

GUÍA PARA EL ANÁLISIS ESTADÍSTICO CON R COMMANDER

Victoria Alea Riera
Ernest Jiménez Garrido
Carmen Muñoz Vaquer
Elizabeth Torrelles Puig
Núria Viladomiu Canela

Departamento de Econometría, Estadística y Economía Española

GUÍA PARA EL ANÁLISIS ESTADÍSTICO CON R COMMANDER

Victoria Alea Riera
Ernest Jiménez Garrido
Carmen Muñoz Vaquer
Elizabeth Torrelles Puig
Núria Viladomiu Canela

Departamento de Econometría,
Estadística y Economía Española

Índice

PRESENTACIÓN	9
Capítulo 1. INTRODUCCIÓN	11
1.1. Instalación	11
1.2. Entorno de trabajo en R Commander	12
1.3. Entrada de datos	14
1.3.1. Entrada directa de datos	14
1.3.2. Importar un fichero de datos	15
1.3.2.1. <i>Importar un fichero de texto</i>	15
1.3.2.2. <i>Importar un fichero Excel</i>	18
1.3.3. Cargar un data frame (load)	19
1.4. Guardar el data frame (save)	20
1.5. Modificar el data frame	21
1.6. Modificar variables	26
1.6.1. Recodificar variables (recode)	27
1.6.2. Calcular una nueva variable	29
1.6.3. Añadir número de observación al conjunto de datos (<code>bd\$obsnumber<-1:n</code>)	30
1.6.4. Tipificar variables (scale)	30
1.6.5. Convertir variable numérica en factor	30
1.6.6. Renombrar variables (names)	31
1.6.7. Eliminar variables del conjunto de datos (<code>bd\$x<-null</code>)	32
1.7. Guardar y exportar resultados	32
1.8. Guardar script	32
Capítulo 2. ANÁLISIS UNIDIMENSIONAL DE DATOS CUALITATIVOS	35
2.1. Resumen e información de las variables (summary)	35
2.2. Distribución de frecuencias (table)	36
2.3. Representaciones gráficas (barplot)	38
2.4. Copiar, guardar e imprimir un gráfico	40
Capítulo 3. ANÁLISIS UNIDIMENSIONAL DE DATOS CUANTITATIVOS (1)	41
3.1. Distribuciones de frecuencias (table)	41
3.1.1. Discreta	41
3.1.2. Continua	43
3.2. Distribución de frecuencias acumuladas (cumsum)	46
3.3. Representaciones gráficas	49
3.3.1. Discreta (barplot)	49
3.3.2. Histograma (hist)	50
3.3.3. Polígono de frecuencias (lines)	53
3.3.4. Polígono de frecuencias acumuladas	56

Capítulo 4. ANÁLISIS UNIDIMENSIONAL DE DATOS CUANTITATIVOS (2)	59
4.1. Medidas estadísticas	59
4.2. Análisis exploratorio	63
4.2.1. Gráfica de tallos y hojas (stem)	63
4.2.2. Diagrama de caja (boxplot)	65
4.3. Bondad del ajuste a la distribución normal.....	68
4.3.1. Gráfica de comparación de cuantiles qq (qqplot).....	68
4.3.2. Test de normalidad Shapiro-Wilk (shapiro.test)	69
 Capítulo 5. ANÁLISIS BIDIMENSIONAL DE DATOS (1)	71
5.1. Tablas de contingencia.....	71
5.1.1. Tablas de doble entrada	71
5.1.2. Introducir y analizar una tabla de doble entrada.....	74
5.1.3. Tabla de entradas múltiples.....	78
5.2. Representación gráfica	80
5.3. Distribuciones condicionadas (tapply)	80
5.3.1. Gráfica de los valores medios.....	82
5.3.2. Diagrama box-plot por grupos.....	84
 Capítulo 6. ANÁLISIS BIDIMENSIONAL DE DATOS (2)	87
6.1. Gráficos	87
6.1.1. Diagrama de dispersión	87
6.1.2. Matriz de diagramas de dispersión.....	89
6.1.3. Gráfica lineal.....	91
6.1.4. Gráfica XY	91
6.2. Análisis de correlación (cor)	94
6.3. Regresión lineal (lm)	95
 Capítulo 7. DISTRIBUCIONES DE PROBABILIDAD	99
7.1. Distribuciones discretas.....	100
7.1.1. Cálculo de probabilidades acumuladas	100
7.1.2. Localización de cuantiles	101
7.1.3. Función de cuantía	103
7.1.4. Representación gráfica	103
7.1.5. Muestras aleatorias	105
7.2. Distribuciones continuas	106
7.2.1. Cálculo de probabilidades	106
7.2.2. Localización de cuantiles	107
7.2.3. Representación gráfica	109
7.2.4. Muestras aleatorias	110
 Capítulo 8. INTERVALOS DE CONFIANZA	113
8.1. Intervalo para la media de una población normal.....	113
8.2. Intervalo para la varianza de una población normal.....	115
8.3. Intervalo para la proporción de éxitos.....	117

