Serie Monografías EASP Nº 54

Descubriendo *R-Commander*

Ricardo Ocaña Riola

JUNTA DE ANDALUCIA
Escuela Andaluza de Salud Pública
CONSEJERÍA DE SALUD Y FAMILIAS

Descubriendo *R-Commander*

Edición 2019

Ricardo Ocaña Riola

Doctor en Ciencias Matemáticas (Estadística e Investigación Operativa)
Profesor de Estadística
Escuela Andaluza de Salud Pública

Tercera Edición, 2019

Depósito Legal: GR 1591-2012
ISBN: 978-84-617-3296-8
Edita: Escuela Andaluza de Salud Pública (www.easp.es)

Índice

PRÓLOGO

Durante los últimos años ha habido un interés creciente entre los profesionales de Ciencias de la Salud por el uso del lenguaje de programación *R* y de la interfaz *R-Commander* en sus investigaciones, más debido al carácter gratuito de los mismos que a la necesidad de programar complejos algoritmos para el análisis estadístico de la información.

En la actualidad existe una amplia bibliografía sobre el lenguaje de programación *R* y sus procedimientos para el análisis de datos. Sin embargo, la documentación sobre *R-Commander* es escasa, especialmente en lengua castellana. Por ello, el propósito de esta monografía es elaborar una guía de ayuda sencilla para el análisis estadístico de datos mediante la interfaz *R-Commander*, dirigida a profesionales no especializados en Estadística que utilizan esta aplicación durante el desarrollo de actividades formativas básicas. A no ser que sea estrictamente necesario no se tratarán, por tanto, cuestiones relacionadas con la programación en *R* o el uso de secuencias de comandos, cuyo abordaje requiere conocimientos informáticos más avanzados y está orientado a especialistas que utilizan métodos estadísticos de forma intensiva en su labor profesional diaria.

Los procedimientos descritos a continuación están basados en *R-Commander* 2.3-2, incluido en la versión 3.3.2 de *R* que fue publicada el 31/10/2016. Anualmente, el número de nuevas versiones y actualizaciones suele ser superior a cuatro, por lo que es posible que algunos procedimientos no estén disponibles o hayan sido modificados en otras versiones. Dada la rapidez con la que se producen las revisiones de este software, es conveniente visitar con frecuencia la web de la *R Foundation for Statistical Computing* (*www.r-project.org*) e instalar la versión más actualizada.

Salvo excepciones, los capítulos siguientes no exponen el fundamento estadístico en el que se basa el procedimiento de análisis de *R-Commander*. Por ello, es aconsejable que el usuario de esta monografía tenga nociones básicas de estadística o bien utilice su contenido como complemento a los conocimientos adquiridos en actividades formativas de Estadística.

Las bases de datos y los casos prácticos utilizados están descritos en el último capítulo, permitiendo así la reproducción de los análisis en cualquier ordenador personal.

1

INTRODUCCIÓN

En el lenguaje común, el azar es sinónimo de casualidad. Sucesos impredecibles que no se pueden anticipar ni evitar. Durante siglos, la ciencia clásica ha negado la presencia de este tipo de sucesos en la Naturaleza. El principio de causalidad, en el que se basa el determinismo científico, afirma que cualquier fenómeno está provocado por una causa en la que el azar no tiene cabida. Esta relación entre la causa y el efecto siempre puede representarse a través de ecuaciones matemáticas capaces de predecir el comportamiento cualquier fenómeno natural una vez cuantificadas las causas que lo provocan. Para el determinismo científico, decir que un suceso ha ocurrido por azar es equivalente a decir que desconocemos las causas que lo provocan.

Si bien es cierto que muchos fenómenos de la Naturaleza pueden predecirse con exactitud, la mayoría de las teorías científicas actuales aceptan la existencia de otros fenómenos que no pueden explicarse mediante modelos puramente deterministas. Fenómenos que, de manera intrínseca, llevan asociados un componente aleatorio en su desarrollo. La propagación de una epidemia, las fluctuaciones bursátiles o el desplazamiento de un ciclón son algunos de los sucesos analizados en diferentes campos científicos en los que el azar juega un papel importante. Aunque pertenezcan a ámbitos distintos, todos estos sucesos tienen en común la imposibilidad de determinar con certeza cuál será su resultado final de entre todos los posibles. Su estudio requiere el uso de un tipo especial de modelos matemáticos denominados aleatorios, cuyo desarrollo tiene en cuenta el efecto del azar.

La Estadística es la ciencia que estudia este tipo de fenómenos. Desde su origen, a mediados del siglo XVII, los métodos estadísticos han permanecido en continuo desarrollo, contribuyendo a la toma de decisiones, al establecimiento de modelos causales y a la descripción de los fenómenos naturales más complejos. No en vano, la Estadística es actualmente un elemento clave en el proceso de investigación de cualquier disciplina fáctica, llegando a consolidarse el lenguaje universal de la ciencia del siglo XXI.

La Estadística no estudia a cada sujeto particular, sino a la población o grupo al que pertenecen esos sujetos. Por ello, es necesario disponer de técnicas que, partiendo de la información individual, sean capaces de extraer conclusiones generales del conjunto. Para este propósito, el desarrollo de múltiples programas informáticos, tanto libres como propietarios, ha permitido durante las últimas décadas la aplicación de complejos modelos estadísticos en

diferentes ámbitos, siendo herramientas fundamentales para el avance del conocimiento científico.

EL CONCEPTO DE SOFTWARE LIBRE

El término software libre se refiere a la libertad de los usuarios para copiar, distribuir, ejecutar y modificar un programa informático accediendo al código fuente del mismo. Sin embargo, el concepto *libre* no es sinónimo de *gratuito* o *no comercial*. Cualquier software libre puede tener un uso y distribución comercial, de manera que, a veces, el mismo software libre puede conseguirse de forma gratuita o pagando un precio determinado. De hecho, el uso comercial de software libre es cada vez más frecuente, siendo lícito la venta de copias o el desarrollo de software comercial a partir de él.[1] De la misma forma, un software gratuito no tiene que ser necesariamente libre, ya que el autor o autores pueden distribuir el producto sin permitir que los usuarios accedan al código fuente para modificarlo o generar nuevas versiones.

A modo de ejemplo, *Openbravo*, un programa informático para la planificación de recursos empresariales, ha sido diseñado como software libre. Su código es abierto, accesible a todo el mundo y el cliente puede modificarlo según sus necesidades. Sin embargo, no es un software gratuito. Existe un modelo de suscripción de pago según uso. Por otro lado, *QuickTime*, sistema de reproducción multimedia desarrollado por Apple, es gratuito, pero no es software libre puesto que el usuario no tiene acceso a su código fuente ni puede modificarlo.

La mayoría de usuarios finales, sin experiencia en programación, no decide utilizar software libre por contribuir a su desarrollo, corregir errores o ampliar sus prestaciones mediante la programación del código abierto. El motivo suele estar más relacionado con la reducción de costes que supone prescindir del pago de licencias, de manera que el software libre será atractivo para este tipo de usuarios siempre que sea gratuito. Aunque esta es una ventaja importante, el software libre también presenta desventajas relacionadas generalmente con la complejidad de instalación, la dificultad de aprendizaje, la ausencia de soporte técnico oficial y la escasez de manuales que permitan una formación sólida estructurada. Por ello, la implicación de la comunidad que utiliza el software es fundamental para su mejora, ya que suelen ser los propios usuarios los que elaboran documentación, informan de errores y comparten conocimiento a través de foros, listas de distribución o redes sociales.

El software libre gratuito puede ser una alternativa al software propietario de pago, pero el coste económico no debe ser el principal factor que decida su utilización. En cada caso particular será necesario realizar una valoración del perfil profesional del usuario final, los

[1] GNU Operating System. *¿Qué es el software libre?* Disponible en: www.gnu.org/philosophy/free-sw.es.html

objetivos a alcanzar, las necesidades que se han de cubrir y las prestaciones que ofrecen las diferentes alternativas de software, ya sea libre, propietario, gratuito o de pago.

EL LENGUAJE DE PROGRAMACIÓN *R*

R es un lenguaje de programación muy flexible orientado a la estadística computacional, el análisis de datos y el desarrollo de gráficos. Es un software libre y gratuito desarrollado bajo las condiciones *GNU General Public License* (www.gnu.org) por el equipo central de la *R Foundation for Statistical Computing* (*www.r-project.org*).

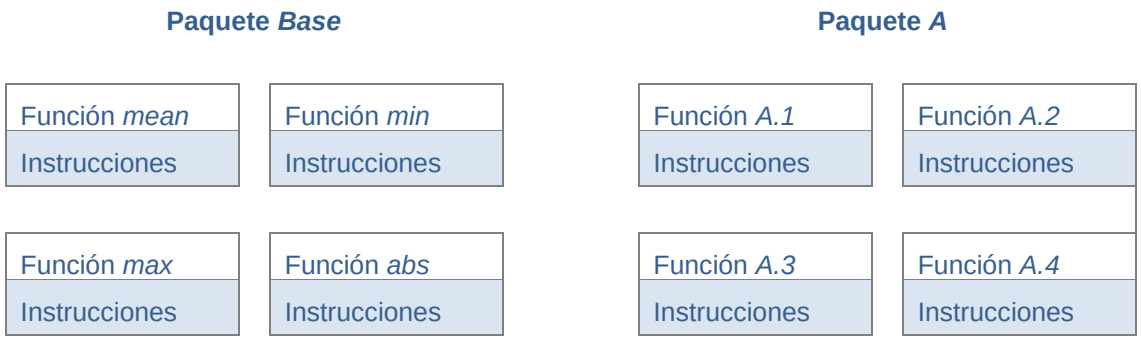
Como en cualquier lenguaje de programación, el usuario debe conocer bien el entorno de trabajo y las funciones básicas implementadas en *R* para, a partir de ellas, realizar el análisis estadístico deseado o desarrollar nuevas funciones. En el ámbito informático, una función es un grupo de instrucciones que procesa los datos introducidos y devuelve un valor final. Así, para calcular la media de una serie de valores en *R* será necesario programar una función que contenga las siguientes instrucciones:

```
function (x, trim = 0, na.rm = FALSE, ...)
{
    if (!is.numeric(x) && !is.complex(x) && !is.logical(x))
    {
        warning("argument is not numeric or logical: returning NA")
        return(as.numeric(NA))
    }
    if (na.rm)
        x <- x[!is.na(x)]
    trim <- trim[1]
    n <- length(x)
    if (trim > 0 && n > 0)
    {
        if (is.complex(x))
            stop("trimmed means are not defined for complex data")
        if (trim >= 0.5)
            return(median(x, na.rm = FALSE))
        lo <- floor(n * trim) + 1
        hi <- n + 1 - lo
        x <- sort(x, partial = unique(c(lo, hi)))[lo:hi]
        n <- hi - lo + 1
    }
    if (is.integer(x))
        sum(as.numeric(x))/n
    else sum(x)/n
}
```

Todas estas instrucciones forman una función denominada *mean*, que por defecto ya viene implementada en *R*. Cuando el usuario hace uso de ella, la función solicita unos datos y devuelve el valor medio de los mismos. Así, al escribir *mean(c(*0,2,4*))* se procesarán automáticamente las instrucciones anteriores y se obtendrá como resultado 2, media aritmética de los valores 0, 2 y 4.

Al igual que *mean*, existen otras funciones que ya han sido programadas por el equipo de desarrollo de *R* y están disponibles para su uso inmediato, como las funciones *min* y *max* que devuelven respectivamente el valor mínimo y máximo de los datos introducidos. Así, *min(c(*0,2,4*))* dará como resultado 0 y *max(c(*0,2,4*))* devolverá el valor 4. Estas funciones forman parte de procedimientos estadísticos básicos, por lo que están agrupadas en un paquete de funciones denominado *Base*. Este paquete, junto a otros que contienen funciones más avanzadas, ha sido desarrollado por el equipo central de *R* y viene incorporado en su instalación.

En general, las instrucciones que permiten realizar un cálculo determinado se programan en una función y a su vez las funciones se agrupan en paquetes temáticos para facilitar su localización. La estructura es similar a la descrita en el siguiente gráfico:



Actualmente, quizá *R* sea uno de los lenguajes de programación con más funciones implementadas para el análisis de datos. Además, su flexibilidad permite programar e incorporar nuevos modelos que han sido desarrollados en el campo de la teoría matemática, cualidad que lo ha convertido en un software muy popular entre estadísticos y matemáticos especializados en estadística computacional.

