Práctica 1. Introducción a R

1. ¿Qué son R y R Commander?

R es un lenguaje y entorno de programación para el análisis estadístico y gráfico. Se trata de un proyecto de software libre que, en la actualidad, se está utilizando cada vez más en diferentes campos de investigación. R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

La página principal del proyecto bajo el que se desarrolla R es `https://www.r-project.org/`. La universidad de Zaragoza a través de su Oficina de software libre también proporciona información y ayuda sobre R.

Si utilizas R en alguno de tus trabajos debes citarlo e incluir en la bibliografía su referencia:

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

R Commander es un paquete de R que proporciona un interfaz con menús desplegables que permiten la aplicación de las técnicas estadísticas más habituales y evita utilizar de forma directa el lenguaje de programación facilitando el uso de R. Si utilizas *R Commander* en algunos de tus trabajos también debes citarlo e incluir en la bibliografía su referencia:

Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. Journal of Statistical Software, 14(9): 1-42.

2. Iniciar R y R Commander

Para iniciar *R* desde el escritorio:

1. Se selecciona en la parte inferior izquierda de la pantalla:

   Inicio → Todos los programas → R → R

2. Aparece la ventana *R Console*, que se muestra en la figura 2.1, y es posible que a continuación se abra automáticamente la ventana de *R Commander* (figura 2.2), que es en la que se trabajará y donde se realizarán los análisis estadísticos.

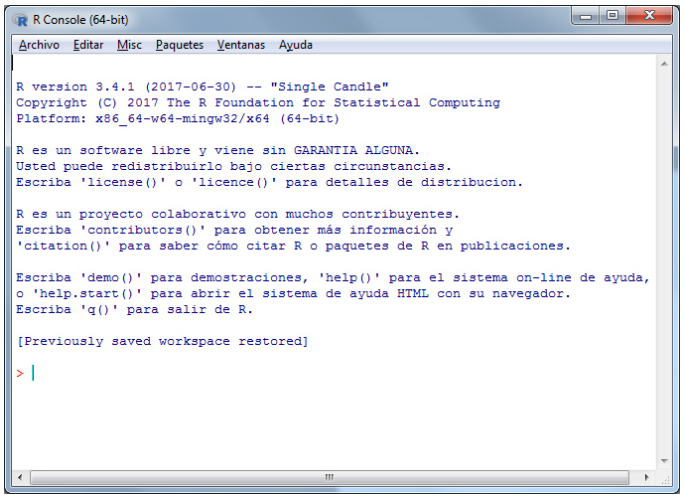


Figura 2.1: Ventana de R.

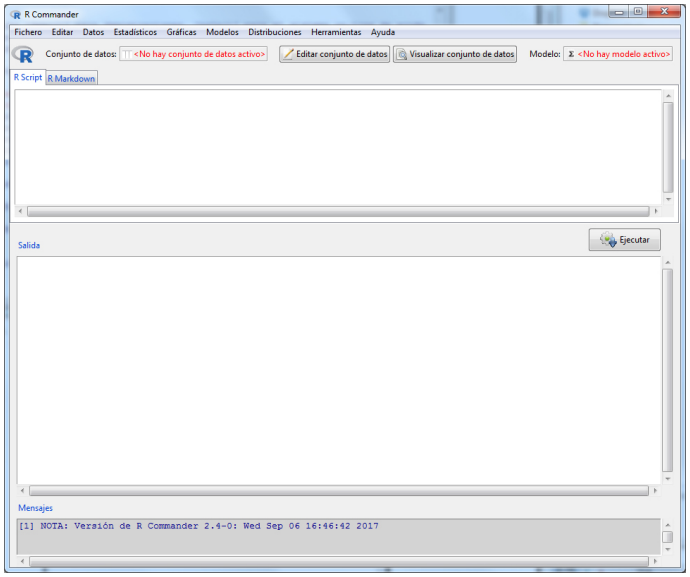


Figura 2.2: Ventana de R Commander.

3. En caso de que *R Commander* no se haya abierto automáticamente, se selecciona en la parte superior de la ventana *R Console* la opción:

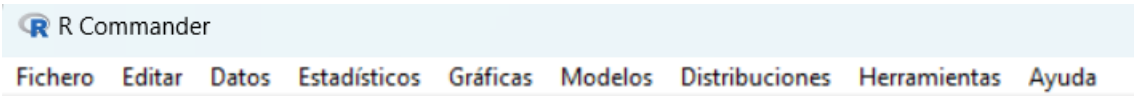
   Paquetes → Cargar paquete

   En la lista se elige el paquete **Rcmdr** y se pulsa el botón OK.