8.4. Intervalo para la diferencia de medias en poblaciones normales con varianzas desconocidas pero supuestamente iguales	121
Capítulo 9. CONTRASTES PARAMÉTRICOS PARA UNA MUESTRA	125
9.1. Contraste para la media de una población normal (t.test)	125
9.2. Contraste para la proporción de éxitos poblacional (prop.test) (binom.test)	129
Capítulo 10. CONTRASTES PARAMÉTRICOS PARA DOS O MÁS MUESTRAS.....	139
10.1. Contraste para la diferencia de medias de poblaciones normales (t.test)	139
10.2. Contraste para la diferencia de medias con muestras apareadas de poblaciones normales (t.test)	142
10.3. Contraste para la igualdad de varianzas de poblaciones normales (var.test)	143
10.4. Contraste para la igualdad de proporciones (prop.test)	144
10.5. Contraste para igualdad de varianzas de más de 2 poblaciones normales (bartlett.test) (leveneTest)	146
10.6. Anova de un factor (aov)	148
Anexo 1. BASE DE DATOS	153
Anexo 2. OPERADORES Y FUNCIONES	157
Anexo 3. FUNCIONES DEFINIDAS EN LA GUÍA.....	161
BIBLIOGRAFÍA	169

Presentación

Esta guía se plantea como una introducción al uso del software estadístico R, utilizando el entorno de R Commander, con el objetivo de proporcionar un complemento informático al contenido teórico de un curso de Estadística aplicada a las Ciencias Sociales.

R y su interfaz R Commander se han convertido en una herramienta básica del análisis descriptivo y de la Inferencia Estadística en muchas universidades y en la mayoría de centros e Institutos de Investigación.

Para la investigación cuantitativa en general, y en las ciencias sociales y económicas en particular, se utilizan desde hace tiempo de forma habitual productos de software estadístico como: SAS, STATA, SPSS, etc. El motivo por el que creemos que debe promoverse la implantación del programa R va más allá de que sea un producto gratuito de código abierto y que se pueda conseguir en unos minutos libremente sin coste económico alguno. La principal motivación es lograr que se sigan sumando esfuerzos colectivos para mejorar el análisis sobre las verdades y mentiras de los datos de la realidad que nos toca vivir. Si R ha conseguido extenderse y mejorar ha sido gracias a esfuerzos desde todos los lugares del planeta, y ese carácter le da un valor tremendamente importante demostrando que hay objetivos diferentes de la obtención de beneficios monetarios, y ello no disminuye la calidad en la cohesión y en la complejidad.

En esta guía todos los procedimientos se han ilustrado fundamentalmente analizando las variables de una única base de datos que contiene 600 observaciones extraídas aleatoriamente de la Encuesta de Estructura Salarial realizada por el INE en el año 2010. Esta base de datos se describe detalladamente en el Anexo 1. El lector puede acceder a esta base de datos y al resto del material en la dirección <http://www.publicacions.ub.edu/liberweb/anEsR-Commander>.

Se recomienda al lector que cree una carpeta de trabajo para guardar las bases de datos utilizadas en la guía y que la defina como directorio activo al iniciar cada sesión con la secuencia Fichero ► Cambiar directorio de trabajo.

