Práctica 1. Introducción a R-Commander.

1- Introducción.
2- Menú de R-Commander.
3- Análisis de una variable cualitativa.
4- Histogramas para variables continuas. Detección de valores atípicos.
5- Filtrado de datos.
6- Transformación de variables.
7- Recodificación de variables.
8- Lectura de ficheros en otros formatos. Guardar datos activos.
9- Salir de R-comander.
10- Apéndice 1: información sobre ficheros.
11- Apéndice 2: instalación de R y R-commander.
12- Actividades propuestas (para entrega voluntaria).
13- Ejercicios propuestos.

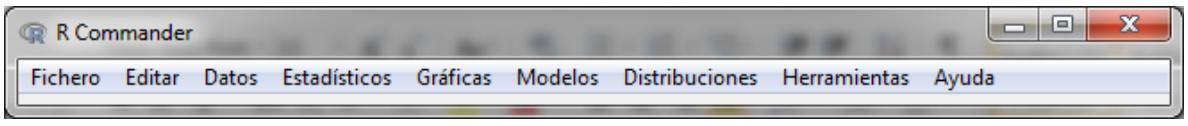
1. Introducción.

El paquete estadístico que vamos a utilizar en la asignatura se llama R-Commander, que es una adaptación de un lenguaje de programación más potente (R) a un sistema de ventanas mediante el cual se pueden realizar de modo sencillo estadísticas. A continuación se presenta una breve introducción a R (basada en el libro libre de R y R-commander http://knuth.uca.es/moodle/course/view.php?id=37 , capítulo 1). R-Commander ya lo tenemos instalado en las aulas de informática, pero es un software libre que se puede instalar en cualquier ordenador.

R es un lenguaje de programación y un entorno para el análisis estadístico de datos y la realización de gráficos. Debido a su naturaleza es fácilmente adaptable a una gran variedad de tareas. Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda. Actualmente R es el resultado de un esfuerzo de colaboración de personas del todo el mundo y abarca una amplia gama de técnicas estadísticas, que van desde los modelos lineales tradicionales a las más modernas técnicas de clasificación, pasando por los test clásicos y el análisis de series temporales. R proporciona una amplia gama de gráficos, que además son fácilmente adaptables y extensibles. La calidad de los gráficos producidos y la posibilidad de incluir en ellos símbolos y fórmulas matemáticas, posibilitan su inclusión en publicaciones que suelen requerir gráficos de alta calidad. El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL. Además está disponible precompilado para una multitud de plataformas. La página principal del proyecto es http://www.r-project.org.

2. Menú de R-Commander.

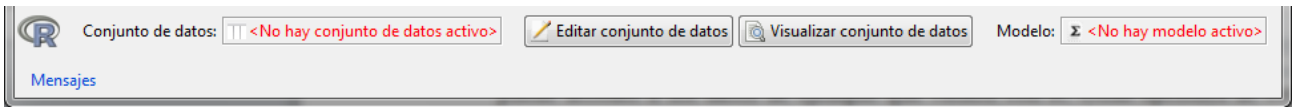
Si pulsamos sobre el icono de R (en el menú de inicio), entramos en la aplicación. En la ventana de R-Comander, observamos una primera franja horizontal que contiene el menú principal. Pulsando sobre las diferentes opciones se despliegan los correspondientes submenús.



Algunas de las opciones del menú principal son:

- **Fichero**: Permite guardar las instrucciones y los resultados de una sesión de trabajo. Además permite terminar la aplicación.
- **Editar**: Contiene las opciones habituales relacionadas con la edición: "Cortar", "Copiar", "Pegar", "Borrar", "Buscar...", "Seleccionar todo", "Deshacer", "Rehacer", "Limpiar ventana".
- **Datos**: Mediante las opciones de este menú se pueden cargar, editar y guardar datos. Además se puede acceder a los datos de ejemplo que vienen con R. Otras opciones de este mismo menú permiten operaciones con los datos, como por ejemplo: recodificación, tipificación, construcción de nuevas variables, ...
- **Herramientas**: Mediante este menú se pueden cargar paquetes y plugins, y se puede configurar la interface.
- **Ayuda**: Muestra ayuda sobre R-Commander.

La segunda franja horizontal contiene la barra de herramientas. Dicha barra muestra información sobre los datos y el modelo en uso.

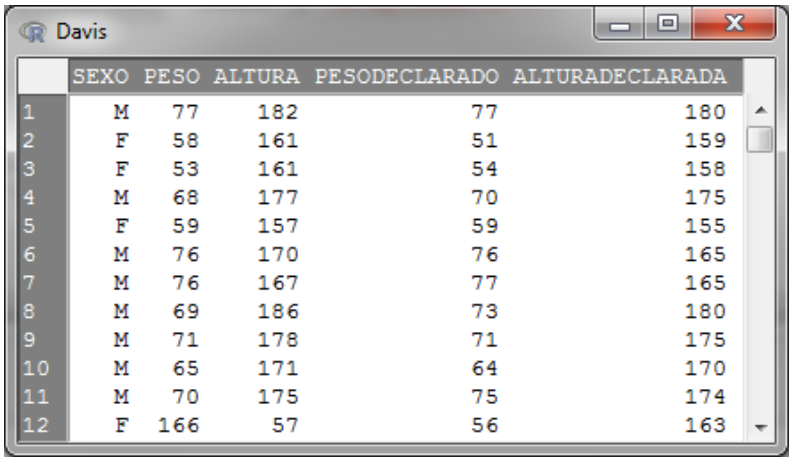


De izquierda a derecha se encuentran cuatro botones:

- **Conjunto de datos**: Este botón muestra el nombre del conjunto de datos activo o la leyenda "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 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 ventana de edición de datos antes de intentar cualquier otra operación.
- **Visualizar conjunto de datos**: Muestra una ventana con los datos del conjunto de datos activo en formato similar al anterior. Esta ventana no permite la modificación de los datos, pero puede mantenerse abierta mientras se continúa haciendo operaciones.
- **Modelo**: La leyenda muestra el nombre del modelo activo o la leyenda "No hay modelo activo" cuando no se ha construido ningún modelo previamente. La pulsación sobre dicho botón permite la selección del modelo en uso de entre los disponibles.

Además del menú principal y la barra de herramientas, observamos tres cuadros. El primero es el cuadro de instrucciones (***R Script)***. Allí van apareciendo todos los comandos o instrucciones de R que se van ejecutando cuando realizamos las diferentes selecciones en los menús. Esto es algo que realiza R-Commander automáticamente. Seleccionando cualquier instrucción de ese cuadro, se puede volver a ejecutar dicha instrucción pulsando el botón "Ejecutar". En el segundo, ***Salida***, van apareciendo los resultados de las diferentes selecciones. En el tercero, ***Mensajes***, aparecen diferentes notas, avisos o errores.