   Otra opción si no nos aparece la ventana de *R Commander* automáticamente es escribir `Commander()` en la venta *R Console*, que es la que aparece en la figura 2.1.

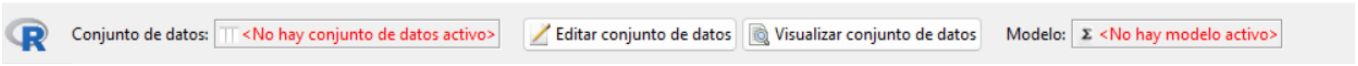
3. Menú de R Commander

La parte superior de la ventana de R Commander contiene el menú principal.



- **Fichero**: Permite guardar las instrucciones y los resultados de una sesión de trabajo.
- **Editar**: Contiene las opciones habituales relacionadas con la edición, *cortar, copiar, pegar, borrar, buscar, seleccionar todo, deshacer, rehacer, limpiar ventana.*
- **Datos**: Se encuentran las opciones para crear, cargar, guardar o modificar un conjunto de datos. Otras opciones de este mismo menú permiten operaciones con los datos como pueden ser la *recodificación*, la *tipificación, construcción de nuevas variables, ...*
- **Estadísticos**: Se encuentran las técnicas estadísticas más habituales.
- **Gráficas**: Se encuentran los diferentes tipos de representaciones gráficas.
- **Modelos**: Se encuentran algunas técnicas de selección de variables y validación de un modelo lineal.
- **Distribuciones**: Se encuentran las herramientas para resolver problemas del cálculo de probabilidades.

En la segunda franja horizontal tenemos la siguiente barra de herramientas:



- **Conjunto de datos**: Este botón muestra el nombre del conjunto de datos activo o *No hay conjunto de datos activo* cuando todavía no se ha cargado o creado ningún conjunto de datos. Pulsando sobre este botón, se despliega un menú que permite activar otro conjunto de datos, entre los que haya disponibles.
- **Editar conjunto de datos**: Permite la edición del conjunto de datos activo en un entorno similar al de una hoja de cálculo. Durante la edición de los datos no es posible realizar ninguna otra operación con R. Por ello es absolutamente imprescindible cerrar la venta de edición de datos antes de intentar cualquier otra operación.
- **Visualizar conjunto de datos**: Muestra una ventana con los datos del conjunto de datos activos en un formato similar al anterior. Esta ventana no permite la modificación de los datos, pero puede mantenerse abierta mientras se continúa haciendo operaciones con R Commander.
- **Modelo**: La leyenda muestra el nombre del modelo activo o *No hay modelo activo*, cuando no se ha construido ningún modelo previamente. La pulsación sobre dicho botón permite la selección del modelo a activar de entre los disponibles.

4. Conjunto de datos

4.1. Abrir un archivo de datos con R

El primer paso cuando se desea analizar un conjunto de datos con R Commander es crear o cargar el archivo que los contiene.

La extensión *.RData* corresponde a un fichero propio de R. Para abrirlo vamos al menú:

Datos → Cargar conjunto de datos

Ejercicio 1. *El archivo ZonasProtegidas.RData contiene de cada una de las comarcas aragonesas: la provincia a la que pertenece la mayor parte de su territorio, su superficie en km^2, su población en 2011 y el número de espacios naturales protegidos, lugares de importancia comunitaria (LIC) y zonas de especial protección para las aves (ZEPA) registrados en 2014, así como la superficie que ocupan (en hectáreas).*

1. *Descarga el archivo ZonasProtegidas.RData y ábrelo en R Commander.*
2. *Observa que se ha escrito una instrucción en las ventanas R script y Salida.*
3. *Observa en la parte superior izquierda de la ventana que aparece el nombre del archivo que has cargado.*
4. *Observa el mensaje que aparece en la ventana de Mensajes.*
5. *Pulsa el botón Visualizar conjunto de datos para ver el contenido del archivo.*
6. *Identifica los individuos, la población, las variables y el tipo de variables que se disponen en el conjunto de datos. ¿Se trata de un censo o de una muestra?*

4.2. Guardar un archivo de datos con R

Cuando se crea un fichero con R Commander, o se realizan modificaciones útiles sobre un fichero de datos existente, conviene guardar el fichero generado. Para ello, estando activo el fichero de datos que se desea guardar se utiliza la siguiente opción:

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

De esta forma el archivo queda almacenado en el formato propio de R lo que puede dificultar su uso con otros programas. Para evitar este problema es recomendable que, además de guardar el archivo en el formato propio de R, se guarde en un formato que puedan leer otros programas que es el formato texto. Para ello:

Datos → Conjunto de datos activo → Exportar el conjunto de datos activos

Ejercicio 2. *Guarda el conjunto de datos utilizado en el ejercicio 1 en formato texto.*

Se pueden guardar los ficheros de instrucciones (texto que aparece en la ventana *"R Script"*) generados en una sesión de trabajo. Para ello en el menú, se utiliza la opción:

Fichero → Guardar las instrucciones...

Si queremos abrir un fichero de instrucciones creado previamente:

Fichero → Abrir archivo de instrucciones...

4.3. Importar ficheros

R utiliza un formato propio para almacenar los datos (extensión *.RData*), sin embargo es capaz de leer datos de otros tipos de ficheros, por ejemplo *Excel, SPSS, SAS, Minitab, STATA* y también desde ficheros de texto con formato *.csv*. Para ello se utiliza el menú *Importar datos* que se encuentra en:

Datos → Importar datos

Si los datos están almacenados en formato Excel, debes asegurarte de que la hoja no contiene más información que la que corresponde al nombre de las variables y a los datos recogidos de cada una de ellas.

Ejercicio 3. *Los datos que corresponden a la cantidad detectada de los isómetros α, β, γ y δ de HCH (hexaclorociclohexano) en el río Gállego en dos puntos de muestreo: pie de prensa del pantano de la Peña y Ardisa entre el 25 de septiembre y el 9 de diciembre de 2014 se encuentran en el archivo HCH.xlsx.*

1. *Abre el archivo con Excel y comprueba qué hoja contiene la información adecuada para ser leída desde R.*
2. *Comprueba si la primera fila contiene los nombres de las variables.*
3. *Comprueba si hay datos faltantes y cómo están indicados.*
4. *Importa el archivo desde R Commander.*

Datos → Importar datos → desde un archivo Excel...

5. *En el cuadro de diálogo que aparece introduce un nombre para el conjunto de datos (Por efecto es Dataset). Se le debe indicar si el nombre de las variables aparece en la primera fija de la hoja de Excel (notar que por norma general esto debería ser así siempre), y que las variables carácter las convierta en factores (variables cualitativas de R). Si pulsamos Aceptar, aparece un diálogo de apertura de fichero, buscamos el fichero HCH.xlsx y lo seleccionamos. Si el archivo contiene más de una hoja, se abrirá una nueva ventana para que seleccionemos la hoja en la cuál se encuentran los datos que queremos importar.*

6. *Lee en la ventana de mensajes de R Commander el mensaje que aparece. Si se ha leído bien, se indica el número de filas y columnas que contiene el archivo. En otro caso, aparecerá un mensaje de error.*

7. *Guarda el conjunto de datos en el formato propio de R.*

Podemos observar que los datos faltantes aparecen representados en *R Commander* con el símbolo *NA* (*Not Available*).

Otra manera de encontrar almacenados los datos es en formato texto (extensión *.txt*).

Ejercicio 4. *El archivo Vinos.txt contiene los resultados de un análisis químico efectuado a los vinos de una región de Italia que provienen de tres cultivos diferentes* [1].

1. *Abre el archivo con un editor de texto.*

2. *Comprueba si la primera fila contiene los nombres de las variables.*

3. *Comprueba si hay datos faltantes y cómo están indicados.*

4. *Comprueba cómo están separados los datos (espacio en blanco, coma, punto y coma, ...).*

5. *Comprueba si el carácter decimal es un punto o una coma.*

6. *Importa el archivo desde R Commander*

Datos → Importar datos → desde archivo de texto, portapapeles o URL ...

7. *En el cuadro de diálogo que aparece introduce un nombre para el conjunto de datos (Introducir el nombre del conjunto de datos). Se le debe indicar si el nombres de las variables aparece en el archivo o no, el carácter separador de la información recogida, así como el carácter decimal.*

8. *Lee en la ventana de mensajes de R Commander el mensaje que aparece. Si se ha leído bien, se indica el número de filas y columnas que contiene el archivo. En otro caso, aparecerá un mensaje de error.*

9. *Guarda el conjunto de datos en el formato propio de R.*

5. Variables en R: Tipos y su modificación

5.1. Tipos de variables

En R las variables se clasifican en *numéricas* (variables cuantitativas) y *factores* (variables cualitativas). En general, son numéricas aquellas cuyos valores figuran en el archivo

[1] Fuente: `https://archive.ics.uci.edu/ml/datasets/Wine`

con números y son factores aquellas cuyos valores figuran en el archivo con caracteres alfanuméricos (letras acompañadas o no de números).