A pesar de sus cualidades técnicas, el uso de *R* puede resultar complejo para personas que no están familiarizadas con los lenguajes de programación. La necesidad de escribir instrucciones y comandos para realizar análisis estadísticos simples hace que *R* no sea el

software elegido por profesionales no especializados en estadística para llevar a cabo proyectos de investigación aplicada.

LA INTERFAZ GRÁFICA R-COMMANDER

En general, el número de personas que usa un software libre no está determinado exclusivamente por cuánto puede hacer el software. La facilidad de uso ha de ser también una de sus características principales, ya que la mayoría de usuarios rehusarán utilizarlo si el software libre no permite realizar de forma sencilla todos los trabajos que necesitan llevar a cabo.[2]

Por este motivo, John Fox, profesor de Sociología de la Universidad McMaster (Canadá), desarrolló en 2005 el paquete *Rcmdr*, una Interfaz Gráfica de Usuario denominada *R-Commander* que permitía a sus alumnos trabajar en un entorno de ventanas similar al de otros programas estadísticos como SPSS.[3] De esta forma se ha facilitado el manejo de *R* en cursos de estadística básica, de manera que actualmente el usuario puede elegir el entorno en el que desea trabajar, ya sea mediante la interfaz *R-Commander* o a través de la consola de instrucciones y comandos de programación *R*.

FORTALEZAS Y DEBILIDADES DE *R-COMMANDER*

La implementación de esta interfaz al software *R* ha mejorado mucho su apariencia, permitiendo que nuevos usuarios no especializados en lenguajes de programación lo utilicen para el análisis estadístico de datos. Desde su introducción en 2005, *R-Commander* ha tenido varias actualizaciones que han aumentado sus posibilidades de forma progresiva. Esta característica, junto a su distribución gratuita, constituyen sus principales fortalezas. Sin embargo, aún presenta algunas debilidades que conviene conocer.

Salida de resultados

Uno de los principales inconvenientes de *R-Commander* es que la ventana de resultados no estructura las salidas en tablas que se puedan copiar, pegar o exportar a documentos de paquetes ofimáticos estándar, como Microsoft Office, Open Office o iWork, por citar algunos. Por ello, es aconsejable que el usuario vaya configurando sus propias tablas de resultados en

[2] GNU Operating System. ¿Qué es el software libre? Disponible en: www.gnu.org/philosophy/free-sw.es.html

3 Fox J. The R Commander: A Basic-Statistics Graphical User Interface to R. Journal of Statistical Software 2005; 11(9): 1-42.

un procesador de textos, transcribiendo la información que *R-Commander* ofrece en la ventana. Aunque el proceso puede resultar tedioso, es la única forma de organizar la información y conseguir un documento final comprensible para personas que no han participado en el análisis de datos o no están familiarizadas con las salidas de resultados de esta interfaz.

Gráficos

Los gráficos realizados con *R-Commander* se visualizan en una ventana independiente denominada ventana gráfica. Cada uno de ellos podrá guardarse en diferentes formatos desde el menú principal de esta ventana, pulsando la opción *Archivo – Guardar como*. También será posible copiarlo pulsando la secuencia *Archivo – Copiar para el área de transferencia* y a continuación pegarlo directamente en un documento de trabajo. Sin embargo, antes de guardar un gráfico o incorporarlo a un documento, es aconsejable mejorar su apariencia y modificar algunos aspectos para que sea autoexplicativo.

Si bien las capacidades gráficas de *R* son enormes, las opciones de *R-Commander* son muy limitadas. Esta interfaz no dispone de un editor de gráficos que permita, entre otros, modificar la leyenda o el color del gráfico antes de exportarlo a un documento, por lo que habitualmente su aspecto original no es el idóneo para informes o presentaciones profesionales. Para modificar su apariencia será necesario recurrir a la sintaxis de programación en *R*, modificando manualmente el código de la gráfica básica, añadiendo nuevos parámetros o ejecutando determinados comandos. Este procedimiento requiere conocer algo más sobre el funcionamiento de este lenguaje, lo que supondrá un esfuerzo adicional para algunos usuarios. En los sucesivos capítulos se explicará cómo realizar esta tarea para cada gráfico particular, aunque existen ciertos parámetros comunes que se suelen incorporar en la mayoría de las líneas de comandos. La siguiente tabla describe los más usuales:

Parámetro	Descripción	Ejemplo
main="Título"	Inserta en la cabecera del gráfico el título entrecomillado.	main="Distribución de la variable sexo"
xlab="Etiqueta"	Inserta en el eje horizontal de un gráfico la etiqueta entrecomillada.	xlab="Sexo"
ylab="Etiqueta"	Inserta en el eje vertical de un gráfico la etiqueta entrecomillada.	ylab="Número de sujetos"
col="color"	Pinta el gráfico del color especificado entre comillas. Se puede elegir entre 657 nombres de colores, todos en inglés.	col="darkblue" (azul oscuro) col="blue" (azul) col="lightblue" (azul claro)
col=número	Pinta el gráfico del color especificado en el número. Este número puede estar entre 0 y 8, repitiéndose cíclicamente los mismos colores a partir del valor 9.	col=0 (transparente) col=1 (negro) col=5 (turquesa) col=2 (rojo) col=6 (violeta) col=3 (verde) col=7 (amarillo) col=4 (azul) col=8 (gris)

Métodos estadísticos

Aunque la mayor parte de las técnicas estadísticas básicas se encuentran incorporadas en la interfaz de ventanas *R-Commander*, algunos métodos más avanzados o determinadas medidas de asociación, como el riesgo relativo o la razón de oportunidades (*odds ratio*), no están implementadas.[4] En estos casos será necesario recurrir a la sintaxis y comandos de *R* o a algún software externo que permita realizar el análisis.

Como complemento puede ser útil disponer de la calculadora estadística *OpenEpi* (www.openepi.com), un software gratuito que permite realizar cálculos estadísticos sencillos. Puede utilizarse desde un servidor web o bien descargarse y ejecutarse posteriormente sin conexión a Internet. Entre sus opciones se encuentran análisis específicos para estudios descriptivos y analíticos, tablas de contingencia, análisis estratificado, análisis de datos apareados, cálculo del tamaño de muestra, números aleatorios, medidas de sensibilidad y especificidad, test de hipótesis e intervalos de confianza, entre otras.

[4] El test de Fisher implementado en *R-Commander* ofrece una estimación de la *odds ratio* para tablas 2 x 2. Sin embargo, ésta no es la *odds ratio* habitual del producto cruzado, sino una estimación obtenida por máxima verosimilitud condicional que difiere de la *odds ratio* convencional.

Gestión de bases de datos

R-Commander no ha sido diseñado para gestionar de forma fácil y eficaz grandes bases de datos, motivo por el que no es aconsejable su uso para registrar y almacenar la información. En su lugar, los desarrolladores de esta interfaz recomiendan utilizar un sistema gestor de bases de datos externo, similar a *Microsoft Access* o *dBase*, y capturar posteriormente la información con *R-Commander* para llevar a cabo el análisis estadístico.

Errores

Algunos fallos de programación hacen que la interfaz no funcione correctamente durante el proceso de edición de datos, produzca resultados erróneos o se bloquee en determinados procedimientos. Quizá estas incidencias queden solucionadas en versiones posteriores de *R-Commander*, al igual que ocurrió en el pasado con otras.

Como advierte la ventana de inicio del programa, *R* es un software libre y viene sin garantía alguna. Por extensión, esta característica también afecta a *R-Commander*, de manera que en ocasiones habrá que recurrir a las listas de distribución o los foros de usuarios para solicitar asistencia sobre incidencias que no se hayan podido resolver. Puesto que estas listas de ayuda son voluntarias no se podrá exigir soluciones a los usuarios ni una respuesta inmediata al problema planteado, lo que en ocasiones puede retrasar el trabajo que se está llevando a cabo.

¿SON FIABLES *R* Y *R-COMMANDER*?

La concepción de *R* como software libre ha permitido que muchas personas programen nuevas funciones y paquetes que se añaden periódicamente a los implementados originalmente por el equipo central de desarrollo de *R*. Estos paquetes se van incorporando a la red CRAN (*Comprehensive R Archive Network*), un repositorio web utilizado por los usuarios para distribuir sus trabajos de forma gratuita.[5] Durante el periodo de redacción de esta monografía había disponibles más de 13.000 paquetes en este sitio web. Algunos de ellos han sido desarrollados por especialistas con experiencia en estadística computacional. Otros, por personas ajenas a esta área de conocimiento y alumnos universitarios no especializados en estadística. Por ello, no todos los paquetes tienen la misma fiabilidad. Ninguno de ellos está exento de posibles errores, ni existe garantía alguna sobre su eficiencia. Es el propio usuario

[5] Hornik K. *R FAQ: Frequently Asked Questions on R*. Viena: The R Foundation, 2018. Disponible en: http://cran.r-project.org/doc/FAQ/R-FAQ.pdf

quien debe decidir si utiliza o no un determinado paquete, establecer los procedimientos de control de calidad apropiados y valorar la forma de utilizarlo durante el proceso de investigación.

La normativa que regula los ensayos clínicos es muy estricta en los aspectos éticos y metodológicos de la investigación. En 1998, el Comité Directivo de la Conferencia Internacional sobre Armonización (ICH) configuró la directriz E9 sobre métodos estadísticos en ensayos clínicos, que se añadió a otras directrices ICH desarrolladas con anterioridad.[6] Esta guía fue adoptada por la EMEA (European Medicines Agency) y la FDA (U.S. Food and Drug Administration) y actualmente es la base de la normativa europea vigente, estableciendo los estándares estadísticos para la investigación sobre nuevos medicamentos. Además de ser un documento clave en este campo, su aplicación se ha extendido a la investigación clínica en general, siendo de gran importancia para todos los profesionales que realicen análisis estadísticos de datos en investigación básica o aplicada.

La ICH-E9 recoge en el apartado "Integridad de los datos y validez del software" lo siguiente: "… El software utilizado para la gestión de datos y el análisis estadístico debe ser fiable y la documentación sobre los procedimientos empleados para chequear el software debe estar disponible". En respuesta a esta normativa, The *R* Foundation for Statistical Computing publicó en 2008 el documento "R: Cumplimiento normativo y cuestiones de validación. Un documento orientativo para el uso de *R* en entornos de ensayos clínicos regulados", cuyo contenido se actualiza periódicamente.[7] En él se explicita que no todos los paquetes de *R* están validados por los creadores de este software y, por tanto, no todos cumplen con la directriz ICH-E9. Así, de los más de 13.000 paquetes que actualmente están disponibles en el repositorio CRAN los creadores de *R* sólo garantizan la fiabilidad de 29, aquellos que han sido desarrollados por el equipo central de *R*. Para el resto, la Fundación *R* no ofrece ninguna garantía. Esta declaración afecta a *R-Commander*, ya que no es uno de los paquetes base que vienen instalados por defecto en *R* ni aparece entre los denominados "*Paquetes Recomendados*". El documento está disponible en la página principal de la web del proyecto (www.r-project.org) dentro del enlace *certification*, mencionando lo siguiente en su apartado 2:

[6] Lewis JA. Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline. *Statistics in Medicine* 1999; 18: 1903-1942.

[7] The R Foundation for Statistical Computing. *R: Regulatory Compliance and Validation Issues. A Guidance Document for the Use of R in Regulated Clinical Trial Environments*. Viena: The R Foundation, 2018. Disponible en: http://www.r-project.org/doc/R-FDA.pdf

> Es importante aclarar que este documento [R: cumplimiento normativo y cuestiones de validación] es ÚNICAMENTE aplicable a los paquetes de R que se suministran junto con R y que llevan el copyright de la Fundación R. Este software se conoce comúnmente como "R Base" más "Paquetes Recomendados" y se publican tanto en código fuente como en formato binario ejecutable bajo Licencia Pública GNU de la Fundación para el Software Libre.
>
> Al escribir estas líneas, "R Base" incluye los siguientes paquetes: base, compiler, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, stats4, tcltk, tools, utils, Por otra parte, "Paquetes Recomendados" incluye los siguientes paquetes: boot, class, cluster, codetools, foreign, KernSmooth, lattice, MASS, Matrix, mgcv, nlme, nnet, rpart, spatial, survival.
>
> Este documento no es de ninguna manera aplicable a otro software relacionado con R, ni a paquetes adicionales disponibles a través de terceros, como los usuarios o miembros del Equipo Central de Desarrollo de R, que pueden, de vez en cuando, hacer disponible su software a través de CRAN u otros repositorios de distribución de software.
>
> Este documento no pretende ser prescriptivo, no presta una opinión legal y no confiere o comunica ninguna obligación legal o de otra índole. Debe ser utilizado por el lector y su organización como un componente en el proceso de toma de decisiones informadas sobre la mejor manera para cumplir con la normativa y las obligaciones pertinentes dentro de su propio entorno de trabajo profesional.
>
> La Fundación R para la Estadística Computacional no ofrece ninguna garantía, expresa o implícita, en este documento.