No se puede acabar esta presentación sin agradecer los esfuerzos tan generosos de toda la comunidad científica que ha creado y que mantiene vivo el proyecto R. Merece la pena visitar la página <http://www.r-project.org>, donde de forma resumida se detallan los objetivos del proyecto.

Barcelona, julio de 2014

Los autores

INTRODUCCIÓN

R es un lenguaje y entorno de programación que permite trabajar con los datos, realizar análisis estadísticos y obtener gráficos. El entorno R ofrece la posibilidad de cargar diferentes paquetes con finalidades específicas de cálculo o representación gráfica. Es, probablemente, uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras.

R forma parte de un proyecto colaborativo y abierto. Se trata de un proyecto de software libre, resultado de la implementación del lenguaje S. Sus usuarios pueden aportar paquetes con funciones o librerías. Dado el enorme número de nuevos paquetes, estos se han organizado en vistas (o temas), que permiten agruparlos según su naturaleza y función. Por ejemplo, hay grupos de paquetes relacionados con estadística bayesiana, econometría, series temporales, entre otros.

En esta guía se detalla el funcionamiento y utilización del paquete R Commander, así como la sintaxis específica del programa R y la utilización de funciones. Este paquete proporciona un entorno de ventanas que permite al usuario realizar muchas de las tareas que se pueden hacer con el lenguaje de comandos, pero de una manera más fácil de usar. Así, mediante las ventanas de R Commander, podemos cargar datos y realizar una serie de análisis estadísticos accediendo a las correspondientes opciones de menú.

1.1. Instalación

Para instalar la versión de R y R Commander adaptada por la Universidad de Cádiz se debe:

1. Bajar el programa que se encuentra en la página:
<http://knuth.uca.es>
2. Elegir la opción:
Paquete R UCA
3. Descargar la última versión desde el servidor 2. Permite descargar el fichero R-UCA-*.exe¹ en el subdirectorío de las bajadas de internet del ordenador al clicar el botón **Guardar archivo**.
4. Ejecutar el programa de instalación clicando sobre R-UCA-*.exe. En el cuadro de diálogo que se abre activar el botón **Ejecutar** para iniciar la instalación.

Usuarios de Mac

En el enlace <http://cran.r-project.org> se debe seleccionar el sistema operativo *Download R for (Mac) OS X*.

¹ El * sustituye el valor de la última versión disponible.

Sistema operativo Mac OS X 10.6²

Es preciso descargar 3 componentes:

- R-3.1.0-snowleopard.pkg
- En tools seleccionar
 - gfortran-4.2.3.pkg y
 - tcltk-8.5.5-x11.pkg

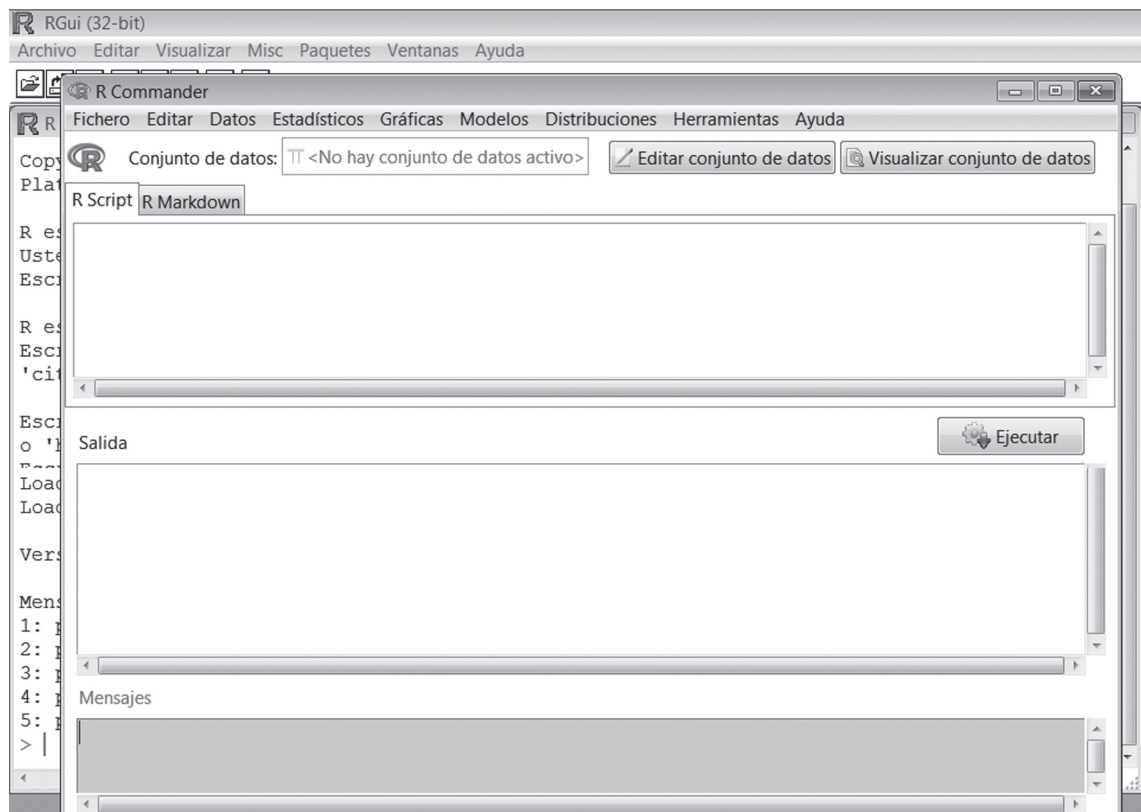
Una vez descargados los 3 archivos:

- Ejecutar primero R-3.1.0-snowleopard.pkg y seguir las instrucciones.
- Ejecutar gfortran-4.2.3.pkg en segundo lugar y seguir las instrucciones.
- Por último, instalar tcltk-8.5.5-x11.pkg

1.2. Entorno de trabajo en R Commander

Al cargar el programa se abren dos ventanas superpuestas:

- RGui
- R Commander

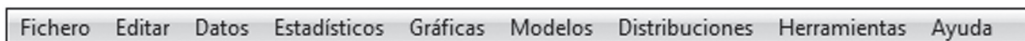


Este manual solo hace referencia a la ventana **R Commander**.

Esta pantalla está dividida en tres zonas:

² Funciona bajo OS X 10.6 y superior (Snow Leopard, Lion, Mountain Lion, Mavericks...). Para sistemas operativos más antiguos, habrá que instalar R desde el enlace <http://cran.r-project.org/bin/macosx/old>

1. En la zona superior de la pantalla se encuentra la **barra de menús**:



- *Fichero*: contiene procedimientos relacionados con la lectura y el almacenaje de ficheros de instrucciones y resultados.
 - *Editar*: permite copiar, mover, pegar, seleccionar, etc., los contenidos de las ventanas de instrucciones y resultados.
 - *Datos*: permite introducir, cargar o importar ficheros de datos. Este menú será el adecuado para manipular y transformar la base de datos y sus variables.
 - *Estadísticos*: contiene una serie de procedimientos estadísticos.
 - *Gráficas*: permite la creación, modificación y edición de gráficos.
 - *Modelos*: contiene procedimientos de inferencia estadística.
 - *Distribuciones*: permite calcular probabilidades, generar muestras aleatorias, realizar gráficos, etc., de diferentes distribuciones de probabilidad.
 - *Herramientas*: permite cargar paquetes R que no están incluidos en el paquete básico. Para cargar un paquete no instalado en R Commander este previamente debe instalarse desde la ventana RGui.
 - *Ayuda*: da acceso a un manual básico de instrucciones.
2. Bajo la barra de menús se encuentra la barra de información de los **Conjuntos de datos** y de los **Modelos activos**:



- El primer botón abre el directorio de las bases de datos o data frames³ cargados y permite elegir el que se quiere activar.
 - El botón **Editar conjunto de datos** abre el *Editor de datos* donde se puede introducir modificaciones en la base de datos activa. Para poder ejecutar cualquier procedimiento el editor debe permanecer cerrado ya que, en caso contrario, R Commander interpreta que la base de datos está en proceso de modificación y no accederá a ella hasta que dicho proceso esté completado y la base resultante quede fijada cerrando la ventana.
 - El botón **Visualizar conjunto de datos** abre una ventana con la base de datos activa, de modo que durante su análisis se pueden visualizar los datos pero no modificarlos.
3. La siguiente zona incluye tres espacios:
 - **R Script** y **R Markdown**⁴ o ventana de instrucciones: muestra los comandos que se ejecutan al realizar las operaciones que se seleccionan mediante la barra de menús. También permite escribir directamente o recuperar conjuntos de instrucciones (Scripts) y ejecutarlos pulsando el botón **Ejecutar**.
 - **Salida** o ventana de resultados: muestra en rojo las instrucciones ejecutadas y en azul los resultados.
 - **Mensajes**: muestra en rojo mensajes de error, en verde avisos y en azul otros mensajes.

³ Data frame en R es sinónimo de base de datos.

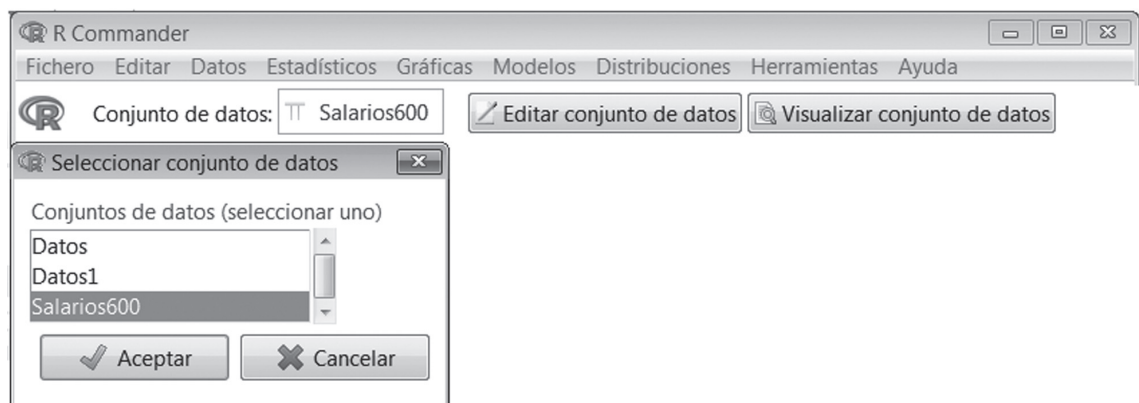
⁴ Ventana de lenguaje HTML.

1.3. Entrada de datos

Cuando se inicia una sesión en R Commander, no hay ningún conjunto de datos activo y el paquete ofrece diversos procedimientos para incorporarlos en la memoria:

- Entrada directa de datos.
- Importación de datos desde
 - un archivo de texto,
 - otro paquete estadístico,
 - hojas de cálculo.
- Cargar un data frame Rda.

Cuando hay varios conjuntos de datos en memoria, se puede elegir cualquiera de ellos como data frame activo seleccionándolo en la ventana que se despliega clicando sobre el *Conjunto de datos* o con la secuencia *Datos ► Conjunto de datos activo ► Seleccionar conjunto de datos activo*. La ventana que se despliega en ambos casos es:



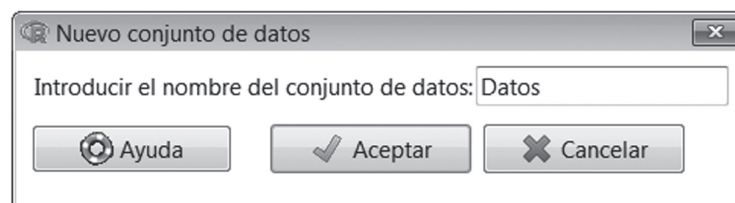
La mayoría de los procedimientos requiere que haya un conjunto de datos activo para ejecutarlos.

1.3.1. Entrada directa de datos

Para entrar datos directamente a R la secuencia que abre el correspondiente cuadro de diálogo es:

Datos

Nuevo conjunto de datos



Hay que indicar el nombre que se quiere dar a la nueva base de datos en *Introducir el nombre del conjunto de datos*; al **Aceptar** se abrirá el *Editor de datos*:

	var1	var2	var3	var4	var5	var6
1						
2						
3						
4						
5						
6						
7						

Por defecto los nombres de las variables son var1, var2 ... Pueden modificarse clicando sobre ellos y aparecerá la ventana *Editor de variable*, que permite introducir el nuevo nombre y seleccionar el tipo de la variable (numérica/carácter).