Se debe tener cuidado con las codificaciones, algunas veces las variables cualitativas se codifican con números por comodidad, pero deben ser tratadas adecuadamente como variables cualitativas.

Para saber de qué manera ha clasificado R las variables, es útil realizar un estudio básico de cada una de ellas con la opción:

Estadísticos → Resúmenes → Conjunto de datos activo

Para las variables de tipo *factor*, R Commander realiza un recuento de los distintos valores y para las *numéricas* calcula algunas medidas numéricas.

Ejercicio 5. *El archivo submuestra_pisa.RData contiene información sobre el Programa Internacional para la Evaluación de Estudiantes o Informe Pisa, llevado a cabo por la OCDE. Evalúa las competencias de los alumnos de 15 años. Para ello se seleccionan algunos institutos de secundaria en lo que los que los alumnos son examinados en varias disciplinas. En particular* 6000000 *estudiantes realizaron la evaluación en 2018, representando a cerca de* 32 *millones de alumnos de 15 años en centros educativos de los* 79 *países y economías participantes. En el archivo adjunto, se encuentran los datos seleccionados al azar de* 30 *alumnos españoles, donde se recoge su género, los estudios de sus padres, el acceso en casa a ordenador, el número de coches que posee su núcleo familiar, y la puntuación en el test normalizado de Pisa en las áreas de Matemáticas y Ciencias.*

1. *Carga el conjunto de datos submuestra_pisa.RData en R Commander.*
2. *Identifica los individuos, la población, las variables y el tipo de variables que se disponen en el conjunto de datos. ¿Se trata de un censo o de una muestra?*
3. *Comprueba si las variables se han clasificado correctamente. ¿Existe alguna variable cuantitativa que se ha detectado como cualitativa? ¿Existe alguna variable cualitativa que se ha detectado como cuantitativa? La siguiente instrucción puede ayudar:*

Datos → Conjunto de datos activo → Convertir variables de caracteres en factores...

5.2. Cambiar variable numérica a factor

Ejercicio 6. *Carga el fichero CCAA0910.RData que contiene información sobre un conjunto de estudiantes. Haz un resumen numérico y clasifica las variables que contiene el fichero. ¿Crees que todo es correcto?*

En el último ejemplo hemos observado que la variable *sexo* está mal codificada, ya que se trata de una variable cualitativa pero está representada con números. Para convertir una variable numérica en una variable factor se utiliza:

Datos → Modificar variables del conjunto de datos activo → Convertir variable numérica en factor...

Este menú permite utilizar el mismo nombre de la variable (o uno nuevo) y que los números sean tratados como caracteres (o bien asignarles otros valores).

Para cambiar las varibale *sexo*, seleccionamos la opción, en el cuadro de diálogo que aparece seleccionamos la variable y dejamos marcada la opción *Asignar nombres a los niveles*. En *Nuevo nombre o prefijo para variables múltiples* lo dejamos tal y como esta, de forma que la codificación correcta sustituirá a la antigua. Pulsando el botón *Aceptar* aparece un diálogo confirmando que queremos sustituir la antigua variable *sexo* por la nueva, y después pide las equivalencias entre los números y los nuevos valores.

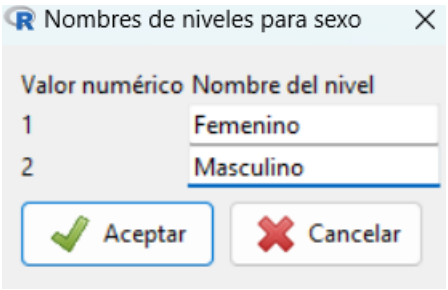


Figura 5.1: Conversión de una variable numérica a factor

Ejercicio 7. *Corrige la variable gafas convirtiéndola a factor sin cambiarla de nombre. Los nuevos valores de las variables deben ser Sí y No, para los valores 1 y 0 respectivamente.*