Según esta información el 99% de los paquetes incorporados a la librería CRAN no se ajustan a las directrices ICH-E9 sobre principios estadísticos para ensayos clínicos. Como solución, el informe elaborado por el equipo central de *R* traslada al investigador y a su organización la obligación de definir los procesos de control de calidad adecuados para cumplir con el marco normativo vigente cuando utilice cualquier paquete elaborado por otros usuarios, incluido *R-Commander*, lo que supone implementar y hacer públicos los procedimientos operativos estándar de control que realizan los ingenieros de informática antes de lanzar cualquier

software al mercado. Evidentemente, este propósito está fuera del alcance de la mayoría de los profesionales no especializados en estadística computacional, por lo que tanto investigadores como instituciones han de tener en cuenta estas consideraciones, y no sólo el carácter gratuito de la aplicación, antes de tomar una decisión sobre el uso de *R* y *R-Commander*. Este y otros aspectos cobran especial relevancia para las organizaciones que necesitan adquirir un software estadístico para uso oficial, debiendo elegir aquel que mejor se adapte a sus capacidades, necesidades y actividades profesionales. Actualmente, la oferta de productos que coexisten en el mercado es muy amplia y su evaluación debe formar parte de un proceso de toma de decisiones asesorado por especialistas con experiencia en el uso de programas estadísticos. Tanto el software libre como el privativo presentan ventajas e inconvenientes que serán diferentes para cada actividad, profesión e institución,[8] por ello no es aconsejable ni apropiado recomendar el uso indiscriminado del lenguaje de programación *R* o la interfaz *R-Commander* a cualquier organización o profesional, especialmente si su actividad principal no es la estadística computacional.

Actualmente, el ensayo clínico es el único diseño de investigación que cuenta con una normativa sobre principios estadísticos y validación de software. Sin embargo, todo lo mencionado anteriormente se hace extensible de forma natural a cualquier tipo de estudio e investigación que requiera llevar a cabo un análisis estadístico de la información.

[8] Culebro M, Gómez WG, Torres S. Software libre vs software propietario: Ventajas y desventajas. México, 2006.

2

INSTALACIÓN DE *R-COMMANDER*

R-Commander es una interfaz gráfica que permite trabajar con *R* a través de un entorno de ventanas similar al de otros programas estadísticos. Para utilizarlo es necesario instalar previamente *R* y configurar algunas opciones que faciliten su manejo. Los siguientes apartados muestran el procedimiento para descargar *R* desde Internet, instalar tanto el software como la interfaz y comenzar una sesión de trabajo con el entorno de ventanas.

El programa dispone de varias versiones que permiten trabajar en sistemas operativos Windows, MacOS X y Linux. Este apartado describe cómo realizar la descarga e instalación bajo Windows y MacOS X por ser los más habituales.

SISTEMA OPERATIVO WINDOWS

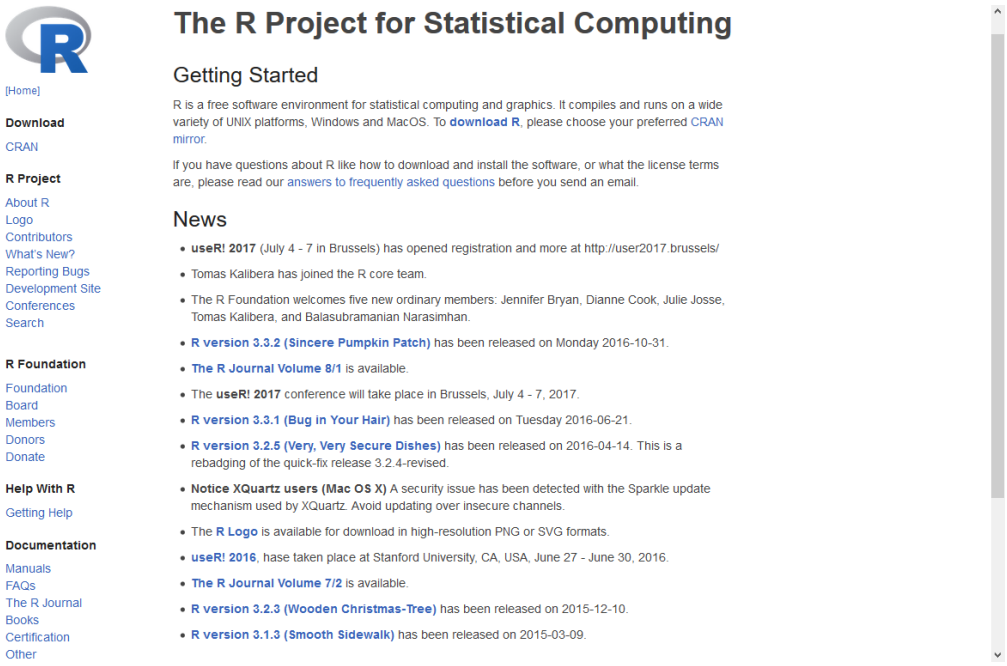
La forma más sencilla de instalar *R-Commander* es a través del *Paquete R-UCA*, un archivo ejecutable desarrollado por el proyecto R-UCA de la Universidad de Cádiz. El acceso se realiza a través de la web *knuth.uca.es/R*. Una vez dentro, hay que localizar en el menú de navegación el enlace *Paquete R-UCA*. Pulsando sobre este enlace aparecerá la página de instalación, donde el apartado *Versiones* contiene el último archivo ejecutable disponible para instalar automática y simultáneamente *R* y *R-Commander*.

La versión de *R-Commander* instalada a través de R-UCA servirá para cubrir los objetivos de la mayoría de usuarios. Sin embargo, no es la última versión disponible. En caso de querer trabajar con la versión más reciente de *R* y *R-Commander* será necesario descargar e instalar ambos programas desde la página original de sus desarrolladores, como muestran las siguientes secciones.

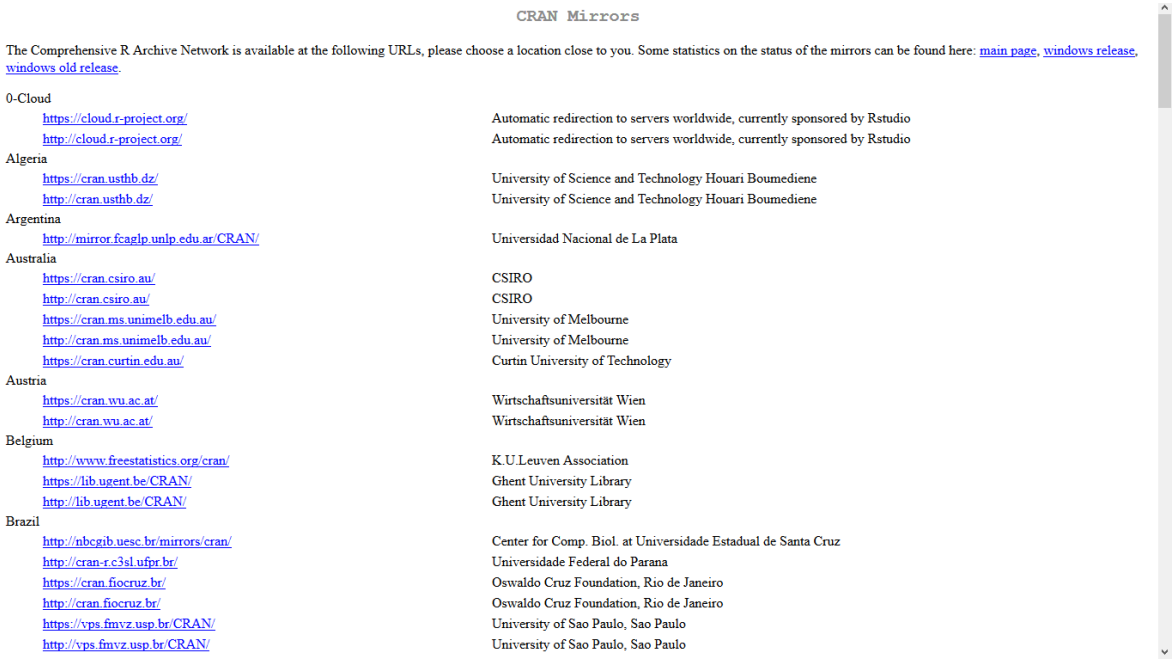
Descarga de *R*

Algunas de las siguientes capturas de pantalla podrían variar dependiendo de la versión de Windows. Sin embargo, el procedimiento de descarga siempre se realiza mediante los hiperenlaces descritos a continuación:

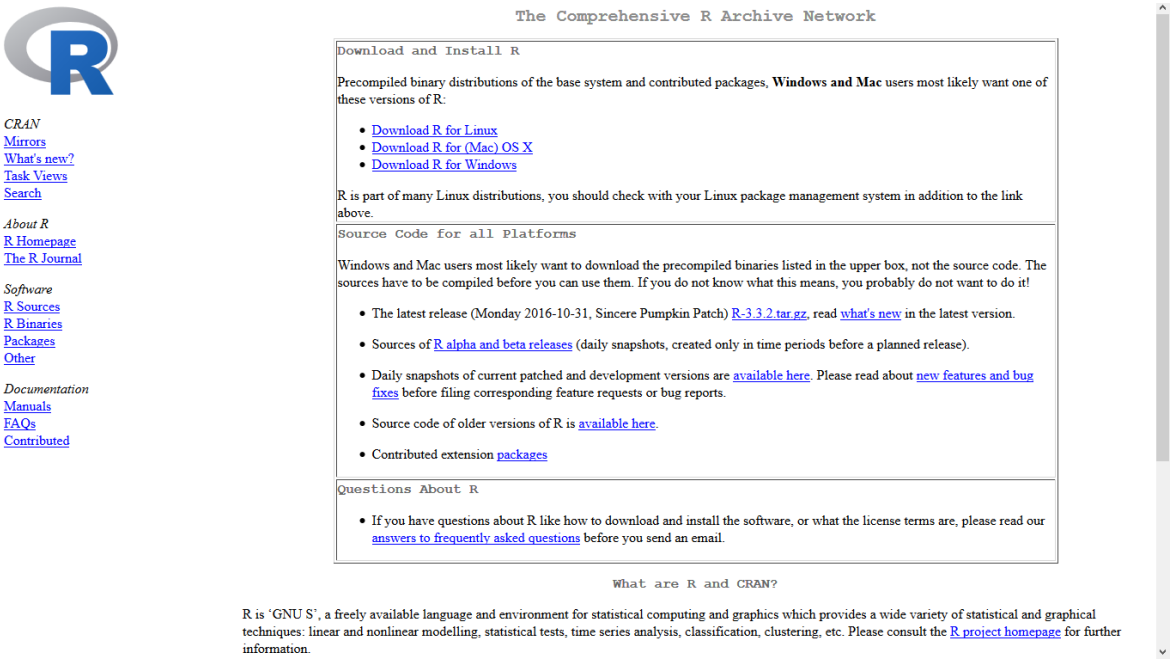
1. Desde el explorador de Internet, entrar en *www.r-project.org*. A continuación, hacer clic con el botón izquierdo del ratón en el enlace *download R*, situado dentro del recuadro *Getting Started*.