Para comenzar, vamos a abrir el fichero Davis.rda. La extensión rda corresponde a un fichero R (no importa que esté en mayúsculas o minúsculas). Para ello, vamos al menú **Datos>cargar conjunto de datos** y en la carpeta correspondiente seleccionamos Davis.rda. Una vez ejecutado, pulsamos al botón 'visualizar conjunto de datos' y se abrirá una ventana como la siguiente.



	SEXO	PESO	ALTURA	PESODECLARADO	ALTURADECLARADA
1	M	77	182	77	180
2	F	58	161	51	159
3	F	53	161	54	158
4	M	68	177	70	175
5	F	59	157	59	155
6	M	76	170	76	165
7	M	76	167	77	165
8	M	69	186	73	180
9	M	71	178	71	175
10	M	65	171	64	170
11	M	70	175	75	174
12	F	166	57	56	163

Este es el aspecto habitual de un conjunto de datos. En las filas aparecen los casos (en este caso individuos) y en las columnas las variables estudiadas. En particular, este fichero proviene de un estudio realizado a personas que practican deporte de modo regular (más información en el apéndice). Como se nos indica en la ventana de los mensajes, el fichero consta de 200 filas (casos) y de cinco columnas (variables). La primera variable es el sexo del individuo (**M**asculino o **F**emenino), las dos siguientes representan mediciones de su peso y altura (en kg y cm, respectivamente). Las dos últimas columnas son el peso y la altura que el individuo declaró que tenía cuando se le preguntó (en kg y cm, respectivamente).

El fichero por tanto consta de una variable cualitativa (sexo) y el resto son variables cuantitativas continuas.

Notas:

- En algunas variables, aparecen algunas casillas con NA. Para R-Commander, éstos son datos que no se conocen (datos perdidos, vienen de las siglas inglesas Not Available)

- A lo largo de las prácticas utilizaremos ficheros de datos ya preparados. Sin embargo es interesante saber que R-Commander, en su menú **Datos,** permite, a través de la opción **Importar datos**, importar de modo sencillo datos contenidos en ficheros creados por otros programas (Excel o SPSS, por ejemplo), y que hay disponibles un gran volumen de ficheros a los que podemos acceder utilizando la opción **Conjunto de datos en paquetes.** Finalmente, podemos crear nosotros un nuevo fichero a través de la opción **Nuevo conjunto de datos.**

3. Análisis de una variable cualitativa.

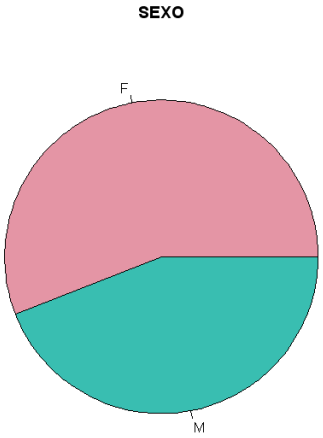
Distribución de frecuencias. La variable sexo es la única variable cualitativa de nuestro estudio. Para analizarla, necesitamos saber su distribución de frecuencias. Para ello, vamos al menú **Estadísticos> Resúmenes>Distribución de frecuencias**. Nos aparecen en la *Salida* las frecuencias absolutas y los porcentajes:

```
counts:
SEXO
 F  M
112 88

percentages:
SEXO
 F  M
56 44
```

(es decir, hay 112 mujeres y 88 hombres, representando el 56% y 44% de la muestra, respectivamente)

Representación gráfica. Su representación se realiza por medio de un diagrama de barras o de sectores. Ambos diagramas aparecen en el menú **Gráficas** (**Gráfica de barras** y **Gráfica de sectores**, respectivamente). Por ejemplo, vamos a realizar un diagrama de sectores con la opción **Gráfica de sectores**. El diagrama aparece en la ventana R graphics y tiene el siguiente aspecto:

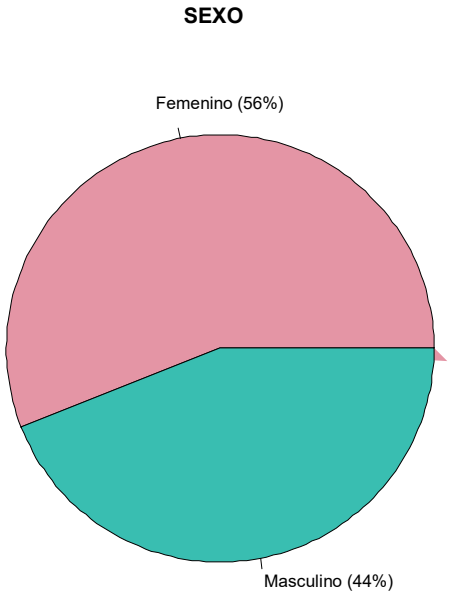


Nota 1: obsérvese que cada vez que ejecutamos una instrucción del menú, su código se añade a la ventana de instrucciones (R Script). En este caso, la instrucción que aparece es la siguiente:

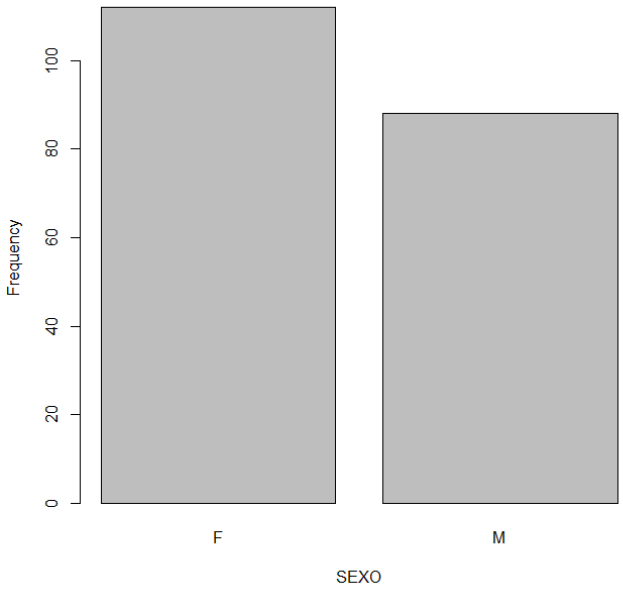
```
with(Davis, pie(table(SEXO), labels=levels(SEXO), xlab="", ylab="", main="SEXO",
  col=rainbow_hcl(length(levels(SEXO)))))
```

Si supiéramos R, podríamos modificarla (y pulsar el botón ejecutar), para cambiar el aspecto del dibujo

Nota 2: Es interesante ir guardando los gráficos que vamos obteniendo, porque cada vez que se hace un gráfico se pierde la información sobre el anterior. Ello lo podemos hacer a través del menú de la ventana de gráficos **Archivo>Guardar como**. No obstante, es recomendable tener abierto un procesador de textos (Word por ejemplo) e ir seleccionando y copiando lo que nos interese guardar. Por ejemplo, si queremos copiar el gráfico de sectores en Word podemos hacer el clásico ctrl+C, ctrl+V, o, si queremos modificar el gráfico en Word (cambiar etiquetas, o poner título, por ejemplo) podemos elegir en la ventana de R Graphics la opción **Archivo>Copiar para el área de transferencia>como un Metafile** (ctrl+W).