5.3. Calcular una nueva variable a partir de otras

En muchas ocasiones resulta interesante analizar nuevas variables de interés, que se pueden construir a partir de las ya existentes. El menú es:

Datos → Modificar variables del conjunto de datos → Calcular una nueva variable

En la siguiente tabla aparecen las principales operaciones que se pueden hacer con las variables:

Operación	Símbolo
Suma, resta, multiplicación, división	+, -, *, /
Potencia	∧
Logaritmo neperiano	log()
Raíz cuadrada	sqrt()

Cuadro 1: Operaciones básicas con variables

Ejercicio 8. *Con los datos del fichero CCAA0910.RData calcula el Índice de Masa Corporal (IMC) mediante la siguiente fórmula:*

$$IMC = peso/altura^2$$

Cabe notar que la altura tiene que estar expresada en metros, y en nuestros datos la tenemos expresada en centímetros.

En la figura 5.2 podemos observar el proceso que se debe realizar:

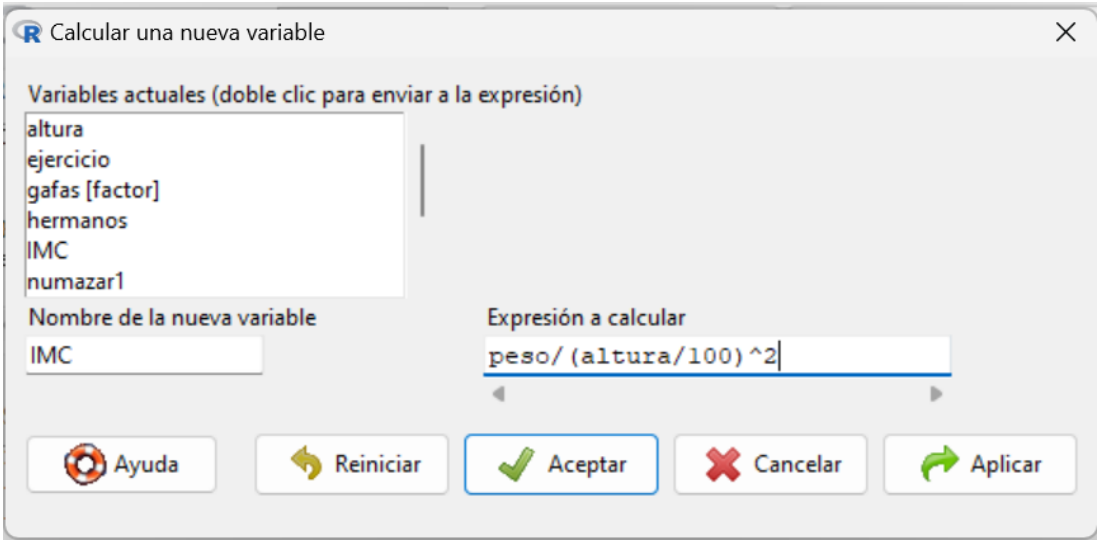


Figura 5.2: Creación de una nueva variable

Ejercicio 9. *Abre el archivo HCH y calcula el contenido de HCH en cada uno de los puntos de muestreo como suma de la cantidad detectada de cada uno de los isómeros considerados.*

5.4. Recodificar variables

En ocasiones, tenemos una variable cuantitativa, y la queremos clasificar según categorías, es decir convertir en una variable categórica. Por ejemplo, el IMC que acabamos de calcular se suele clasificar por categorías:

Valores IMC	Categoría en IMC
Menor que 18.5	BajoPeso
[18.5, 25)	PesoNormal
[25, 30)	Sobrepeso
30 o más	Obesidad

El menú que permite realizar esta transformación es:

Datos → Modificar variables del conjunto de datos activo → Recodificar variables

Los pasos a seguir son los siguientes:

1. Elegir la variable que se quiere recodificar (se pueden elegir varias variables a la vez siempre que sean del mismo tipo y se quieran recodificar de la misma manera).
2. Escribir el nombre de la nueva variable
3. Si la nueva variable tiene carácter numérico, se elimina la marca de la opción *Convertir cada nueva variable en factor*. En otro caso, se debe dejar marcada esta opción.
4. Escribir las reglas de recodificación

A continuación se muestran las reglas de recodificación. Es **muy importante** tener en cuenta que R recodifica en el orden en el que se incluyen las directrices de recodificación:

- **lo:a="Categoría 1"** indica que la categoría 1 de la nueva variable serán los valores $\leq$ **a**, si el valor **a** no ha sido previamente recodificado. Si ha sido incluido en una recodificación previa, entonces incluye $<$ **a**. El valor Categoría 1 se escribe entre comillas si la variable nueva es de tipo factor.
- **a:b="Categoría 2"** indica que los valores de la variable original dentro del intervalo **[a,b]** pasarán a ser **Categoría 2** en la nueva variable. Si el valor a (o bien b o ambos) ha sido utilizado en alguna regla previa, entonces incluye $(a, b]$ (o bien $[a, b)$ o (a, b)).
- **b:hi="Categoría 3"** indica que la categoría 3 comprenderá los valores $\geq$**b**, siempre que **b** no se haya recodificado previamente.

En nuestro caso, que vamos a recodificar el IMC, deberíamos escribir lo siguiente:

```
30:hi = "Obesidad"
25:30 = "Sobrepeso"
18.5:25 = "PesoNormal"
lo:18.5 = "BajoPeso"
```

Notar que *lo* indica el límite inferior de la variable y *hi* indica el límite superior de la variable.

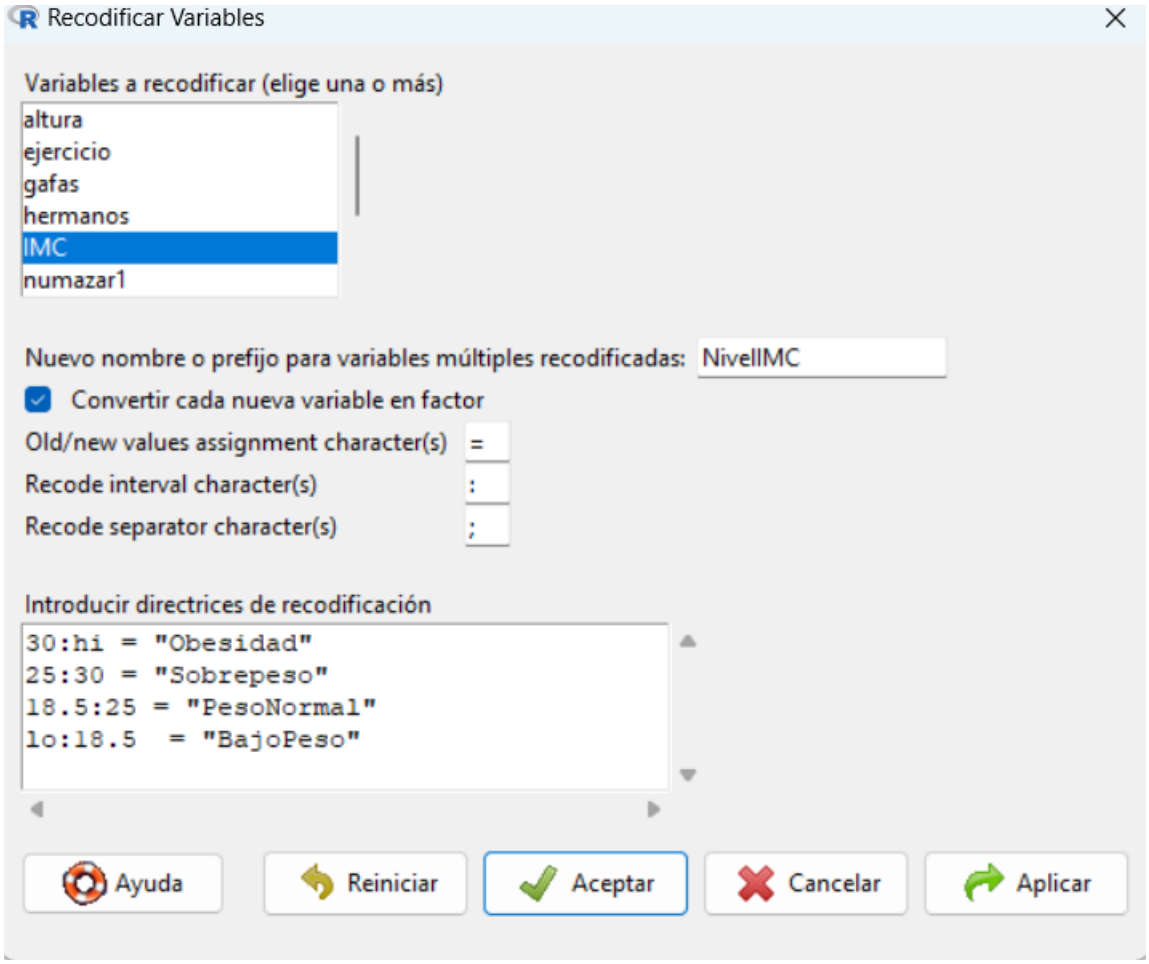


Figura 5.3: Recodificación de los valores de la variable IMC

Ejercicio 10. *Recodifica la variable IMC calculada en el conjunto de datos CCAA0910, asignándole el nombre NivelIMC, como se muestra en la Figura 5.3.*