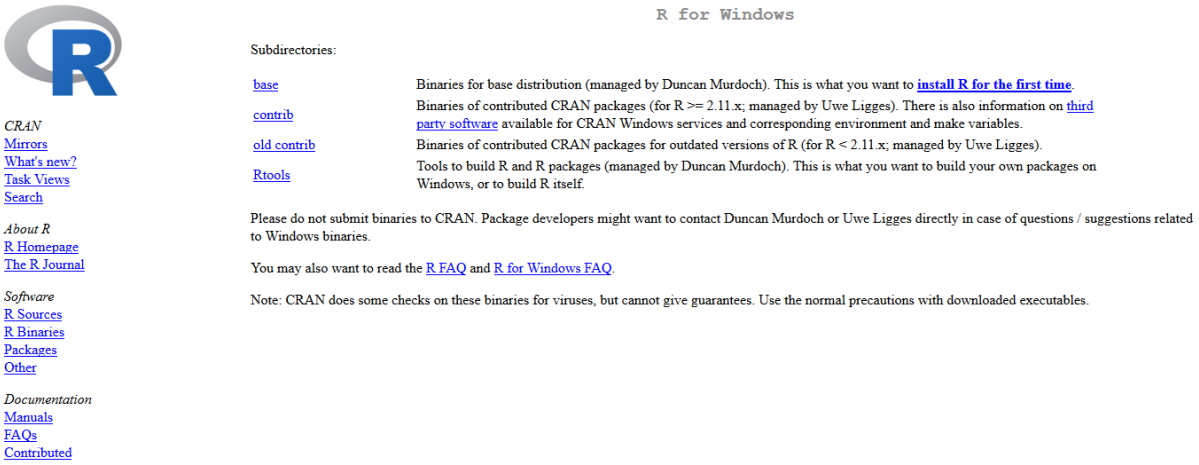
2. Localizar España (*Spain*) en el listado de Servidores y pinchar sobre el enlace correspondiente a Madrid (*Spanish National Research Network, Madrid*), habitualmente dado por http://cran.rediris.es



3. En el cuadro *Download and Install R*, seleccionar la opción *Download R for Windows*.

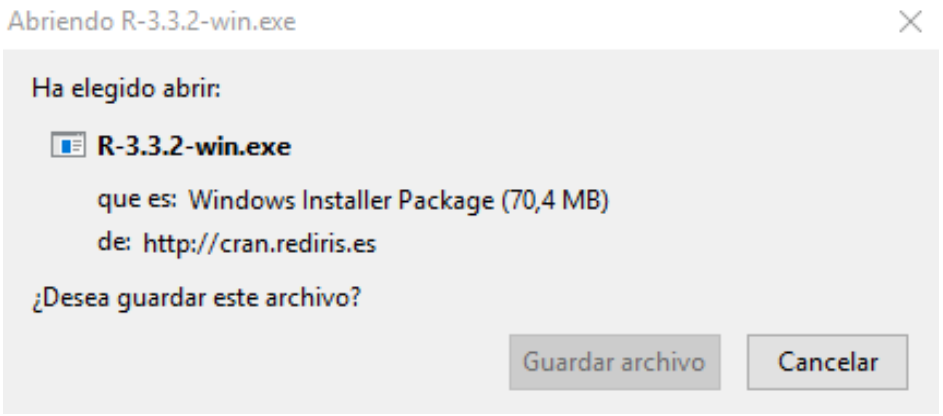


4. Hacer clic sobre el enlace *base*.



5. En el cuadro superior de la pantalla aparecerá un enlace que lleva a la última versión de R. Hacer clic sobre este enlace, titulado *Download R (...) for Windows*. En lugar de los puntos suspensivos aparecerá la numeración de la última versión de *R* disponible.

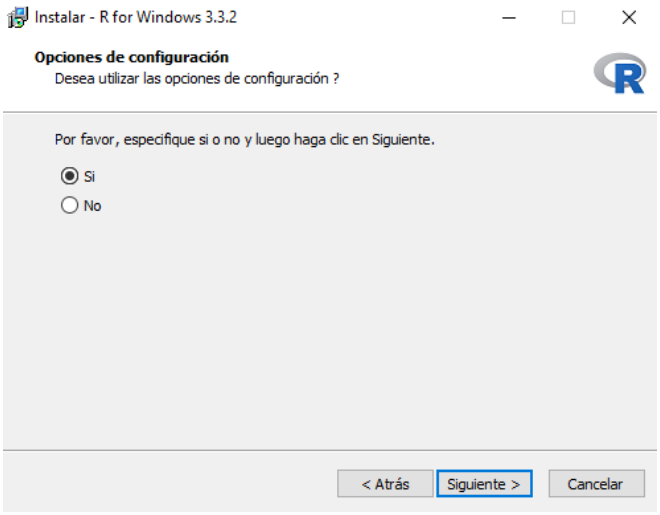
6. Pulsar el botón *Guardar archivo* para almacenar el archivo *R-(...)-win.exe* en la carpeta *Mis documentos* o en cualquier otra que el usuario decida. Este archivo ejecutable se utilizará posteriormente para instalar *R* en el ordenador.



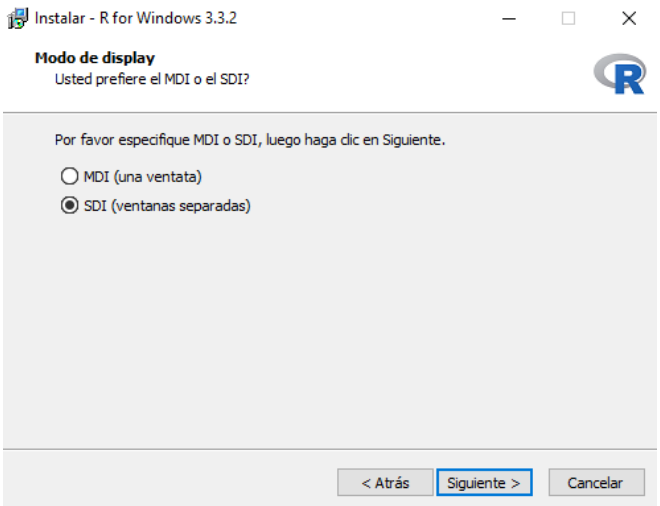
Instalación de *R*

Una vez descargado el archivo ejecutable *R-(...)-win*, la instalación de *R* es muy sencilla. Bastará con ir a la carpeta donde se almacenó el archivo, hacer doble clic sobre él y seguir las instrucciones que aparecerán en pantalla.

La configuración que aparece por defecto es la más frecuente, aunque es aconsejable asegurar la instalación de un parámetro en particular. Para ello, cuando el cuadro de diálogo de la instalación pregunte si se desea utilizar las opciones de configuración se responderá *Sí*.



Tras pulsar el botón *Siguiente* aparecerá la ventana *Modo de display*. En ella es conveniente marcar siempre la opción *SDI* (ventanas separadas), ya que la selección *MDI* alternativa suele tener algunos problemas de compatibilidad con *R-Commander*. El resto de opciones puede quedar como aparece por defecto.

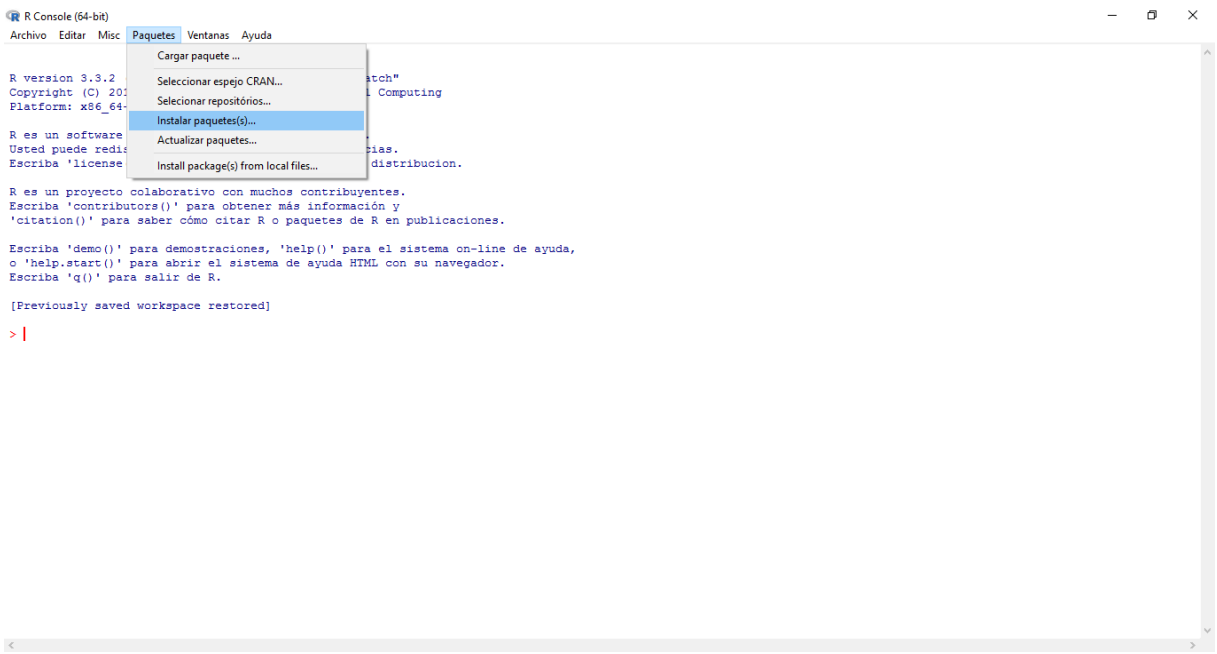


Instalación de *R-Commander*

Tras la instalación del software *R* aparecerán en el escritorio dos iconos con la forma . Uno de ellos llevará a pie de imagen el nombre *Ri386* y el otro *Rx64*, haciendo referencia respectivamente a la versión 32-bit o 64-bit de *R*. En general, es recomendable usar la primera versión (*Ri386*) si se trabaja bajo Windows 32-bit y la segunda (*Rx64*) si el sistema operativo es Windows 64-bit.[9] El tipo de sistema operativo instalado en cada ordenador puede consultarse pulsando el botón *Inicio* de Windows y a continuación *Configuración–Acerca de-Tipo de sistema*.

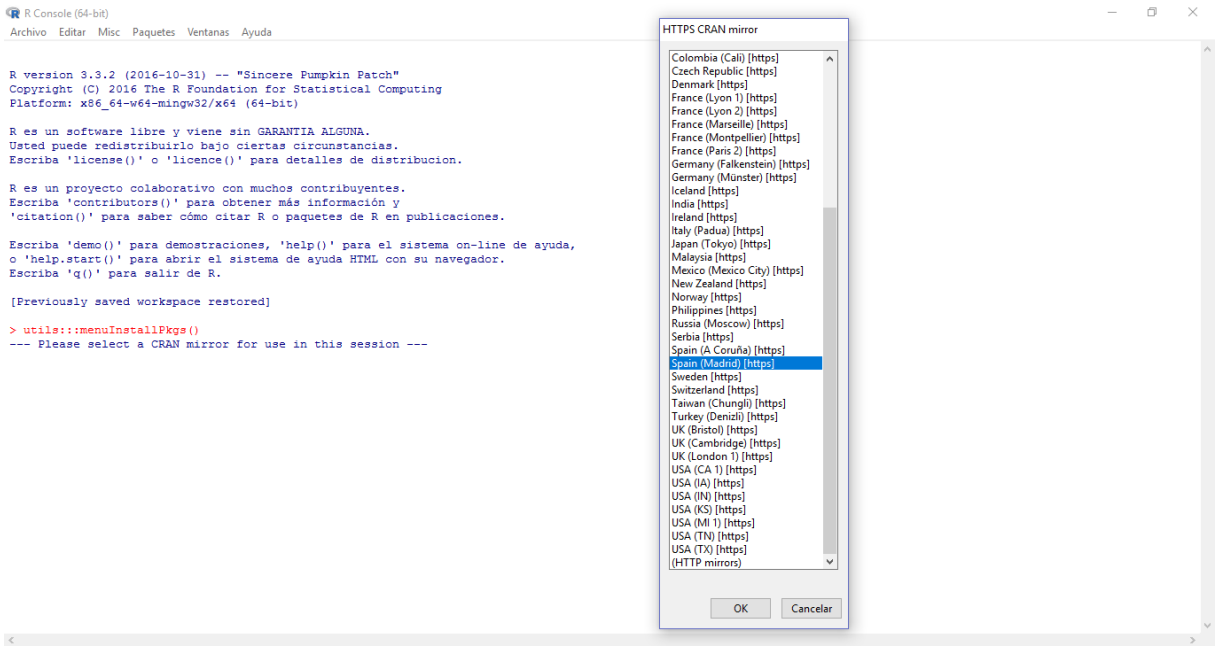
Fase 1: Procedimiento de instalación obligatorio

Para instalar *R-Commander* se seleccionará la pestaña *Paquetes* del menú principal y posteriormente la opción *Instalar paquetes(s)*.