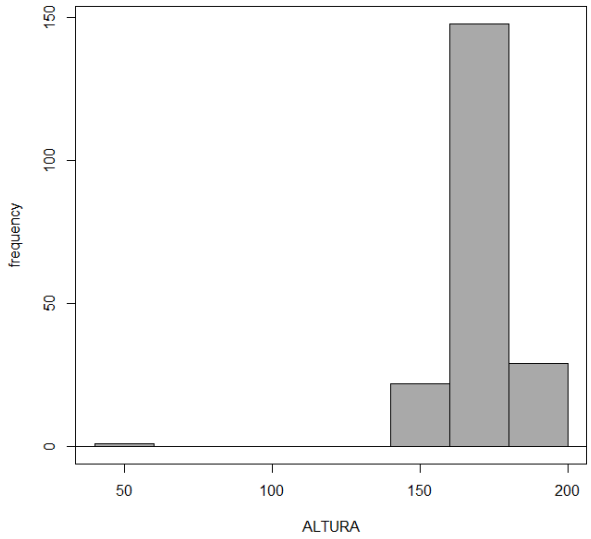


Si en lugar del gráfico de sectores, se hubiese elegido el gráfico de barras, el resultado que se hubiese obtenido sería el siguiente:



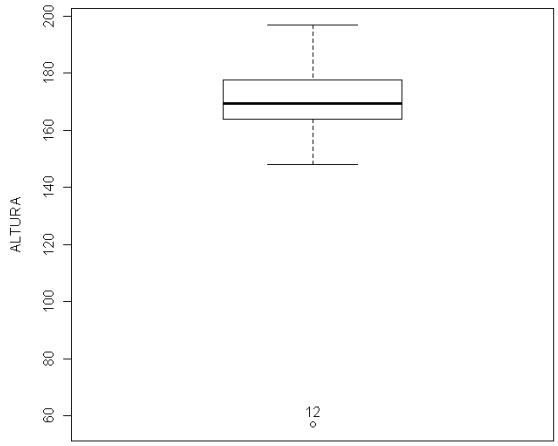
4. Histogramas para variables continuas. Detección de valores atípicos.

El resto de las variables de nuestro fichero son continuas, por lo que su representación gráfica se realiza mediante un histograma (**Gráficas>Histograma**). El histograma para la variable altura quedaría de la siguiente forma:

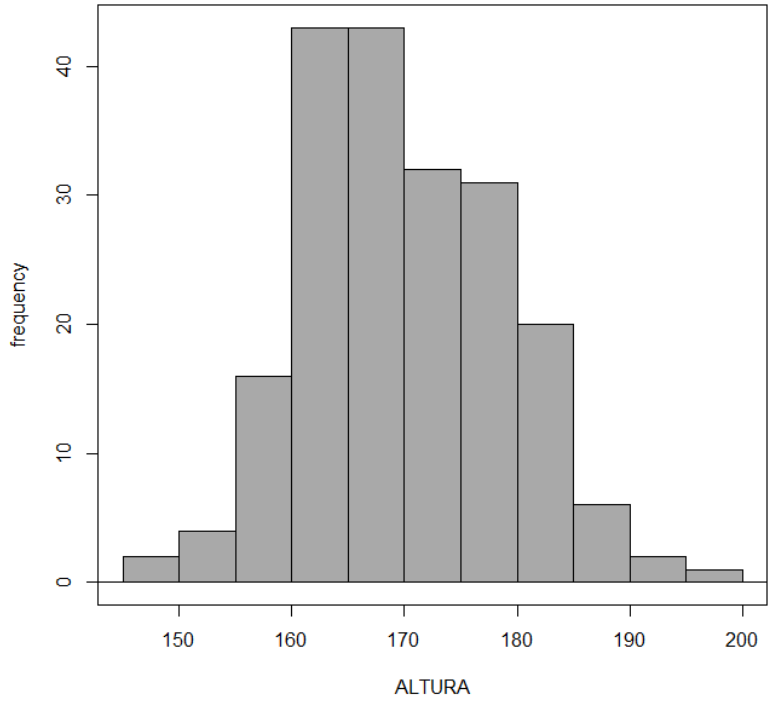


El aspecto resulta 'sospechoso'. Obsérvese que es imposible que la altura de una persona esté en el intervalo de 50 centímetros. Ello ha hecho que la división automática de las clases 'concentre' los 199 datos restantes en 3 intervalos. Deberíamos detectar ese dato, y ver si es un posible error de registro. Es posible buscarlo en el fichero de datos, pero con muchos datos, la búsqueda puede ser bastante tediosa. Otra opción

es dibujar un diagrama de caja (**Gráficas>Diagrama de caja**). Su construcción la explicaremos en clase más adelante. Seleccionamos la variable altura y obtenemos el siguiente gráfico.



El gráfico nos indica que la mayoría de los datos se concentran entre las alturas 150-200 cm, y que hay un dato atípico (en este caso imposible), que es el número 12. Para ver qué sucede con ese dato, vamos a "**Editar conjunto de datos**" y observamos ese caso. Es obvio que al teclear esos datos se ha 'intercambiado' el peso con la altura. Modificamos los datos intercambiados y salimos del editor. Al realizar otra vez el histograma, ahora queda de la siguiente forma, con una distribución de las alturas más natural.



Observamos que la mayor parte de las alturas se acumulan entre 155 y 185 cm. Notamos una bimodalidad (hay dos intervalos de alturas, el (160,165] y el (165,170], que son los que más valores acumulan y están empatados a número de observaciones). Nótese que R, por defecto, dibuja los histogramas con los intervalos abiertos por la izquierda y cerrados por la derecha, aunque esta opción se podría cambiar. La bimodalidad será probablemente debida a la 'mezcla' de alturas masculinas y femeninas. Observamos

también una cierta asimetría positiva (si situamos el centro en la mitad de los datos, las observaciones se extienden más hacia la derecha que hacia la izquierda de dicho centro).

Nota 1: Al eliminar el dato atípico en el editor de datos, hemos hecho modificaciones sobre el fichero inicial que nos puede interesar guardar (las modificaciones realizadas en el editor no se guardan automáticamente en el fichero inicialmente abierto). En cualquier momento podemos guardar un fichero activo en **Datos>Conjunto de datos activo>Guardar el conjunto de datos activos…,** y podemos elegir el mismo nombre (y así sustituimos el inicial) o darle otro nombre (por ejemplo, para no confundirnos, podemos guardar el fichero con las modificaciones como Davis2).