Una vez introducidos los datos se debe cerrar el *Editor de datos*.

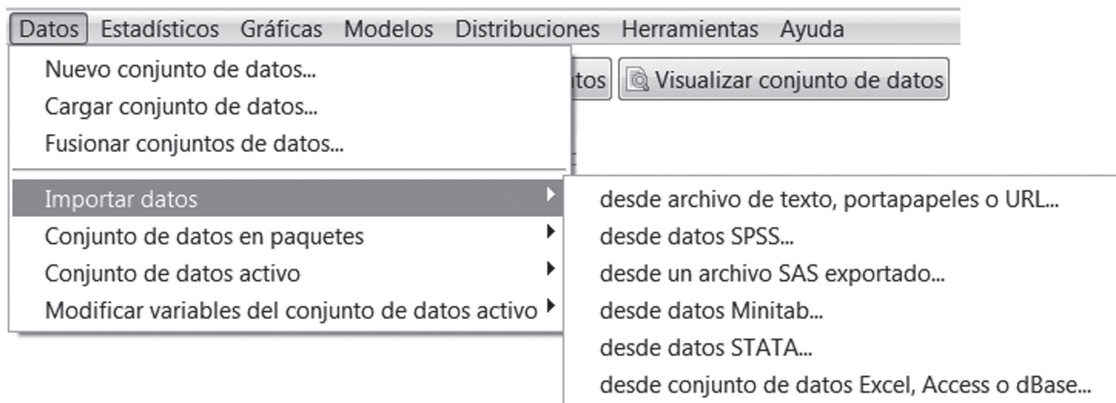
Recuerde que, para poder ejecutar cualquier procedimiento, el *Editor de datos* debe permanecer cerrado.

1.3.2. Importar un fichero de datos

El programa R Commander permite analizar bases de datos creadas con otros programas. Para ello se deberán importar mediante la secuencia:

Datos

Importar datos



Esta opción permite leer datos desde:

- Ficheros de texto.
- Ficheros de otros programas estadísticos como SPSS, SAS, Minitab, STATA.
- Hojas de cálculo o bases de datos como Excel, Access o dBase.

1.3.2.1. Importar un fichero de texto

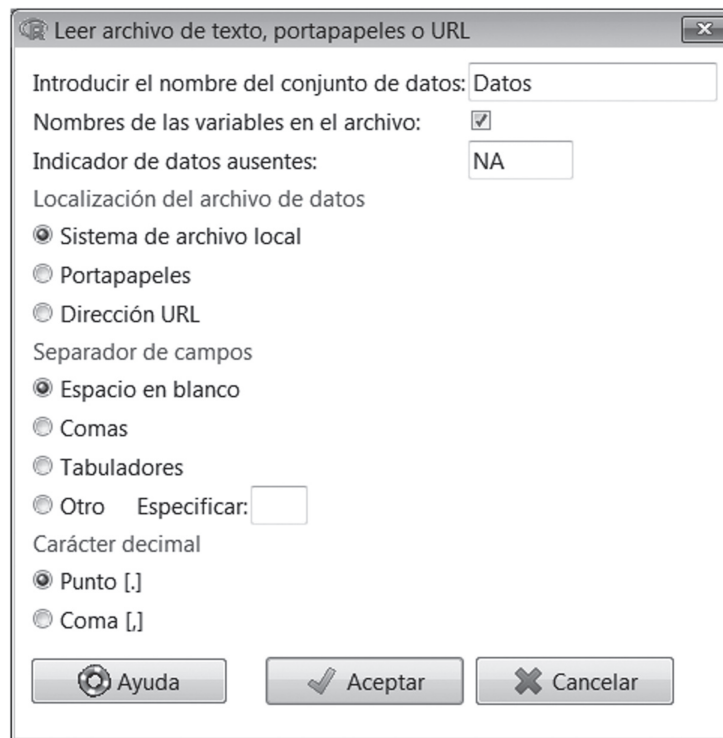
Para importar datos desde un fichero txt la secuencia es:

Datos

Importar datos

desde archivo de texto, portapapeles o URL

El cuadro de diálogo que se abre es:



- *Introducir el nombre del conjunto de datos:* se debe indicar el nombre que se quiere dar en R Commander al data frame.
- *Nombre de las variables en el fichero:* esta opción se mantendrá activada si la primera fila del fichero txt contiene los nombres de las variables; en caso contrario se debe desactivar.
- *Indicador de datos ausentes:* se indicará el código utilizado en la fichero txt para los valores missing. Se recomienda utilizar el código de R, **NA** (not available).
- *Localización del archivo de datos:* se seleccionará la localización del fichero txt, entre las siguientes opciones:
 - *Sistema local* si los datos se encuentran en un subdirectorio del ordenador.
 - *Portapapeles* si los datos están temporalmente en el portapapeles.
 - *Dirección URL* si están en una dirección de internet.
- *Separador de campos:* hay que seleccionar el tipo de separador utilizado en el fichero txt.
- *Carácter decimal:* se indicará el elemento escogido en el archivo txt para separar la parte entera de la parte decimal. Se aconseja utilizar como indicador decimal el punto.

Ejemplo 1.1: Importación de los ficheros: Marathon_peor.txt y Marathon_mejor.txt.

Antes de realizar la importación se debe comprobar, editando estos ficheros con el Bloc de Notas, si el fichero contiene:

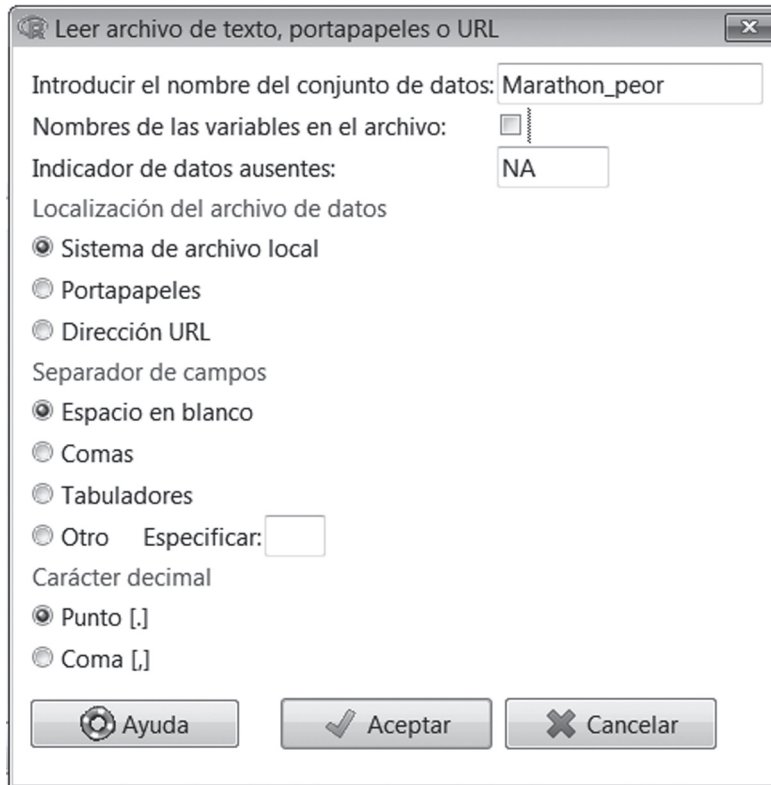
- Nombres de las variables en la primera fila.
- Valores ausentes y codificación de los mismos.
- Separador de campos.
- Carácter decimal.

