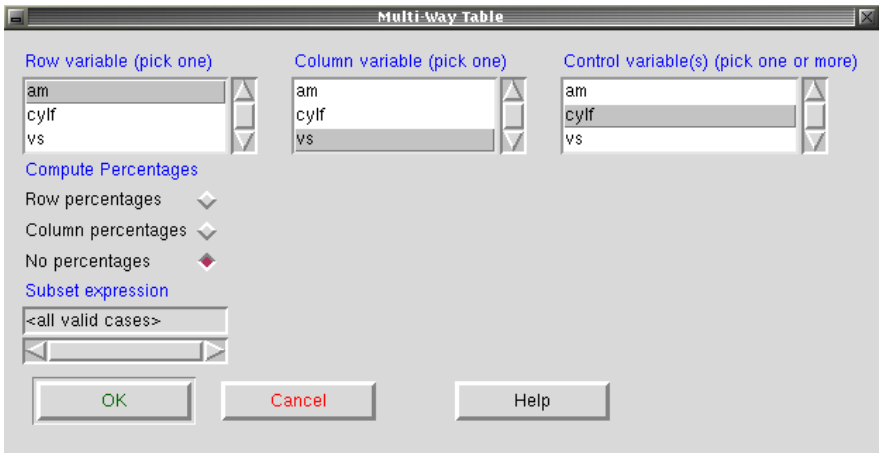




Téngase en cuenta que, si se quiere introducir directamente una tabla de frecuencias en R, se ha de emplear la opción *Estadísticos / Tablas de contingencia / Introducir y analizar una tabla de doble entrada.*



Si se trata de estudiar la distribución conjunta de más de dos variables, utilice *Estadísticos / Tablas de contingencia / Tabla de entradas múltiples.*

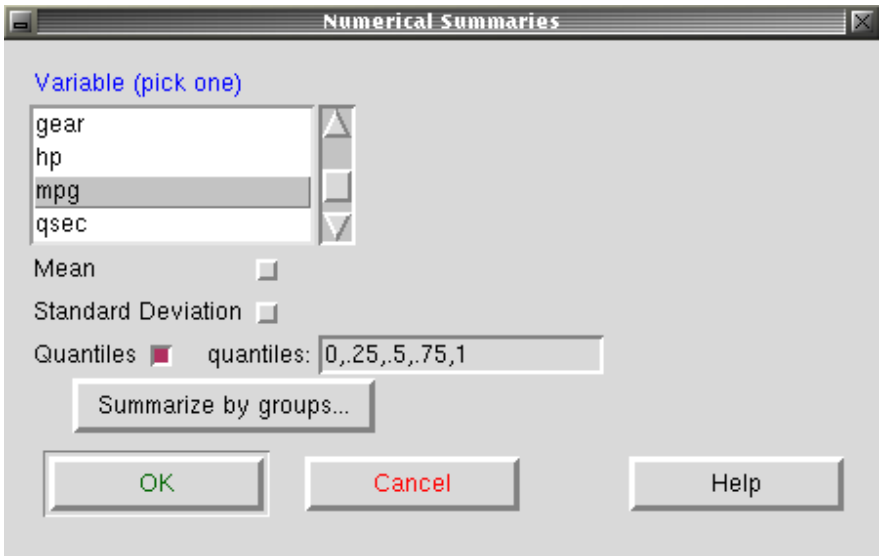


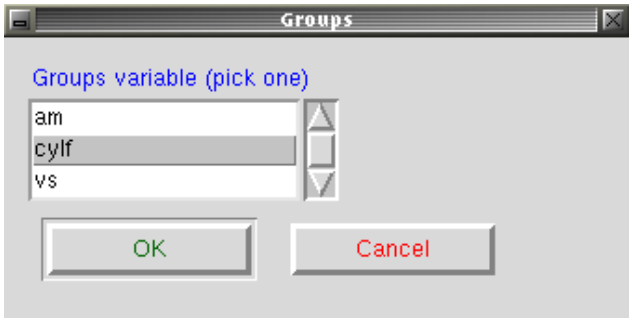
La salida presenta una tabla de frecuencias por cada combinación de las restantes variables:

```
, , cyl = 4

   vs
am  0 1
  0 0 3
  1 1 7

, , cyl = 6

   vs
am  0 1
  0 0 4
  1 3 0

, , cyl = 8

   vs
am  0 1
  0 12 0
  1 2 0
```

4.2. Variable cuantitativa frente a variable cualitativa

A menudo interesa analizar el comportamiento de una variable cuantitativa según los niveles de un factor (por ejemplo, por sexos). En los cuadros de diálogo que lo permiten, existe un botón para los cálculos por grupos (*Resumir por grupos*). Por ejemplo, en el cálculo de descriptivos:

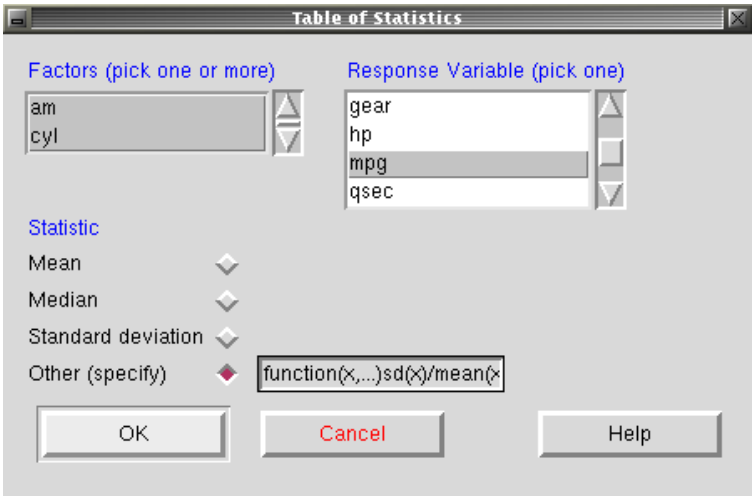




En el ejemplo, se solicitan cuartiles para cada nivel de la variable `cyl`. Los niveles de `cyl` son 4, 6 y 8:

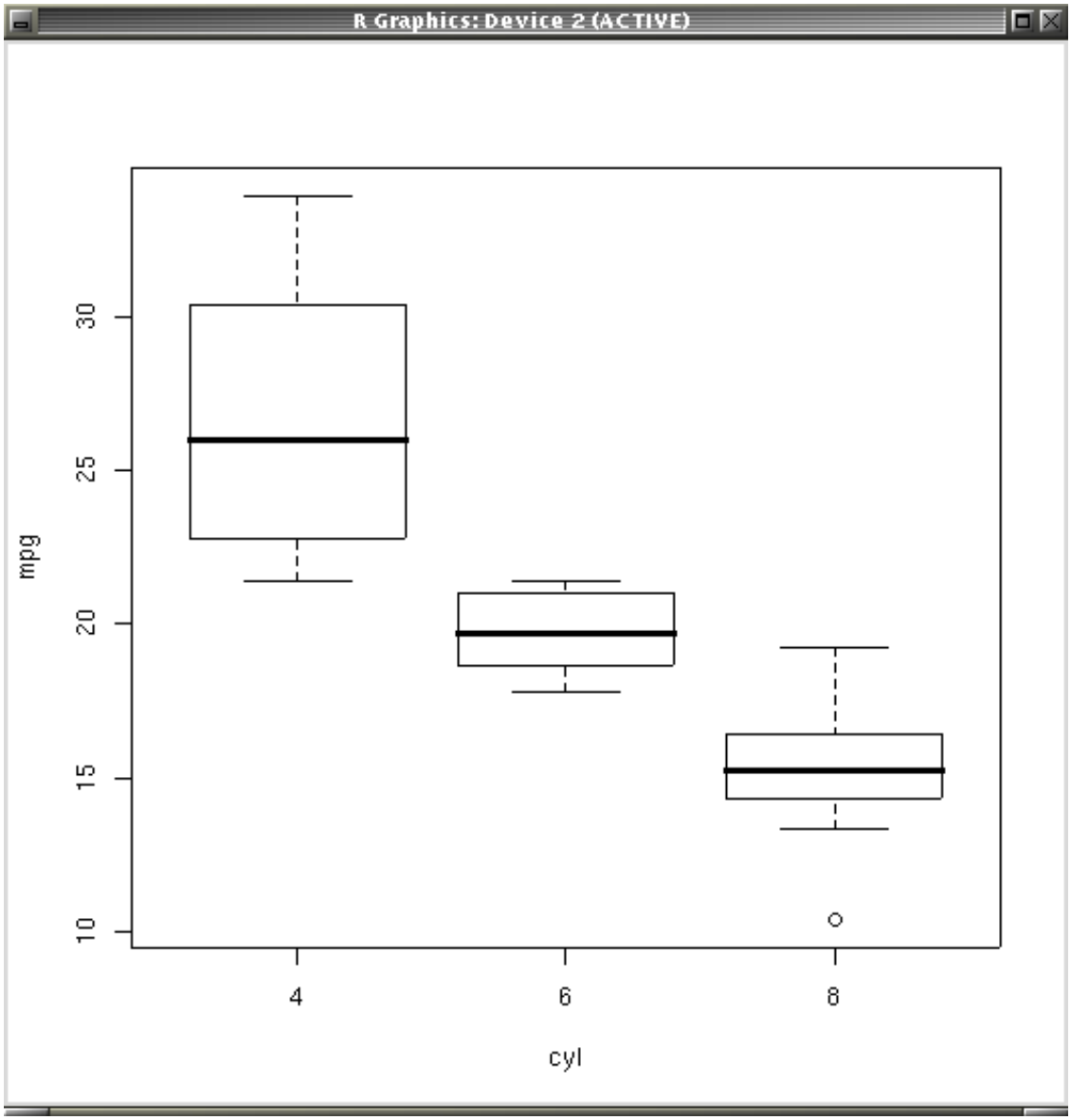
```
    0%    25%    50%    75%   100%   n
4  21.4  22.80  26.0  30.40  33.9  11

6  17.8  18.65  19.7  21.00 21.4   7

8  10.4  14.40  15.2  16.25  19.2  14
```

Para hallar descriptivos según grupos definidos a partir de combinaciones de dos o más factores, se recurre a *Estadísticos / Resúmenes / Tabla de estadísticas*.




```
          cyl
am                4          6          8
  Autom. 0.06343161 0.08531853 0.18434524
  Manual 0.15971006 0.03649378 0.03673282
```

En cuanto a los gráficos, se pueden realizar gráficos de cajas (*Gráficas / Diagrama de cajas*) de una variable cuantitativa respecto de un factor.


R Graphics: Device 2 (ACTIVE)
mpg
10
15
20
25
30
4
6
8
cyl