Nota 2: Con frecuencia nos interesa realizar un análisis sobre un subconjunto de datos del fichero, por ejemplo, nos puede interesar realizar estudios separados para hombres y mujeres. Muchas opciones, como iremos viendo, permiten introducir como factor una variable (en este caso el sexo) y automáticamente realizan estudios separados para cada nivel de dicha variable. En el caso de los histogramas, es posible realizar directamente un estudio por separado según los diferentes niveles de un factor. Por ejemplo, en el cuadro de diálogo asociado al histograma, basta pulsar el botón **Gráfica por grupos** y elegir el factor que interese (en este caso SEXO) para obtener histogramas por separado según los niveles del factor.

Nota 3: Como hemos indicado, por defecto R dibuja histogramas con intervalos (a,b]. Si se quisiese modificar esa opción y dibujar el histograma con intervalos [a,b), habría que realizar lo siguiente. Al obtener el histograma de la variable ALTURA, el comando R que se ha ejecutado y que aparece en la ventana de instrucciones es el siguiente.

```
with(Davis, Hist(ALTURA, scale="frequency", breaks="Sturges", col="darkgray"))
```

El comando se modificaría de la siguiente forma.

```
with(Davis, Hist(ALTURA, scale="frequency", breaks="Sturges", right=FALSE, col="darkgray"))
```

Seleccionando esta última instrucción y pulsando el botón **Ejecutar**, aparece un histograma con intervalos de la forma indicada.

Nota 4: El dibujo de un histograma o un diagrama de cajas tiene sentido cuando el número de datos con el que trabajamos es grande. En otro caso, es mejor utilizar representaciones gráficas alternativas, donde cada dato aparezca representado individualmente. Una buena opción sería utilizar el llamado Diagrama de puntos (**Gráficas>Diagrama de puntos**). Puedes probar a representar gráficamente la altura según sexo mediante esta opción.

5. Filtrado de datos.

Partiendo del fichero original, se trata de crear otros ficheros que contengan una parte de los datos. Vamos a considerar, por ejemplo, los datos relativos a las mujeres. Para seleccionar las filas de las mujeres, vamos a **Datos>Conjunto de datos activo> Filtrar conjunto de datos activo.** En el cuadro de diálogo que aparece, hay que tener en cuenta lo siguiente.

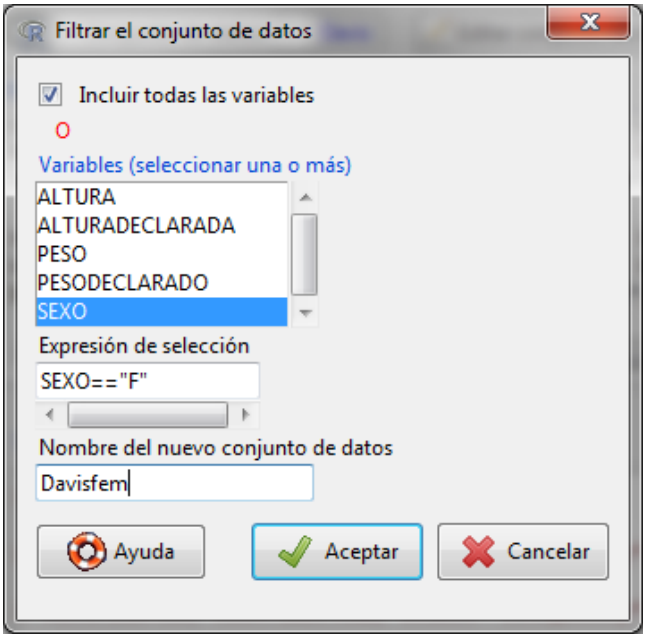
- En el cuadro “Expresión de selección” se pone SEXO==”F” para indicar que sólo se quieren los casos donde la variable SEXO es igual a F.

- Si en la casilla ‘Nombre del nuevo conjunto de datos” se pone un nombre, se generará un nuevo conjunto de datos (en este caso Davisfem) con los datos seleccionados.

- Hay que tener cuidado con las mayúsculas y las minúsculas. Para R son diferentes.

- “F” va entre comillas, porque es un código (alfanumérico) y no un número. Cuando se utilizan variables numéricas, los números van sin comillas.

- Si se desea copiar todas las variables, hay que marcar la casilla “Incluir todas las variables”. En otro caso, hay que seleccionar las variables que se desea copiar con la combinación del ratón y la tecla Ctrl.



Después de pulsar “Aceptar” en el cuadro de diálogo anterior, ahora tenemos dos conjuntos de datos activos: el original (Davis) y el nuevo (Davisfem). Para pasar de uno a otro, se utiliza el botón “Conjunto de datos” en la barra de herramientas.

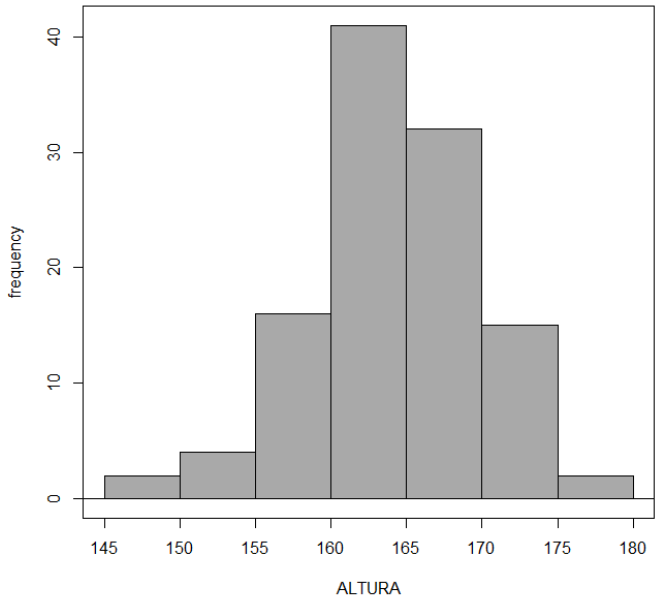
Nota 1: Si queremos conservar el fichero Davisfem para posteriores sesiones, lo deberíamos guardar, como se ha indicado anteriormente con **Datos>Conjunto de datos activo> Guardar el conjunto de datos activos.**

Nota 2: En expresión de selección se podrían utilizar más opciones, según los datos que queramos filtrar. Se detallan a continuación las expresiones y operaciones lógicas más habituales:

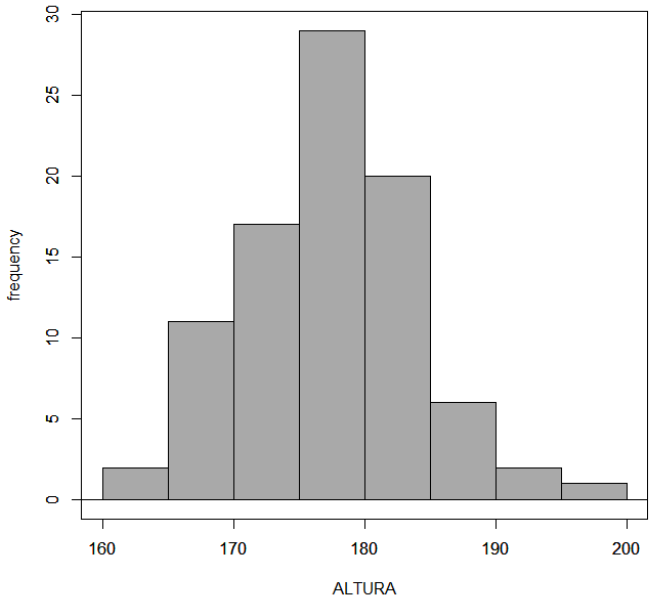
Expresiones lógicas	
<	Menor
<=	Menor o igual
>	Mayor
>=	Mayor o igual
= =	**Igualdad lógica**
!=	Distinto
Operadores lógicos	
&	Conjunción (y)
|	Disyunción lógica (o)
!	Negación