Con la secuencia:

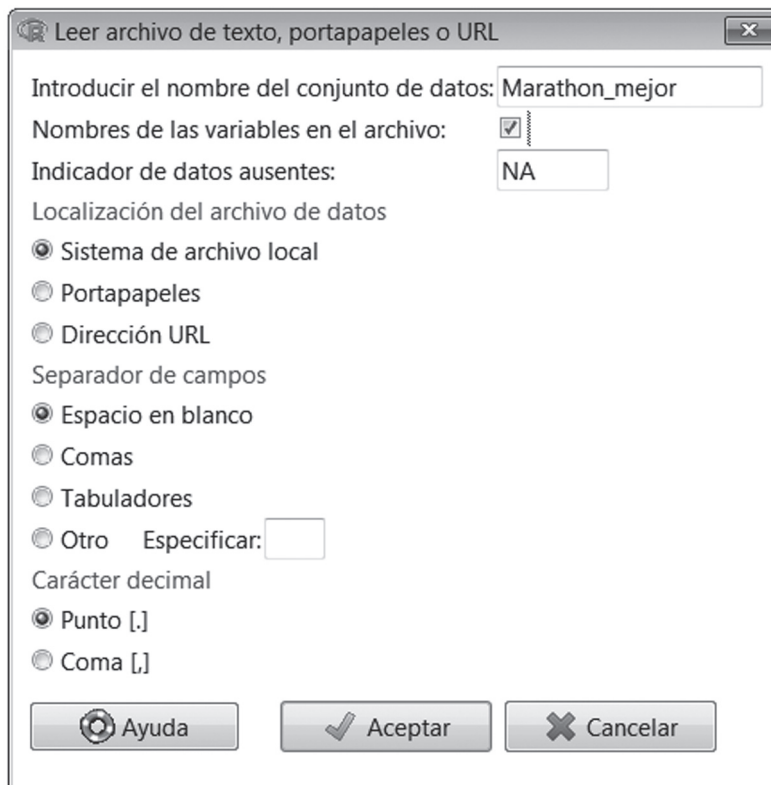
Datos ► Importar datos ► desde archivo de texto, portapapeles o URL

Se abre el cuadro de diálogo donde:

- Para leer el archivo `Marathon_peor.txt` se ha desactivado *Nombres de las variables en el fichero* ya que no están en el fichero `txt`.



- Para leer `Marathon_mejor.txt` se deja activada la opción *Nombre de las variables en el archivo*.



Se pueden guardar las bases de datos creadas con los mismos nombres, mediante la secuencia:

Datos ► Conjunto de datos activo ► Guardar el conjunto de datos activo

Estas bases de datos quedarán guardadas con extensión **Rda**.

1.3.2.2. Importar un fichero Excel

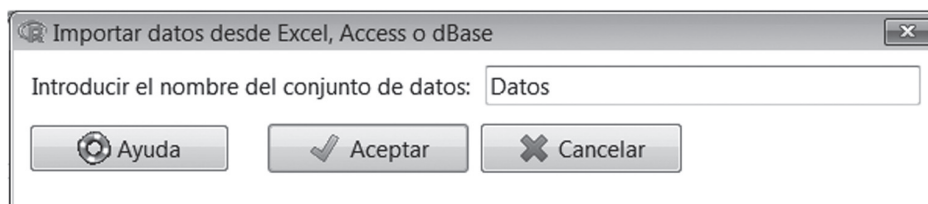
Para importar datos desde un fichero Excel la secuencia es:

Datos

Importar datos

desde conjunto de datos Excel, Access o dBase

Aparece el cuadro de diálogo *Importar datos desde Excel, Access o dBase*,



en el que únicamente se debe indicar el nombre que se quiere dar al data frame; al **Aceptar** se abre el Explorador de archivos donde se seleccionará el fichero Excel que se desea importar. Al clicar **Abrir** se despliega un cuadro de diálogo en R en el que se selecciona la hoja del fichero Excel que se quiere importar.

Si se quiere importar varias hojas se debe repetir el proceso.

Ejemplo 1.2: Importación de la hoja Full1 del fichero Salarios600.xlsx.

Para importar la hoja con nombre Full1 del fichero Salarios600.xlsx y guardarla con el nombre Salarios se debe seguir la secuencia:

Datos ► Importar datos ► desde conjunto de datos Excel, Access o dBase

En el cuadro de diálogo se debe indicar el nombre que se quiere dar a la base de datos, en este caso Salarios.

Al **Aceptar**, en el Explorador de archivos se elige la carpeta donde está guardado el documento Salarios600.xlsx y se selecciona.