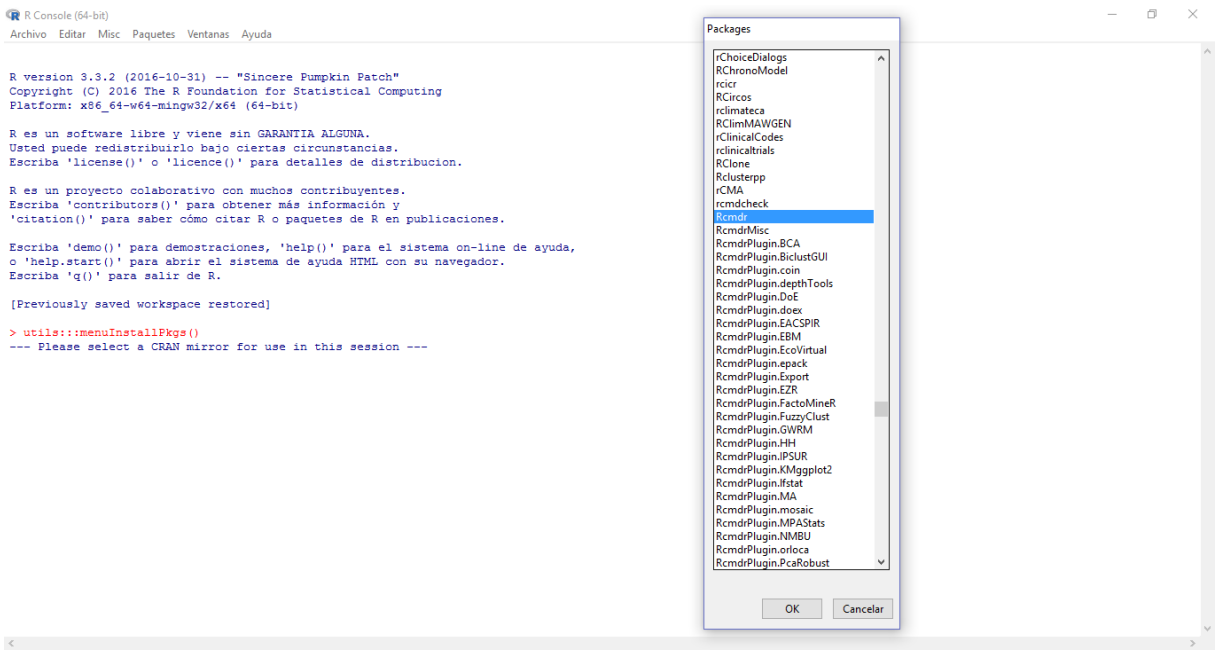


Se abrirá una nueva ventana que contiene los servidores desde los que se pueden descargar nuevos paquetes. En ella habrá que hacer clic sobre la opción *Spain (Madrid)*, o bien *Spain (A Coruña)*, y pulsar el botón *OK*, como muestra la siguiente imagen:

[9] Ripley BD, Murdoch. *R for Windows FAQ*. Disponible en http://cran.rediris.es/bin/windows/base/rw-FAQ.html.



A continuación, aparecerá en orden alfabético un listado con los paquetes disponibles. En esta ventana deberá seleccionarse con el ratón el paquete *Rcmdr*.



Por último, pulsando sobre el botón *OK* de la ventana de paquetes comenzará la instalación de *R-Commander*. Es posible que la ventana de instalación pregunte si se desea usar una librería personal (*Would you like to use a personal library instead?*) o si se quiere crear (*Would you like to create a personal library?*). En ambos casos se pulsará *Sí*.

Cuando termine la instalación del paquete *R-Commander*, será necesario comprobar si se ha realizado correctamente. Para ello, junto al símbolo > de la consola de *R*, se deberá escribir la instrucción *library(Rcmdr)* y pulsar la tecla *Intro* del teclado. Si la instalación fue correcta,

se abrirá la ventana de *R-Commander*. En caso contrario, se deberán repetir desde el principio los pasos descritos en esta Fase 1 de instalación.

Fase 2: Procedimiento de instalación adicional optativo

Una vez efectuada la instalación descrita en la Fase 1, es posible programar algunas instrucciones para que *R-Commander* se abra automáticamente al iniciar una sesión de trabajo e incorpore algunas utilidades para el análisis de datos. Aunque no es obligatorio, realizar este paso adicional es aconsejable para agilizar la entrada a *R-Commander*, especialmente si se va a utilizar con frecuencia.

Para llevar a cabo este procedimiento se ejecutará en primer lugar el Bloc de Notas, disponible en la carpeta de accesorios de Windows, o cualquier otro editor de texto. Para tener permisos de edición, es necesario abrir el editor con la opción "Ejecutar como administrador". Utilizando uno de estos editores se abrirá el archivo *Rprofile.site*, situado en la carpeta C:\Archivos de programa\R\R-(…)\etc (o bien C:\Program Files\R\R-(…)\etc), y a continuación se escribirán las siguientes instrucciones al final del contenido de este archivo:

```
local({
old <- getOption("defaultPackages")
options(defaultPackages = c(old, "Rcmdr"))
})
```

Es muy importante escribir el texto tal como aparece, respetando las letras mayúsculas y minúsculas, sin olvidar ningún paréntesis, corchete o entrecomillado.

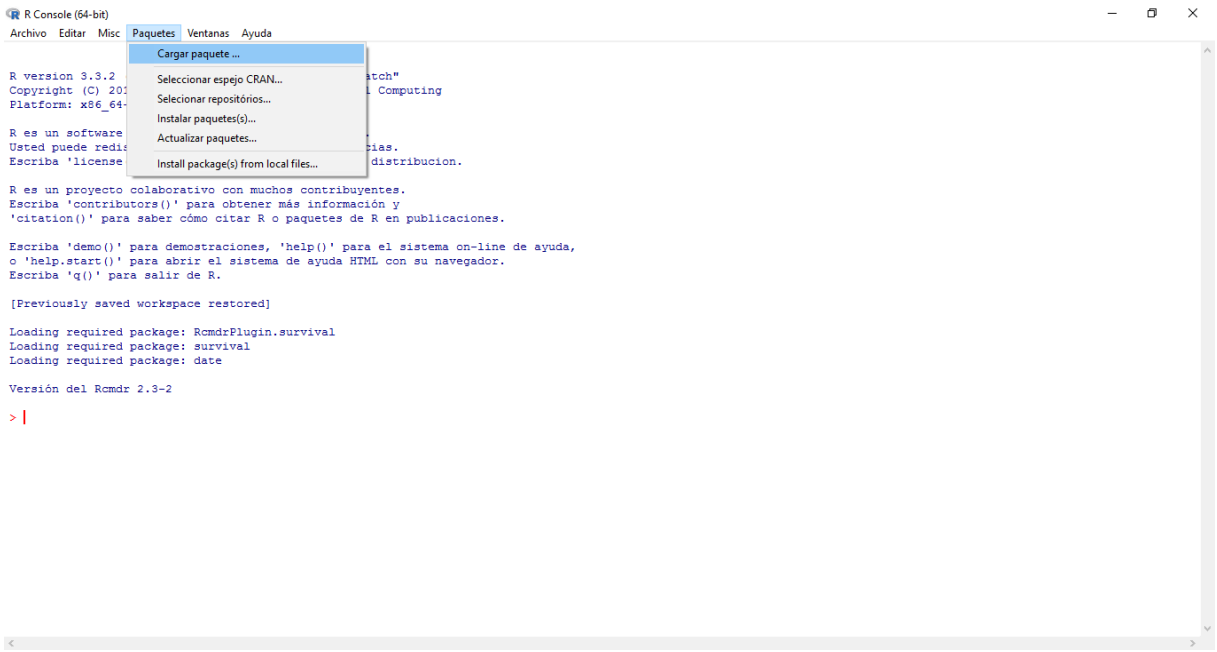
Por último se guardará el archivo *Rprofile.site* dentro de la misma carpeta, reemplazando al archivo original, teniendo en cuenta que su extensión ha de ser *.site* y no *.txt* o *.dat* como puede aparecer por defecto en algunos editores de texto.

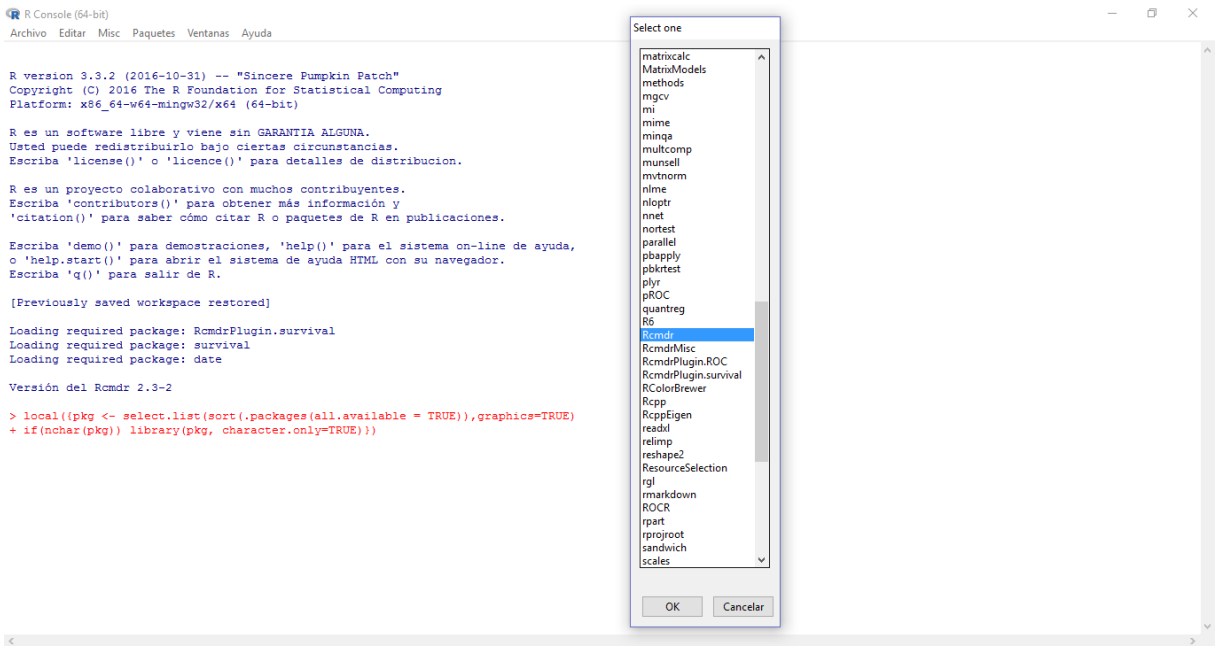
Comenzar una sesión de trabajo con *R-Commander*

Si en la instalación de *R-Commander* se realizó el procedimiento adicional optativo descrito en la Fase 2, bastará con hacer doble clic en el icono del escritorio para comenzar una sesión de trabajo con *R-Commander*.

Si durante el procedimiento de instalación de *R-Commander* sólo se realizó la Fase 1, entonces tras pulsar el icono se abrirá la consola de *R* pero no la de *R-Commander*. En este

caso, para activar *R-Commander* habrá que escribir *library(Rcmdr)* en la consola de *R* y pulsar la tecla *Intro* del teclado. Otra opción es pulsar la secuencia de pestañas *Paquetes* - *Cargar paquete* del menú principal. Se abrirá entonces una ventana con los paquetes disponibles, sobre la que habrá que seleccionar *Rcmdr* y pulsar el botón *OK*:



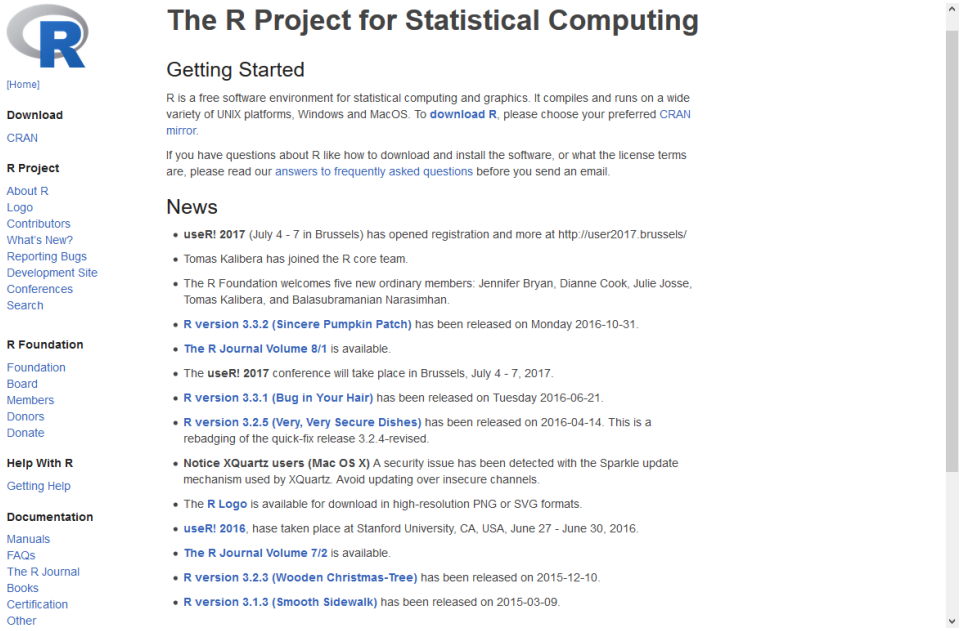


Cualquiera de los dos procedimientos tendrá que repetirse en cada sesión de trabajo, siempre que se desee trabajar con *R-Commander*. La nueva ventana abierta será el entorno de trabajo *R-Commander*, que podrá maximizarse para tener un campo visual más amplio.