Por ejemplo, si quisiéramos dibujar un histograma de chicas con altura inferior o igual a 170, utilizaríamos la siguiente expresión de selección: SEXO= =”F” & ALTURA<=170

Teniendo seleccionado el conjunto Davisfem, dibujamos el histograma de las mujeres.



Observamos que la moda (intervalo que más se repite) está en las alturas 160-165 y que hay una leve asimetría a izquierda (o negativa) en los datos. De modo análogo, volviendo al fichero Davis, se podría crear el conjunto de datos de los varones (Davismasc) y dibujar su histograma.



Se observa que la moda está en las alturas 175-180, y que parece haber una leve asimetría positiva (o a derecha) en los datos.

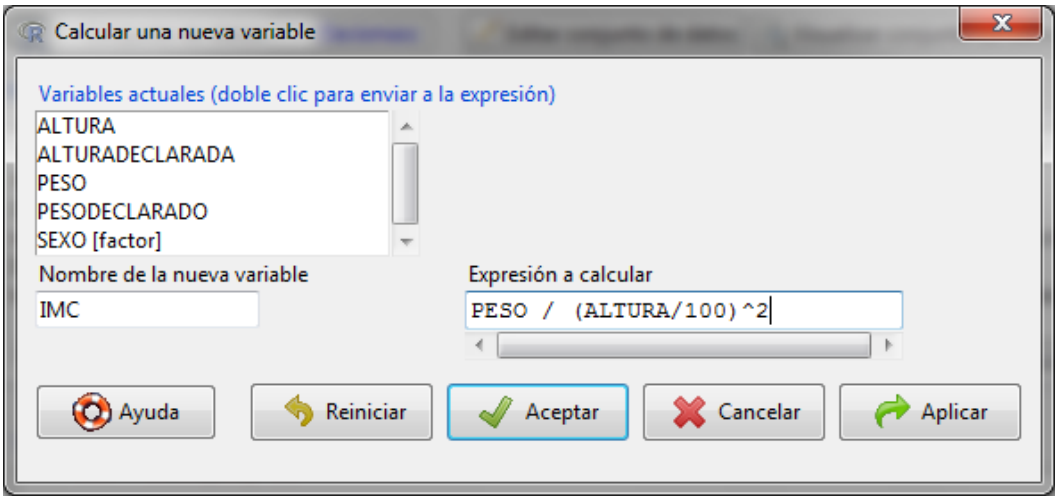
Nota 3: En ocasiones, suele ser de interés cambiar el número de intervalos y sus longitudes en un histograma. Para ello, basta con sustituir en el comando que genera el histograma (comando Hist) la opción breaks=Sturges", por la opción breaks=c(p_1, p_2, …, p_k), donde los valores p_i los elegimos nosotros. Así, la opción breaks=c(40, 80, 120, 160, 200) generaría cuatro intervalos, con longitud 40 cada uno de ellos, comenzando en el valor 40 y terminando en el valor 200. Está opción se puede utilizar cuando se quieren representar dos histogramas con la misma escala en el eje de abscisas. Si hay muchos intervalos, resultaría más cómoda la opción breaks=seq(40, 200, by=40).

6. Transformación de variables.

A veces, resulta interesante crear y analizar nuevas variables de interés que se pueden construir a partir de las ya existentes. Por ejemplo, una medida de interés que relaciona peso y altura es el índice de masa corporal (IMC), que no presenta tanta dependencia del sexo del individuo como ocurre con los pesos y las alturas, y que se define como (peso medido en kg y altura en metros): IMC=peso/altura^2.

A través del IMC se definen los parámetros 'oficiales' de población con sobrepeso o con peso inferior al normal, para más información véase http://www.who.int/mediacentre/factsheets/fs311/es/index.html.

Vamos a definir esta nueva variable en el fichero Davis2 (recordad que es fichero original Davis con la modificación del peso y altura del caso 12). Para ello, vamos a **Datos>Modificar variables del conjunto de datos activo>Calcular una nueva variable** y rellenamos el cuadro de diálogo que nos aparece de la siguiente forma.



Las variables a incluir en "Expresión a calcular" se pueden teclear directamente o seleccionar con el ratón en el cuadro superior haciendo doble click sobre la variable. Los signos aritméticos son los habituales. Nótese que hemos dividido la altura por 100 para pasarla de cm a metros.

La nueva variable se habrá incorporado a nuestro fichero y ahora la podremos utilizar en nuestro estudio.

Nota: En éste y otros cuadros de diálogo, observareis un botón con la opción de "Reiniciar". Si presionamos sobre él, R coloca en los campos del cuadro de diálogo las opciones que tiene establecidas por defecto.

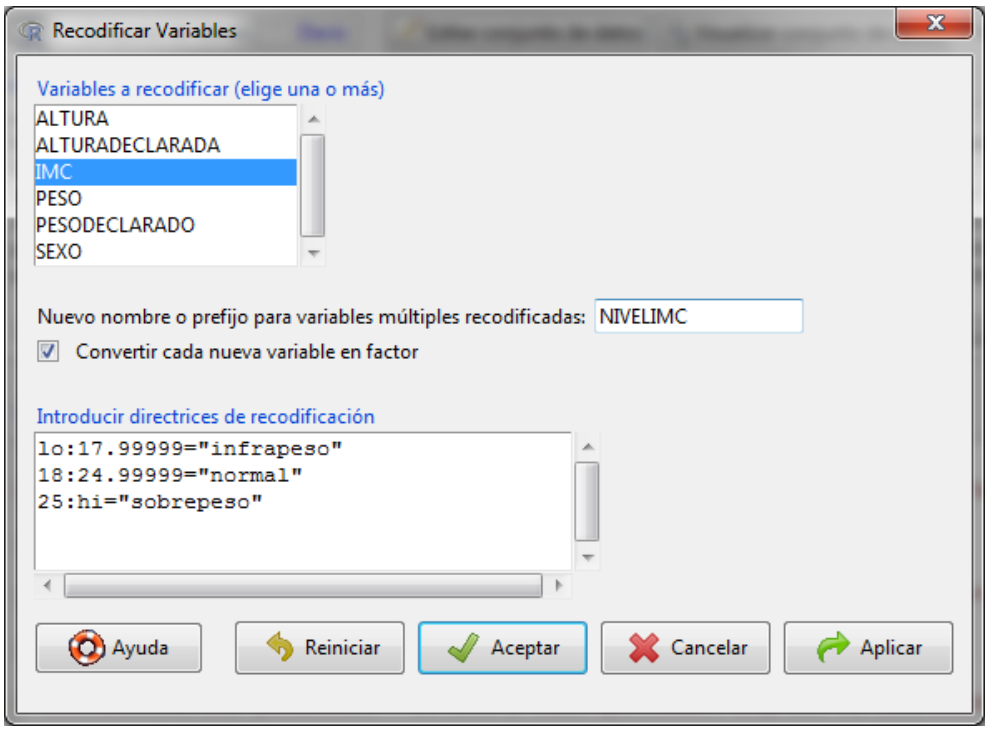
7. Recodificación de variables.