Ejercicio 11. *El Real decreto 140/2003, del 7 de febrero, por el que se establecen los criterios sanitarios de la calidad del agua de consumo humano* [2] *establece que la cantidad total de HCH no debe superar los* 0.4 *ug/L. Calcula una nueva variable que recoja si en un día de medición la cantidad de HCH detectada en La Peña y en Ardisa ha superado dicho límite. En la recodificación debe aparecer Menor de* 0.4 *o Mayor o igual que* 0.4

6. Ordenar los valores de un factor

R ordena los valores de los factores en orden alfanumérico y cuando estos valores tienen asociado un orden es posible que no coincida por el elegido por R. En estos casos es recomendable reordenarlos utilizando la opción:

Datos → Modificar variables del conjunto de datos activo → Reordenar niveles de factor

[2] BOE nº 145 del 21 de febrero de 2003, pág. 7228-7244

En el cuadro de diálogo que aparece se debe elegir la variable cuyos valores se quieren reordenar, puedes elegir un nombre nuevo (entonces se creará una nueva variable) o mantener el mismo (que es lo más recomendable) y seleccionar la opción *Factor de tipo ordenado*.

Si has elegido mantener el mismo nombre, debes responder *Sí* a la pregunta *¿Sobreescribir la variable?*. En el nuevo cuadro de diálogo debes asignar números a los valores de la variable que respeten el orden deseado.

Ejercicio 12. *Las variables que has obtenido en el ejercicio anterior son de tipo factor con un orden asociado: Menor que* 0.4 *y Mayor o igual que* 0.4 *que no coincide con el orden alfabético que utiliza R. Reordena estos valores manteniendo la misma variable.*

7. Filtrar el fichero de datos

En algunas ocasiones, es interesante estudiar un subconjunto de la población o de la muestra por separado. Aunque R permite hacer estudios por grupos (definidos por una variable) en algunas ocasiones es mejor construir un fichero de datos independiente con la submuestra de interés. Además, esta opción de *Filtrado de ficheros*, permite también eliminar del fichero datos erróneos o atípicos. Para filtrar el conjunto de datos se debe seguir la siguiente instrucción:

Datos → Conjunto de datos activo → Filtrar conjunto de datos activo

Para realizar el filtrado de los casos, es necesario indicar en el cuadro *Expresión de selección* la condición que deben cumplir los datos que queremos extraer. Tras ello hay que indicar un nuevo nombre para el nuevo conjunto de datos.

Vamos a crear, a partir del conjunto de datos *CCAA0910*, un conjunto de datos llamado *Mujeres* que contenga los datos de las estudiantes. La condición que impondremos es que la variable *sexo* tome valor *Femenino* y lo indicamos de la siguiente manera:

$$sexo == "Femenino"$$

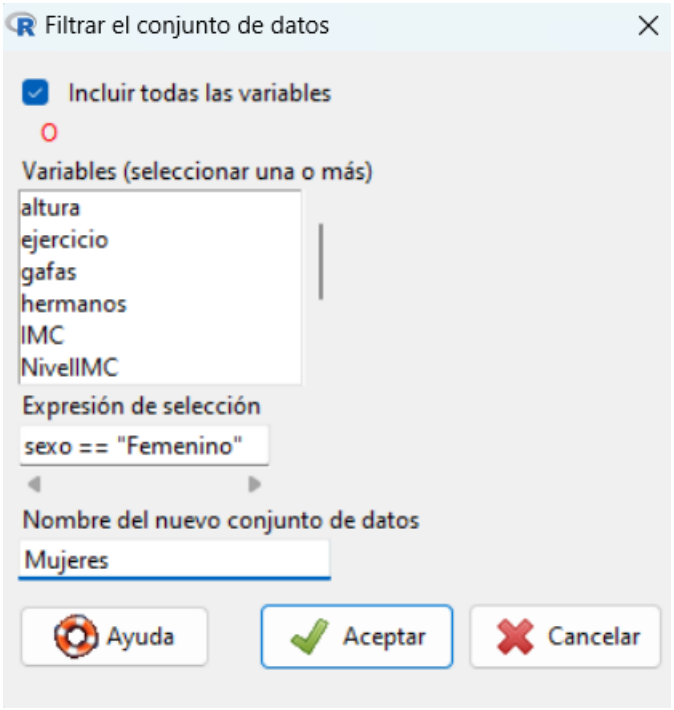


Figura 7.1: Filtrado del conjunto de datos *CCAA0910*

Tras darle a *Aceptar*, se observa que, si no se ha producido ningún error, el nombre del conjunto de datos activo ha cambiado a *Mujeres*, y si lo visualizamos, solo se muestran los registros pertenecientes al sexo femenino. En este momento el conjunto de datos está creado pero no se ha guardado, esto quiere decir, que si cerrásemos R perderíamos este conjunto de datos filtrado.

Notar que en la expresión de selección puede ser necesario utilizar formulas más complejas. A continuación se detallan las expresiones y operaciones lógicas más habituales:

Expresiones lógicas

$<$	Menor	$<=$	Menor o igual
$>$	Mayor	$>=$	Mayor o igual
$==$	Igualdad lógica	$!=$	Distinto

Operadores lógicos

&	Conjunción (y)
|	Disyunción (o)
!	Negación

Otra observación a tener en cuenta, es que cuando la variable es tipo cualitativa (factor) el valor debe ponerse entre comillas, mientras que si la variable es de tipo cuantitativo (numérico) el valor se pone sin comillas.

Ejercicio 13. *Crea un conjunto de datos con los estudiantes que utilizan la talla de camiseta L, tienen dos hermanos y que usan gafas.*

8. Creación de un conjunto/fichero de datos

Por último, veamos cómo introducir un conjunto de daros directamente en R Commander. Tenemos los siguientes datos y los queremos introducir como una variable.

7 5 3 15 8 14 4 4 8 15 12 2 1

Para realizarlo, seleccionamos:

Datos → Nuevo conjunto de datos

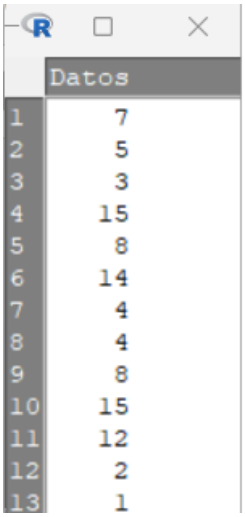
En la ventana que aparece introducimos el nombre del conjunto de datos que vamos a crear, por ejemplo, *DatosPrueba*. Pulsamos en *Añadir columna* hasta tener las columnas que necesitemos y en *Añadir fila* hasta tener las filas que necesitemos. Observa que la primera columna se llama *rowname* y su contenido corresponde al nombre de la fila, que si no se indica lo contrario contiene el orden en la tabla.

Ahora si pulsamos sobre la cabecera de la primera columna, donde aparece *V1* podemos escribir el nombre, *Datos*. Si tuviéramos más columnas deberíamos proceder análogamente introduciendo el nombre correspondiente a cada columna. A partir de aquí se introduce en cada casilla el valor de la tabla. Cabe notar que en R el separador decimal es el punto.

Una vez introducidos los datos, si pulsas *Aceptar* observa cómo el conjunto de datos introducido es el nuevo *Conjunto de datos activo* (recuerda que este conjunto de datos no está guardado en el ordenador).

Ejercicio 14. *Crea el conjunto de datos mencionado anteriormente y guárdalo para poder usarlo posteriormente con el nombre de DatosPrueba.RData.*

Nota: Si visualizas el conjunto de datos una vez creado, deberíais ver algo similar a la siguiente figura:



	Datos
1	7
2	5
3	3
4	15
5	8
6	14
7	4
8	4
9	8
10	15
11	12
12	2
13	1

9. Salir de R Commander

Para salir (de un modo ortodoxo) de R y R Commander se utilizará:

Fichero → Salir → De Commander y R

Previamente el programa preguntará si se desean guardar las ventanas abierta. Si se quiere guardar algún conjunto de datos creado ya se ha indicado cómo hacerlo.

Nota: Si en algún momento se cierra por error la ventana de R Commander, quedando abierta la de R, se puede volver a cargar R Commander ejecutando la instrucción ***Commander()***