SISTEMA OPERATIVO MAC OS X

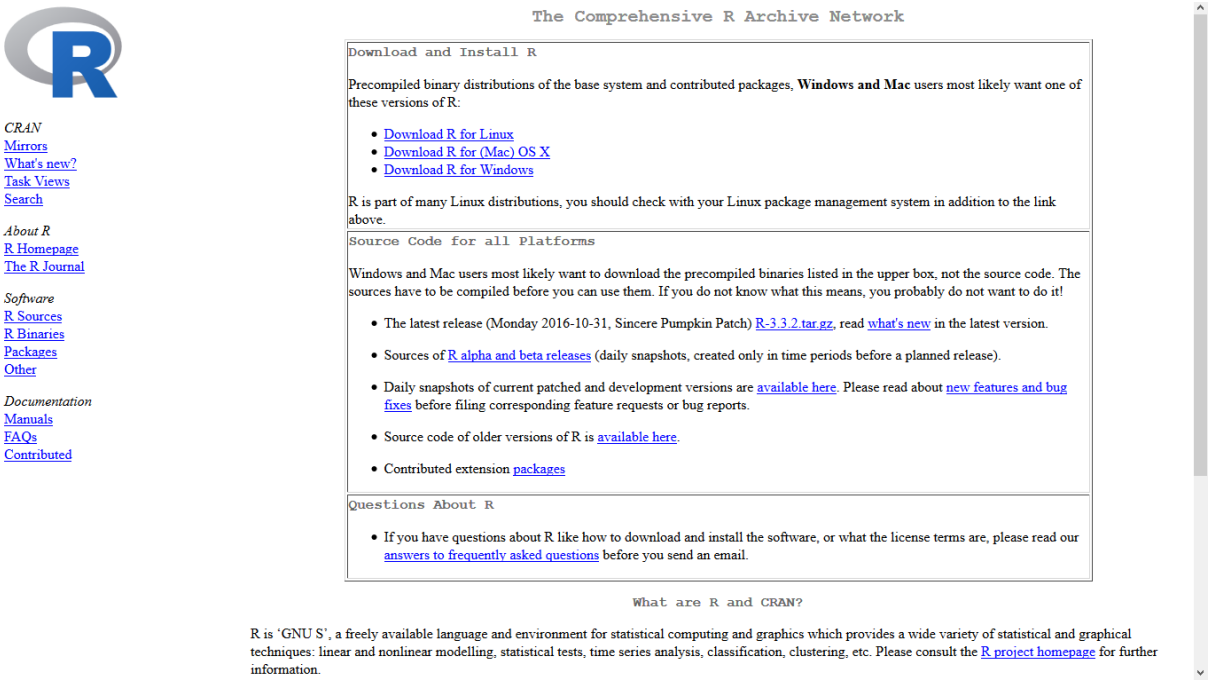
Descarga de *R*

1. Desde el explorador de Internet, entrar en *www.r-project.org*. A continuación, hacer clic con el botón izquierdo del ratón en el enlace *download R*, situado en la parte inferior de la ventana dentro del recuadro *Getting Started*.



2. Localizar España (*Spain*) en el listado de Servidores y pinchar sobre el enlace correspondiente a Madrid (*Spanish National Research Network, Madrid*), habitualmente dado por http://cran.rediris.es

3. En el cuadro *Download and Install R*, seleccionar la opción *Download R for (Mac) OS X*.



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Monday 2016-10-31, Sincere Pumpkin Patch) R-3.3.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

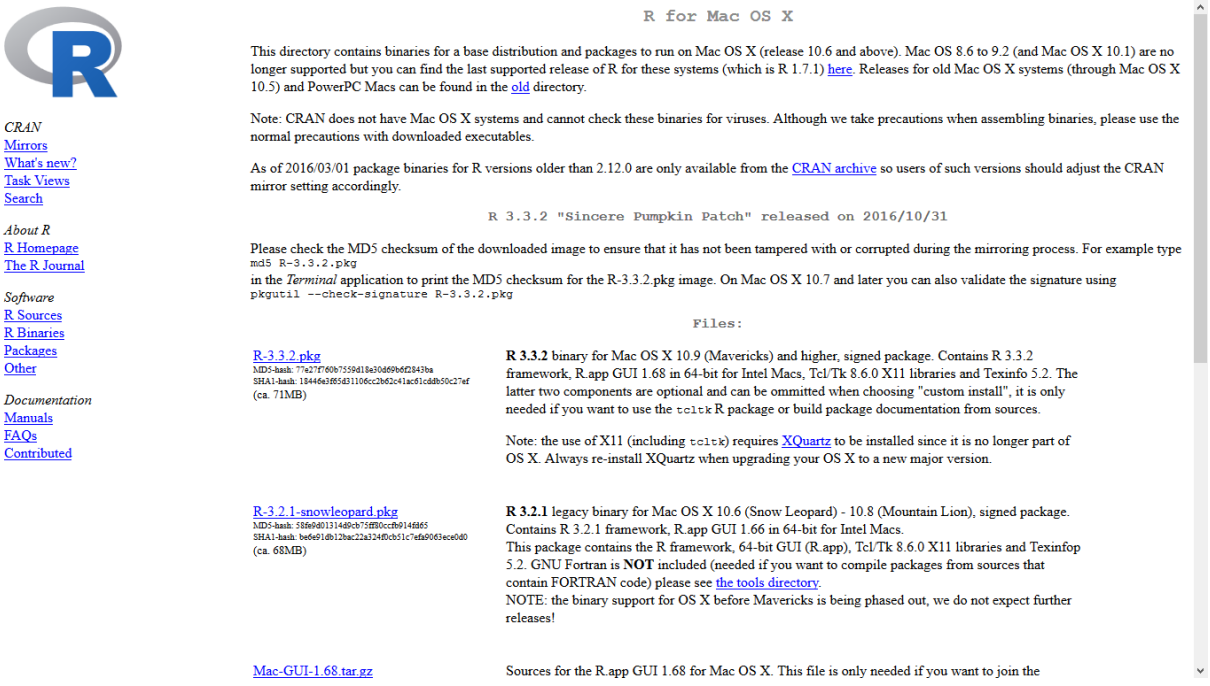
Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

4. En el apartado *Files* de la pantalla, hacer clic sobre el enlace *R-(...).pkg*. En lugar de los puntos suspensivos aparecerá la numeración de la última versión de *R*, según el Sistema Operativo instalado.



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS X 10.1) are no longer supported but you can find the last supported release of R for these systems (which is R 1.7.1) here. Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the old directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, please use the normal precautions with downloaded executables.

As of 2016/03/01 package binaries for R versions older than 2.12.0 are only available from the CRAN archive so users of such versions should adjust the CRAN mirror setting accordingly.

R 3.3.2 "Sincere Pumpkin Patch" released on 2016/10/31

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type `md5 R-3.3.2.pkg` in the *Terminal* application to print the MD5 checksum for the R-3.3.2.pkg image. On Mac OS X 10.7 and later you can also validate the signature using `pkgutil --check-signature R-3.3.2.pkg`

Files:

R-3.3.2.pkg
MD5-hash: 77e27f760b7559d18e30d69b6f2843ba
SHA1-hash: 18446e3f05d31106cc2b62c41ac61cddb50c27ef
(ca. 71MB)

R 3.3.2 binary for Mac OS X 10.9 (Mavericks) and higher, signed package. Contains R 3.3.2 framework, R.app GUI 1.68 in 64-bit for Intel Macs, Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. The latter two components are optional and can be ommitted when choosing "custom install", it is only needed if you want to use the `tcltk` R package or build package documentation from sources.

Note: the use of X11 (including `tcltk`) requires XQuartz to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your OS X to a new major version.

R-3.2.1-snowleopard.pkg
MD5-hash: 58fe9d013144b9cb75ff80ccfb914fd65
SHA1-hash: be6e91db12bac22a324f0cb51c7efa9063ece0d0
(ca. 68MB)

R 3.2.1 legacy binary for Mac OS X 10.6 (Snow Leopard) - 10.8 (Mountain Lion), signed package.
Contains R 3.2.1 framework, R.app GUI 1.66 in 64-bit for Intel Macs.
This package contains the R framework, 64-bit GUI (R.app), Tcl/Tk 8.6.0 X11 libraries and Texinfo 5.2. GNU Fortran is **NOT** included (needed if you want to compile packages from sources that contain FORTRAN code) please see the tools directory.
NOTE: the binary support for OS X before Mavericks is being phased out, we do not expect further releases!

Mac-GUI-1.68.tar.gz

Sources for the R.app GUI 1.68 for Mac OS X. This file is only needed if you want to join the

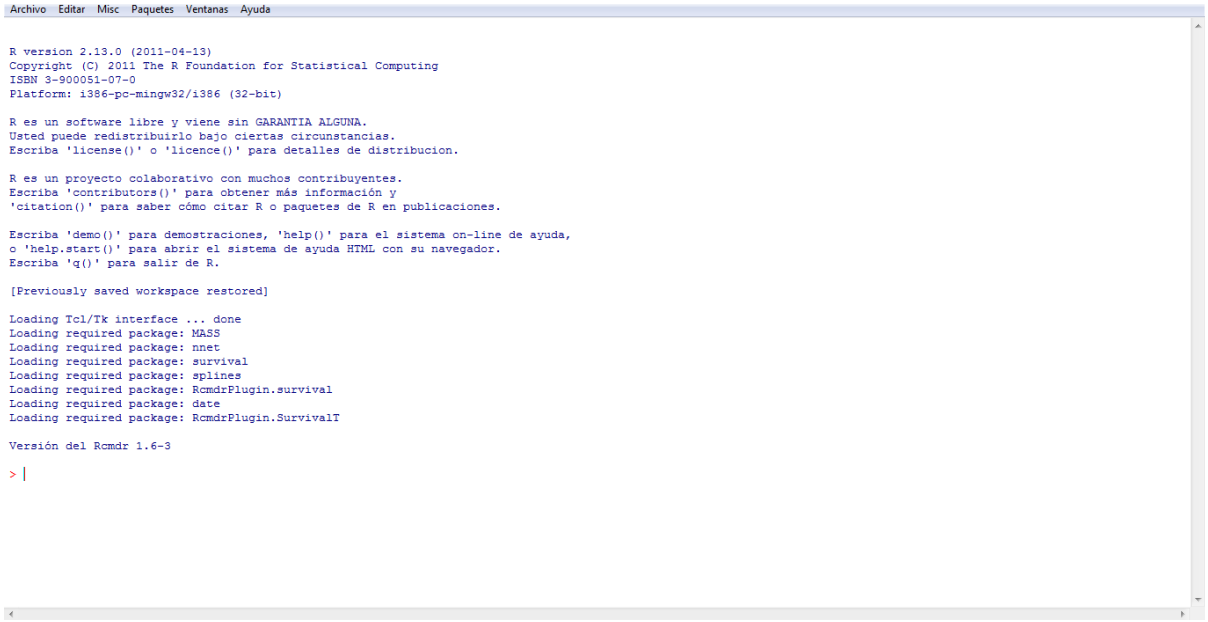
Instalación de *R*

Una vez descargado el archivo *R-(...).pkg*, el paquete se abrirá automáticamente para proceder a la instalación de *R*. Si no es así, será necesario ir a la carpeta donde se almacenó el archivo, hacer doble clic sobre él y seguir las instrucciones que aparecerán en pantalla.

Si la versión de Mac OS X es 10.9 o superior, tras la instalación de *R* será necesario volver a la misma página web en la que estaba el archivo *R-(...).pkg* y pulsar el hiperenlace *XQuartz*. Allí, habrá que descargar e instalar esta aplicación, que ya no es parte de OS X, y reiniciar el ordenador.

Instalación de *R-Commander*

Tras la instalación del lenguaje de programación *R* aparecerá en la carpeta *Aplicaciones* del *Finder* un icono con la forma  . Haciendo doble clic sobre este icono se accederá a la consola o pantalla de inicio de *R*, cuya apariencia es similar esta:



El símbolo >, en color rojo, indica que *R* está preparado para recibir instrucciones y comenzar a trabajar utilizando los comandos del lenguaje de programación.