En ocasiones tenemos una variable cuantitativa y la queremos clasificar según varias categorías. Como ejemplo, vamos a utilizar la variable índice de masa corporal calculada en el ejercicio anterior, y que se encuentra en el fichero Davis2. La clasificación de personas según su IMC es la siguiente:

Valores IMC	Categoría
Menos de 18	infrapeso
[18,25)	normal

25 y más	sobrepeso

Vamos a crear una variable (podemos llamarla NIVELIMC), que para cada persona indique la categoría a la que pertenece (infrapeso, normal o sobrepeso). Para ello, tenemos que ir a **Datos>Modificar variables del conjunto de datos activo> Recodificar variables**. Se rellenaría el cuadro de diálogo de la siguiente manera (*activamos "Convertir cada nueva variable en factor", para indicar que el resultado es una variable cualitativa)*:



Explicación
a:b es la forma de indicar el intevalo **[a,b]**
lo:a la expresión **lo** indica el límite inferior de la variable. Así pues, estamos indicando el intervalo **[lo,a]**
b:hi la expresión **hi** indica el límite superior de la variable. Así pues, estamos indicando el intervalo **[b,hi]**
Para más posibilidades de recodificación, consultar las notas abajo, o bien la ayuda del cuadro de diálogo (está en inglés).

Notas:

- Según la tabla informativa sobre el sobrepeso, debemos realizar intervalos de la forma [a,b). Sin embargo, la opción **a:b** realiza intervalos cerrados en ambos extremos. Como en el fichero aparecen los datos con 5 cifras decimales, para indicar el primer intervalo (que correspondería a menos de 18, sin incluir este último), hemos utilizado **lo:17.9999**. Para el resto de los intervalos, se procede de la misma forma.
- Cuando un valor aparece clasificado en dos grupos, se toma como valor 'recodificado' el del grupo en el que aparece en primer lugar.
- Se puede recodificar un solo valor, en vez de un conjunto de valores. Por ejemplo, si queremos sustituir una altura 'rara' de 57 centímetros por un dato perdido, pondríamos 57=NA.
- Para el último grupo se puede usar la opción **else**, esto es, la tercera línea de las directrices la podríamos haber sustituido por:

  NA=NA

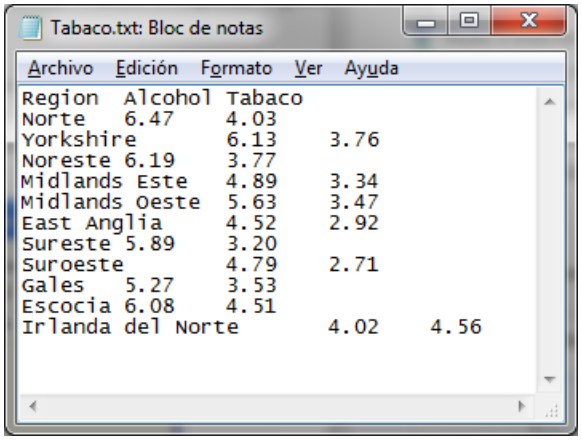
else=”sobrepeso”

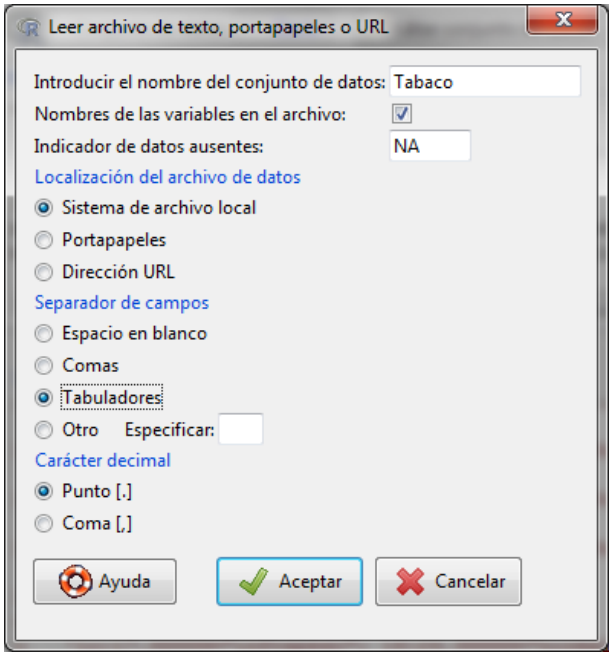
De esta manera, indicaríamos primero que los valores perdidos se dejen tal cual y que los valores restantes, después de haber hecho los dos primeros grupos, serían los individuos con “sobrepeso” (pero hay que tener cuidado con esta instrucción y estar seguros de incluir en else sólo a las personas con sobrepeso).

8. Lectura de ficheros en otros formatos.

R permite la lectura de ficheros de texto (txt), ficheros creados en una base de datos (Excel, por ejemplo) o ficheros adaptados a otros paquetes estadísticos (SPSS, MINITAB, …). Vamos a explicar alguna de estas opciones.

Lectura de datos desde un fichero de texto (txt, por ejemplo). Tenemos en un fichero .txt (Tabaco.txt) los datos relativos a consumo de tabaco y alcohol en diferentes regiones de Gran Bretaña (promedios semanales por vivienda, en libras, año 1981). Haz doble click sobre el fichero de datos y observa que su aspecto es el siguiente.





Para leer dicho fichero en R-Commander, se va a la opción **Datos>Importar datos>desde archivo de texto, portapapeles o URL**. Teniendo en cuenta que vamos a llamar al nuevo conjunto de datos Tabaco, que el fichero txt lleva los nombres de las variables en la primera línea, que los datos están separados por tabuladores y que los separadores decimales son puntos, el cuadro de diálogo correspondiente se rellenaría como indica la figura superior. Posteriormente, se puede guardar el fichero en formato R mediante la opción habitual.