R-Commander es un paquete adicional que deberá instalarse a continuación para trabajar en un entorno de ventanas más sencillo. Para ello deberá ejecutarse la siguiente instrucción, que ha de escribirse en la consola de *R* tal como aparece a continuación, respetando mayúsculas y minúsculas, y pulsando la tecla *Intro* tras su escritura:

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

Comenzar una sesión de trabajo con *R-Commander*

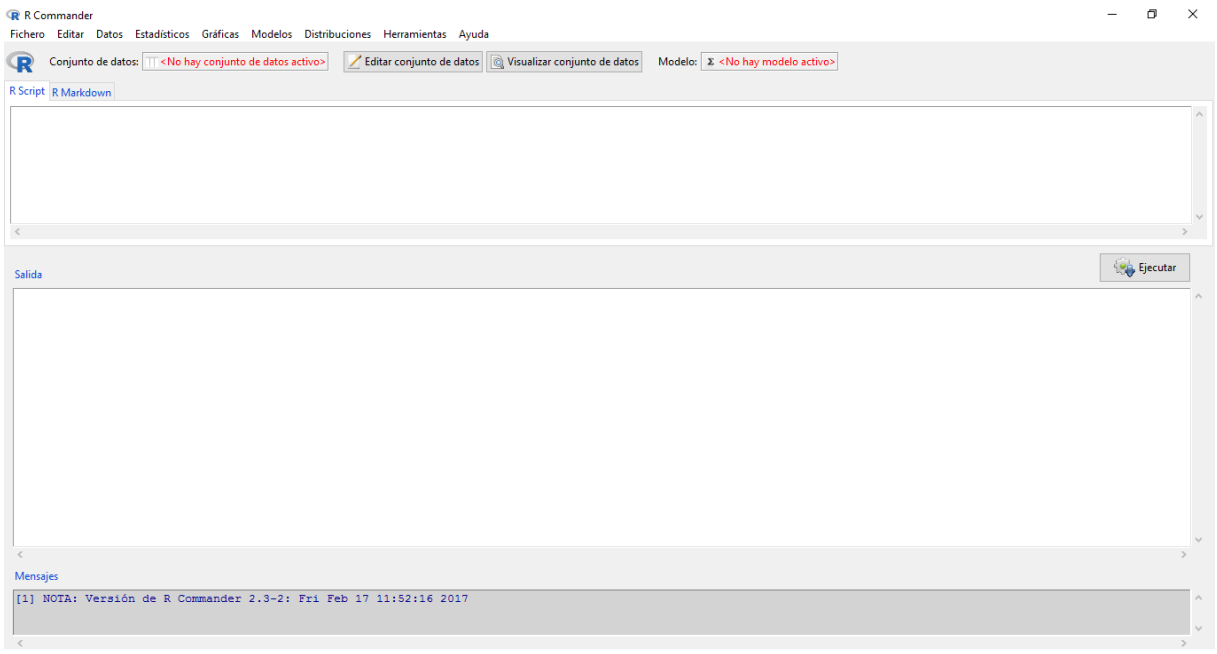
Con los pasos seguidos anteriormente, *R-Commander* quedará instalado en el Mac de forma permanente. Sin embargo, siempre que se pulse el icono se abrirá por defecto la consola de *R* pero no la de *R-Commander*. Para activar esta interfaz habrá que escribir *library(Rcmdr)* en la pantalla de inicio de *R* y pulsar la tecla *Intro* del teclado. Este procedimiento tendrá que repetirse en cada sesión de trabajo, siempre que se desee trabajar con *R-Commander*.

La nueva ventana abierta será el entorno de trabajo *R-Commander*, que podrá maximizarse para tener un campo visual más amplio. A veces, esta interfaz es lenta en Mac. Si esto ocurriera, se puede solucionar instalando el Entorno de Desarrollo Integrado *RStudio* (www.rstudio.com) y activando *R-Commander* dentro de él, en lugar de abrirlo desde *R*.

NOCIONES BÁSICAS

Explorar el menú de opciones y las ventanas de *R-Commander*

La barra de menú de *R-Commander*, situada en la parte superior de la pantalla, es similar a la de otros programas estadísticos, intuitiva y relativamente sencilla de manejar.



En el menú principal se visualizan diferentes pestañas, cuya utilidad es la siguiente:

Fichero	Permite cambiar el directorio de trabajo, guardar archivos de instrucciones y resultados y salir del programa, entre otras opciones.
Editar	Realiza las funciones propias de cualquier editor de texto.
Datos	Crea una base de datos en formato *R* o importa bases de datos de otros programas (SPSS, Minitab, Stata, Excel, Access y Dbase). Además, contiene opciones para calcular variables nuevas o recodificar, tipificar y modificar las variables activas.
Estadístico	Cubre la mayoría de los análisis estadísticos básicos, incluyendo modelos de regresión multivariante para variables dependientes cuantitativas y cualitativas.
Gráficas	Realiza análisis exploratorio de datos y descripción de la información mediante gráficos.
Modelos	Contiene opciones para realizar el diagnóstico de los modelos y comprobar su bondad de ajuste.
Distribuciones	Muestra la función de densidad o la función de probabilidad de las distribuciones continuas y discretas más usuales.
Herramientas	Modifica la configuración por defecto de *R-Commander*, carga nuevos paquetes de *R* e instala complementos de *R-Commander* (*plugins*) para realizar análisis estadísticos que no están incorporados por defecto.
Ayuda	Ofrece ayuda sobre el funcionamiento de *R-Commander*, incluyendo la versión en Castellano del documento *Iniciación a R-Commander* elaborado por John Fox.

Bajo el menú principal hay un submenú con dos botones, uno para editar y otro para visualizar la base de datos activa. Junto a ellos se muestran dos etiquetas en las que aparecerá el nombre del conjunto de datos y el nombre del modelo estadístico que el usuario está utilizando en cada momento.

Por último, debajo del submenú, se encuentra la ventana de trabajo dividida en tres partes. La primera corresponde a la ventana de instrucciones, donde automáticamente aparecerán la

sintaxis y los comandos de todos los análisis realizados. La segunda es la ventana de resultados, espacio donde se mostrarán sucesivamente los resultados de cada análisis estadístico. Finalmente, la parte inferior recogerá los mensajes que el software genere durante la sesión de trabajo. Esta última es especialmente importante para monitorizar los mensajes de error, localizar su procedencia y proceder a la corrección.

Definir el directorio de trabajo

Habitualmente, los archivos que se utilizan en una investigación suelen estar almacenados en una carpeta de proyecto. Para facilitar la búsqueda de estos archivos durante una sesión de trabajo con *R-Commander* es aconsejable definir la carpeta o directorio de trabajo en el que se encuentran. De esta forma, *R-Commander* buscará y guardará allí, agilizando el proceso de análisis. Esta acción se realiza desde el menú principal a través de la secuencia:

Fichero – Cambiar directorio de trabajo

La carpeta correspondiente se buscará en el cuadro de diálogo abierto. Una vez localizada quedará activada en memoria pulsando el botón *Aceptar*.

Limpiar la ventana de trabajo

A menudo, las ventanas de instrucciones, resultados y mensajes se llenan de información que deja de ser necesaria una vez que se ha realizado el análisis de datos y los resultados se han pasado a un procesador de textos. Para limpiar cualquiera de estas ventanas bastará con hacer clic con el ratón sobre ella y pulsar la siguiente secuencia desde del menú principal:

Editar – Limpiar ventana

Esta acción borrará toda la información de la ventana, aunque también es posible seleccionar sólo una parte del texto con el ratón y pulsar posteriormente la tecla *Suprimir* (*Supr*) del teclado para eliminarlo.

El procedimiento se repetirá para borrar el contenido del resto de ventanas. En caso de limpiar una ventana por error es posible restaurar su información pulsando *Editar – Deshacer*.

Salir de *R-Commander* y de *R*

Antes de salir del programa puede ser útil guardar los resultados de la sesión de trabajo mediante la secuencia:

Fichero – Guardar los resultados como

Esta opción almacenará en un archivo de texto todos los resultados que se encuentren en la ventana correspondiente. Además, es conveniente guardar la base de datos en formato *R-Commander* a través de la opción:

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

De esta forma se archivarán tanto los datos originales como las nuevas variables y transformaciones realizadas durante la sesión. Para recuperar esta base de datos bastará con abrir de nuevo el archivo mediante la secuencia *Datos – Cargar conjunto de datos*.

Una vez guardado lo necesario, se puede salir del programa haciendo clic sobre:

Fichero – Salir – De Commander y R

Al cerrar, el programa recordará si se desea guardar alguna información.

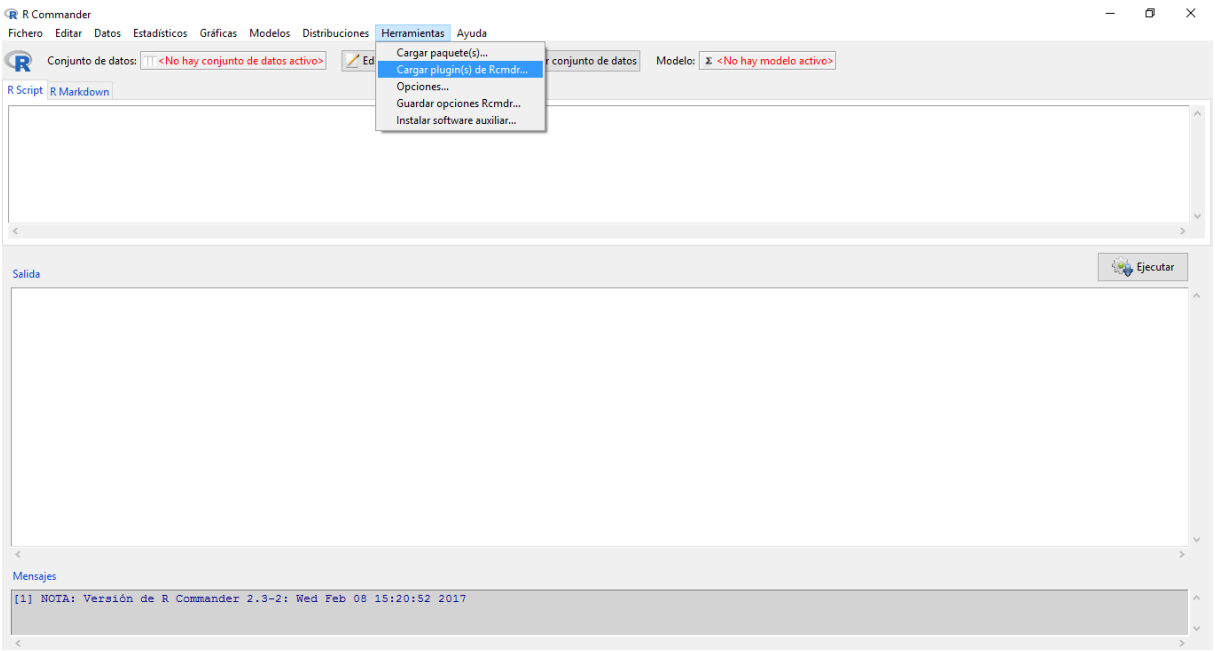
Instalar nuevas aplicaciones para *R-Commander*

En ocasiones, *R-Commander* dispone de paquetes opcionales que pueden ser instalados para realizar análisis estadísticos específicos. Cada uno de estos paquetes se denomina *Plugin*, y entre ellos se pueden encontrar, a modo de ejemplo, los desarrollados para análisis de supervivencia (*RcmdrPlugin.survival*) y análisis de curvas ROC (*RcmdrPlugin.ROC*). Para su instalación deberán ejecutarse secuencialmente las siguientes instrucciones desde la consola de *R*, escribiéndolas tal como aparece a continuación, y pulsando la tecla *Intro* tras la escritura de cada una de ellas:

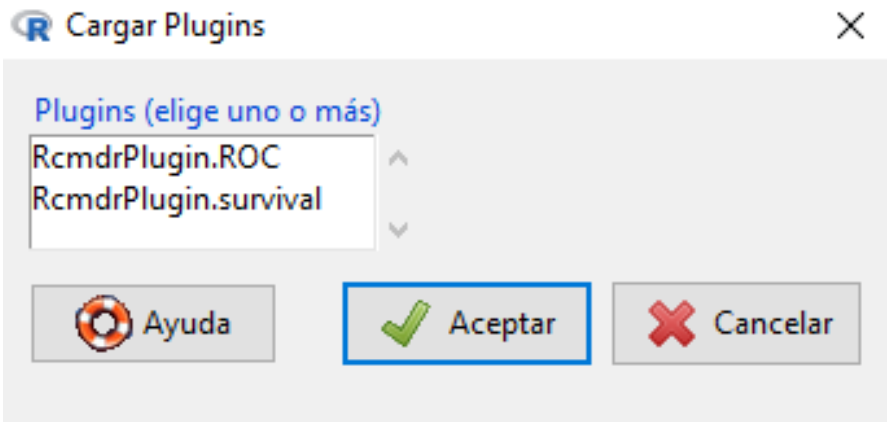
```
install.packages("RcmdrPlugin.survival", dependencies=T)
install.packages("RcmdrPlugin.ROC", dependencies=T)
```

En caso de que *R* solicite la conexión con un servidor, se seleccionará alguno de los disponibles para España (*Spain*). Una vez instalados ambos paquetes, será necesario salir de *R* y de *R-Commander* para que, al volver a abrir la interface, ésta reconozca los nuevos *Plugin* instalados. Para comprobar que la instalación se ha realizado correctamente, será necesario abrir de nuevo *R-Commander* y ver que los *Plugin* aperecen dentro del menú de *R-*

Commander pulsando la secuencia: *Herramientas – Cargar plugin(s) de Rcmdr*. De no ser así, habrá que volver a realizar la instalación.