Lectura de datos desde un fichero Excel. El fichero Alumnos12.xls, contiene los datos relativos a la encuesta del primer día de clase, realizada el año pasado. Para abrirlo en R, utilizar la opción **Datos> Importar datos> desde un archivo de Excel.** Únicamente hay que dar un nombre al conjunto de datos que se va a crear, seleccionar el fichero Excel a leer, e indicar la hoja de dicho fichero donde se encuentran los datos. Si se ha tenido la precaución de definir el nombre de las variables en la primera fila (como es el caso, procurando no introducir caracteres extraños), R-Commander tomará estos nombres como nombres de las variables.

Cuando leas un fichero en otro formato, te resultará conveniente guardarlo como archivo de R-Commander, si lo quieres utilizar posteriormente. Recuerda la opción **Datos>Conjunto de datos activo> Guardar el conjunto de datos activo** para guardar un conjunto de datos.

9. Salir de R commander.

Para salir (de modo ortodoxo) de R y R-commander se utilizará la opción **Fichero> Salir> De Commander y R.** Previamente el programa preguntará si se desean guardar las ventanas abiertas. Si se quiere guardar algún conjunto de datos creado en la sesión (antes de salir de ella), ya se ha indicado cómo hacerlo **(**utilizando **Datos> Conjunto de datos activo> Guardar el conjunto de datos activo,** asegurándose de que el fichero que queremos guardar está seleccionado con el botón "Conjunto de datos"). Otra opción interesante es la de **Fichero> Guardar el entorno de trabajo de R como,** que nos permitiría guardar todos los subficheros abiertos en la sesión en un único fichero.

Nota: Si en algún momento se cierra por error la ventana de R commander, quedando abierta la de R, se puede volver a cargar el R-Commander ejecutando en la R-Console (la ventana que apenas utilizamos) la instrucción: >Commander().

10. Información sobre ficheros.

Información sobre el fichero Davis (extraído y modificado del paquete car, incluido en R).

- **Descripción**: contiene 200 filas y 5 columnas. Los individuos son hombres y mujeres que practican ejercicio de manera regular. Hay algunos datos perdidos.

- **Formato**: este fichero contiene las siguientes variables:

    - Sexo: Un factor con los niveles: F, femenino; M: masculino.
    - Peso: peso medido en kg.
    - Altura: altura medida en cm.
    - Pesodeclarado: lo que cada persona dijo que pesaba, en kg.
    - Alturadeclarada: lo que cada persona dijo que medía, en cm.

Referencias

[1] Davis, C. (1990) Body image and weight preoccupation: A comparison between exercising and non-exercising women. *Appetite*, **15**, 13–21.

[2] Fox, J. (1997) *Applied Regression, Linear Models, and Related Methods.* Sage.

Información sobre el fichero Wesrd

- **Descripción**: contiene 669 datos, y cuatro variables. Esta información proviene del denominado estudio epidemiológico de retinopatía diabética de Wisconsin, EE.UU. (Wisconsin Epidemiological Study of Diabetic Retinopathy). En él, se estudió una cohorte (grupo) de diabéticos recibiendo atención médica en la región de Wisconsin. Una descripción detallada de los datos puede verse en [1]. Al principio del estudio se midieron un gran número de variables que pueden tener relación con

el desarrollo de la enfermedad (en nuestro fichero sólo aparecen algunas de ellas), y al final del estudio se comprobó si la enfermedad se había llegado a desarrollar.

- **Formato**: este fichero contiene las siguientes variables:

  - dur: el tiempo que llevaba la persona con diabetes al principio del estudio.
  - gly: el porcentaje de **hemoglobina glicosilada** (HbA1C) en la sangre (hemoglobina que se ha adherido a la glucosa). 2.5-3.5% es normal para los no diabéticos y 6.5% es considerado generalmente buen control para los diabéticos.
  - imc: indice de masa corporal al principio del estudio (un imc entre 18 y 25 se considera normal).
  - Ret: nos dice si el diabético al final del estudio desarrolló retinopatía (1) o no (0).

Referencias

[1] Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1988). Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy, *Journal of the American Medical Association* **260**: 2864-2871.

11. Instalación de R-Commander.

En este apartado se indican las instrucciones para instalar R-Commander en Windows (para otras plataformas, véase http://knuth.uca.es/repos/ebrcmdr/pdf/actual/ebrcmdr.c1.pdf). La dirección web donde se puede conseguir el programa es http://knuth.uca.es/R/doku.php?id=instalacion_de_r_y_rcmdr:r-uca. En dicha página podemos descargar la versión que queramos (se recomienda instalar la versión R-UCA 3.4.1, que es igual o similar a la que tenemos en el aula de informática). Al elegir con el ratón R-UCA-3.4.1.exe en esa página web, se nos descargará el ejecutable de dicha versión en nuestro ordenador.

- Procedemos a la instalación ejecutando el fichero ejecutable recién descargado. En el proceso de instalación, es conveniente dejar todas las opciones por defecto.

- Al abrir por primera vez el paquete, al final de la ventana de R-Commander (mensajes) nos aparece lo siguiente:
  [1] NOTA: Versión de R Commander….
  [2] AVISO: La versión de Windows de R Commander funciona mejor bajo RGui con la interface de documento único (SDI); vea ?Commander

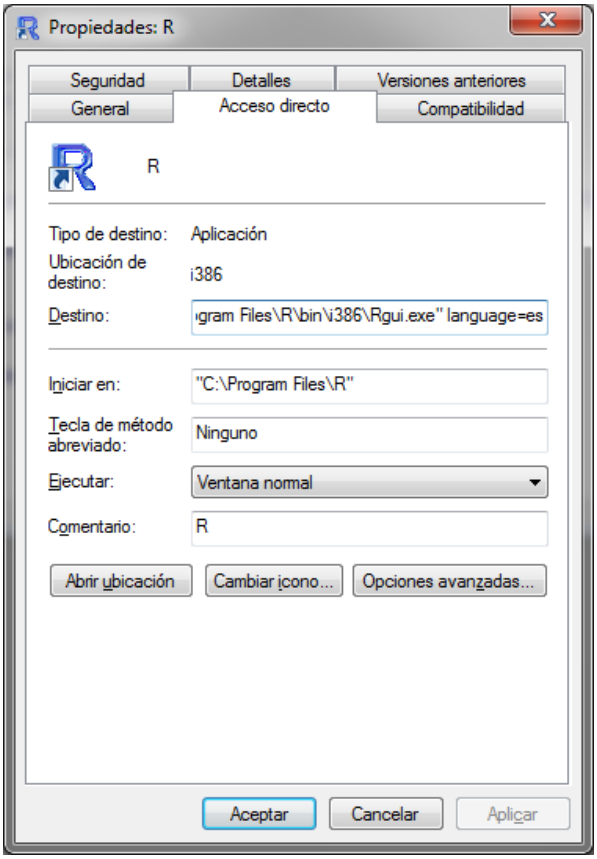
- Si la interface de documento único no viene puesta por defecto, es conveniente ponerla. Para ello, buscar el icono del programa R en el menu de inicio .

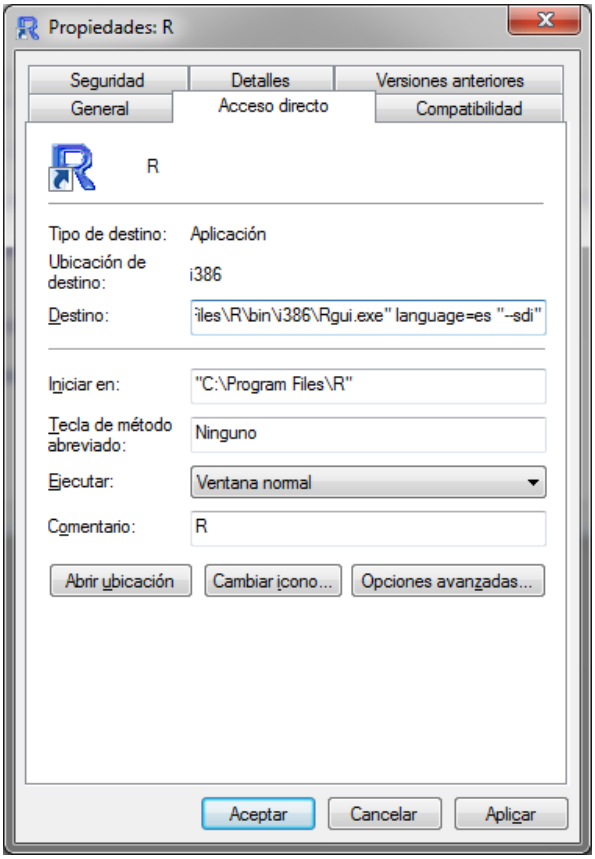
- Nos situamos en dicho icono y vamos a 'Propiedades' mediante el botón derecho del ratón.

- En la casilla Destino, al final debería poner sdi. Si no lo pone, al final del todo, (sin modificar nada de lo que hay), añadir lo siguiente: --sdi (dejando un espacio entre el final del texto ya escrito y lo que se quiere añadir).

- Pulsamos aceptar y ya tenemos el programa listo para su uso.

Nota: Si se quiere comprobar que esta operación se ha realizado correctamente, en el menú de R-Console, opción **Editar>Preferencias de la interface gráfica**, debe aparecer activada en la primera línea la opción SDI.





12. Actividades propuestas (para entrega voluntaria).

La actividad se puede hacer individualmente o en grupos de dos. Este ejercicio continuará en la práctica siguiente. Se pueden hacer todos o algunos de los apartados propuestos.
En el fichero Excel Alumnos, aparecen los datos de la encuesta realizada el primer día de clase, para este curso y cursos anteriores.

1- Lee este fichero en R-Commander. Selecciona, además:
- Una variable cualitativa. Estúdiala según el apartado 3.
- Una variable numérica continua. Represéntala gráficamente, en global y separada según alguna de las variables del fichero que pueda resultar de interés.
  Nota: Cada uno de vosotros elegirá la variable continua a estudiar siguiendo la siguiente regla. Si tu NIP acaba en 1 o 2, elige la variable Internet (Bloque 1). Si tu NIP acaba en 3 o 4, elige la variable tiempo en llegar a la universidad (Bloque 1). Si tu NIP acaba en 5 o 6, elige la variable Nota Univers. (nota de acceso a la Universidad, bloque 2). Si tu NIP acaba en 7 u 8 elige la variable Mejor Nota (mejor nota de primero, bloque 2). Si tu NIP acaba en 9 o 0 elige la variable NotaMatematicas (Bloque 2).
- Una variable numérica discreta. Represéntala gráficamente.

  Nota: Si se hace en grupo, cada persona utilizará una variable diferente en cada apartado.

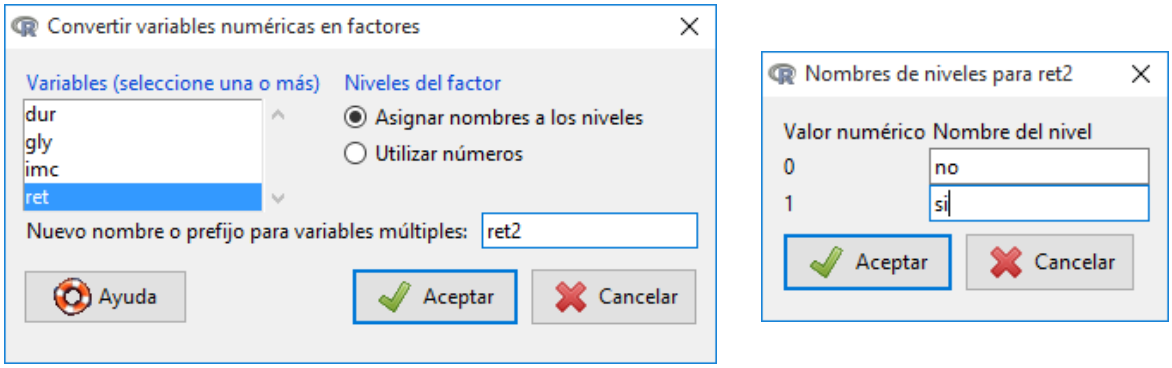
2- Añade en el fichero alguna transformación o recodificación de variable que pienses puede resultar de interés. Estudia dicha variable.
  Nota: Si se hace en grupo cada persona elegirá una transformación/recodificación diferente.

13. Ejercicios propuestos.

1) Abre el fichero Wesrd. Este fichero proviene de un estudio realizado a personas diabéticas, a las que se les realizó un seguimiento para observar cuántas de ellas desarrollaban una enfermedad ocular llamada retinopatía diabética. El fichero contiene las siguientes variables, cuya descripción aparece en el apartado 10: duración de la diabetes (dur), porcentaje de hemoglobina glicosilada en la sangre (gly), índice de masa corporal (imc): peso/altura^2 y si el paciente desarrolló retinopatía (ret) 0: no 1: sí.

   - Realiza una descripción de la variable ret a partir de su distribución de frecuencias y de su diagrama de barras.

     Nota: Esta actividad no podrás hacerla tal cual, puesto que R-commander sólo permite los análisis indicados en el apartado 3 cuando dicha variable está especificada como cualitativa (factor) y, en nuestro caso, está codificada como una variable numérica. Para evitar este problema, debemos convertir la variable ret en una variable cualitativa o factor (que llamaremos, por ejemplo, ret2). Ello se realiza mediante la opción **Datos> Modificar variables del conjunto de datos activo> Convertir variable numérica en factor.**



     Notemos que, mediante la opción 'asignar nombres a los niveles', podemos indicar que el valor 0 indica no y el 1 sí, con lo cual las estadísticas resultarán más claras.

   - Dibuja el histograma de hemoglobina glicosilada para todos los pacientes. Describe lo que observas. Dibuja por separado los histogramas relativos a los pacientes que desarrollaron retinopatía y los que no lo hicieron. Compara los resultados.