Para cargar cualquiera de los *Plugin* instalados bastará con seleccionarlo desde el menú mencionado anteriormente y pulsar el botón *Aceptar*. Esta acción reiniciará automáticamente *R-Commander*, por lo que es aconsejable realizarla antes de comenzar una sesión de trabajo.



Si, en Sistemas Operativos Windows, se desea que al ejecutar *R* se active automáticamente tanto *R-Commander* como cualquiera de los *Plugin* instalados, será necesario modificar el archivo *Rprofile.site* mencionado en el apartado de instalación de *R-Commander* para Windows (Fase 2) de esta monografía. Para ello se abrirá el Bloc de Notas, o cualquier otro editor de texto, con permisos de administrador. A continuación, se abrirá el archivo

Rprofile.site, situado en la carpeta C:\Archivos de programa\R\R-(…)\etc (o bien C:\Program Files\R\R-(…)\etc), y se añadirán las siguientes instrucciones:

```
local({
old <- getOption("defaultPackages")
options(defaultPackages = c(old, "Rcmdr"))
options(Rcmdr=list(plugins=c("RcmdrPlugin.survival", “RcmdrPlugin.ROC”)))
})
```

Al escribir las instrucciones, es muy importante respetar las letras mayúsculas y minúsculas y no olvidar ningún paréntesis, corchete o entrecomillado.

Por último se guardará el archivo *Rprofile.site* dentro de la misma carpeta, reemplazando al archivo original, teniendo en cuenta que su extensión ha de ser *.site* y no *.txt* o *.dat* como puede aparecer por defecto en algunos editores de texto.

3

GESTIÓN DE BASES DE DATOS CON *R-COMMANDER*

Habitualmente, el diseño, elaboración y gestión de bases de datos se realiza mediante programas informáticos específicos que permiten el procesamiento de la información de forma rápida y estructurada. Este tipo de software se denomina sistema gestor de bases de datos, siendo Microsoft Access, dBase o FileMaker algunos de los más populares.

El propósito de un sistema gestor de bases de datos es almacenar y organizar la información. Sin embargo, no permite realizar análisis estadísticos avanzados con los datos disponibles. Para ello es necesario disponer de otro programa informático que capture la información procedente del sistema gestor de bases de datos y realice el análisis estadístico apropiado. Actualmente existe una oferta muy amplia de software estadístico. De ellos, quizá SPSS, Stata y SAS sean los más utilizados.

Muchas veces, la gestión de bases de datos consume una parte importante del tiempo invertido en un proyecto de investigación, por lo que contar con una herramienta potente que ayude a procesar eficazmente la información es tan importante como disponer del software estadístico apropiado para analizar los datos. Conscientes de esta necesidad, los principales desarrolladores de software estadístico diseñan sus programas para que cumplan la doble función de gestionar grandes bases de datos y analizar estadísticamente la información en una fase posterior. De esta forma se evita que el usuario necesite aprender dos sistemas informáticos diferentes.

R-Commander no ha sido diseñado para funcionar como sistema gestor de bases de datos, por lo que no es aconsejable su uso para almacenar la información. En su lugar, es preferible utilizar un sistema gestor de bases de datos externo y capturar posteriormente la información para llevar a cabo el análisis estadístico. En cualquier caso, el uso del editor de datos *R-Commander* puede ser de utilidad para introducir directamente pequeños conjuntos de datos, motivo por el que los siguientes apartados describen cómo realizar este proceso además de importar bases de datos elaboradas con otros programas informáticos.

Los contenidos de este capítulo están basados en el caso práctico Accidentes por pinchazo en profesionales de enfermería.

CONCEPTOS BÁSICOS

La información correspondiente a cada uno de los profesionales que participó en el estudio de accidentes por pinchazo se recogió en una ficha individual con un código personal de identificación. En ella se registraron, además, las siguientes características del profesional: grupo al que había sido asignado (formación o no formación), estado al final del seguimiento (accidentado o no accidentado), edad y sexo (hombre o mujer). La principal hipótesis de investigación era que el programa de formación implementado es eficaz para disminuir los accidentes por pinchazo, de manera que la proporción de accidentes sería menor en el grupo de profesionales que recibió formación sobre medidas preventivas. La comprobación de esta hipótesis requerirá el uso de métodos estadísticos concretos, sin embargo, antes de proceder con el análisis de datos es necesario organizar, procesar y almacenar la información en una base de datos electrónica.

Estructura de una base de datos

Aunque existen muchos tipos y modelos de bases de datos, las utilizadas para el análisis estadístico de la información tienen estructura rectangular, con una apariencia similar a esta:

Código	Grupo	Estado	Edad	Sexo
00004	Formación	No accidentado	45	Hombre
00006	No Formación	No accidentado	50	Hombre
00014	No Formación	No accidentado	55	Hombre
00015	Formación	No accidentado	26	Mujer
00018	Formación	No accidentado	58	Mujer
00019	Formación	No accidentado	21	Mujer
00022	Formación	No accidentado	52	Mujer
00024	Formación	No accidentado	51	Mujer
00001	Formación	Accidentado	22	Hombre
00002	No Formación	Accidentado	22	Hombre
00003	No Formación	Accidentado	22	Hombre
00005	Formación	Accidentado	30	Hombre
00007	Formación	Accidentado	34	Hombre
00008	Formación	Accidentado	23	Hombre
00009	No Formación	Accidentado	28	
00010	No Formación	Accidentado	21	Hombre
00011	No Formación	Accidentado	40	Hombre
00012	Formación	Accidentado	30	Hombre
00013	No Formación	Accidentado	35	Hombre
00016	No Formación	Accidentado		Mujer
00017	No Formación	Accidentado	50	Mujer
00020	No Formación	Accidentado	25	Mujer
00021	Formación	Accidentado	47	Mujer
00023	No Formación	Accidentado	23	Mujer
00025	No Formación	Accidentado	23	Mujer

Base de datos con información numérica y caracteres de texto.

Cada columna de la base de datos corresponde a una característica de los individuos incluidos en el estudio de accidentes por pinchazo. En esta investigación se recogió información sobre cinco características de los profesionales, siendo el código de identificación la situada en la primera columna y el sexo de los sujetos en la quinta. El nombre de cada una de ellas aparece en la cabecera de la base de datos, sombreada en color. El orden en el que se disponen las columnas es indiferente para organizar la base de datos.

Debajo de la cabecera de la base de datos aparece la información registrada, donde cada fila almacena las características de un único sujeto. Así, la primera fila de la base de datos muestra la información del profesional con código de identificación 00004, perteneciente al grupo que recibió formación, no accidentado al finalizar el seguimiento, 45 años de edad y sexo masculino. Cuando no se tiene información de alguna característica la celda correspondiente de la base de datos queda vacía, como el valor del sexo para el sujeto con código 00009 o la edad para el sujeto 00016. Es lo que se conoce como un valor perdido, dato faltante o *missing*.

Tipos de variables

Los valores de cada característica difieren de un sujeto a otro. Así, la edad del profesional de la primera fila es diferente a la edad del profesional de la segunda fila. Debido a esta variabilidad de los valores registrados, las características se denominan variables.

Habitualmente existen dos tipos de variables que pueden ser utilizadas en un análisis estadístico de datos: Cualitativas y cuantitativas.

Una variable es cualitativa cuando sus valores recogen una cualidad del individuo que no puede medirse con un instrumento ni lleva asociada unidades de medida. Así, el sexo es una variable cualitativa con dos valores, hombre y mujer, denominados categorías. Estas categorías deben estar definidas de tal forma que cada sujeto de la base de datos pueda incluirse sólo en una de ellas, de forma exclusiva e inequívoca. El sexo es una variable cualitativa nominal porque sus categorías, hombre y mujer, no tienen un orden natural preestablecido. Si se hubiese recogido la variable gravedad del accidente, con categorías leve, moderado y grave, se tendría una variable cualitativa ordinal, ya que registra una cualidad cuyos valores o categorías pueden ordenarse de forma natural de menor a mayor severidad. Aunque no es la terminología usual, *R-Commander* denomina a las variables cualitativas *factores* y a sus categorías *niveles*.

Una variable cuantitativa es una característica de los sujetos que puede expresarse mediante valores numéricos, con una unidad de medida asociada a ellos. La edad es una variable cuantitativa cuya unidad de medida es el año. Además, esta variable es continua, ya que el valor de la edad asignada a cada individuo puede tener tantos decimales como se desee dependiendo de la precisión requerida. Otras variables cuantitativas, como el número de hijos,

se denominan discretas porque sus valores solo pueden ser números enteros, sin decimales. *R-Commander* denomina *numérica* a cualquier tipo de variable cuantitativa.

Esta clasificación de las variables no sólo es importante para procesar y registrar adecuadamente la información. También lo es para aplicar el análisis estadístico apropiado en función del tipo de variable analizada, ya que requerirá técnicas diferentes.

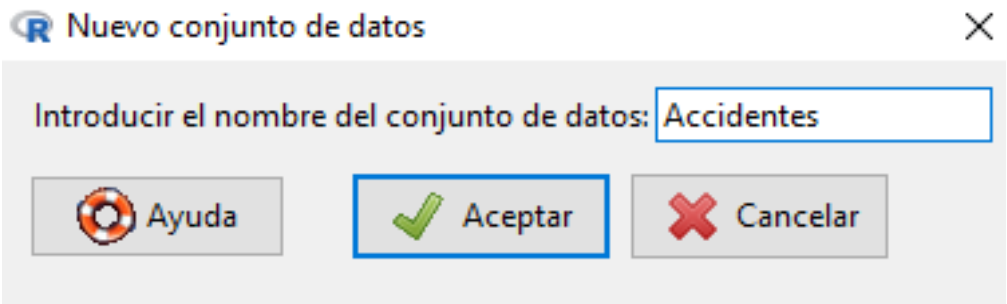
En el estudio de accidentes por pinchazo, las variables grupo, con categorías formación-no formación, estado al final del seguimiento, con categorías accidentado-no accidentado y sexo, con categorías hombre-mujer, son variables cualitativas, mientras que la edad es cuantitativa. Aunque la variable código de identificación es una variable numérica, no cuantifica ninguna medición. Sólo se utiliza para identificar a los sujetos de estudio, cumpliendo la misma función que podría hacer el DNI o el número de Seguridad Social. Por este motivo no tiene interés utilizarla en un análisis estadístico de datos.

ELABORACIÓN DE UNA BASE DE DATOS

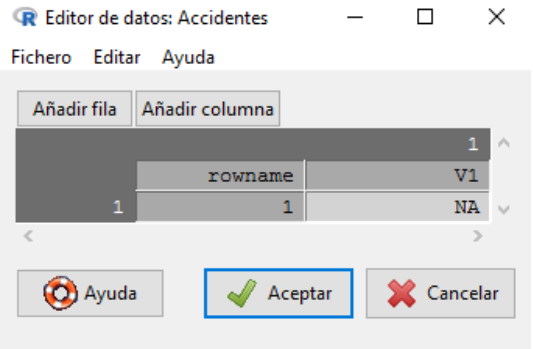
El aspecto del editor de datos de *R-Commander* es similar al de una hoja de cálculo. El acceso para crear una nueva base de datos se realiza desde el menú principal, seleccionando:

Datos – Nuevo conjunto de datos

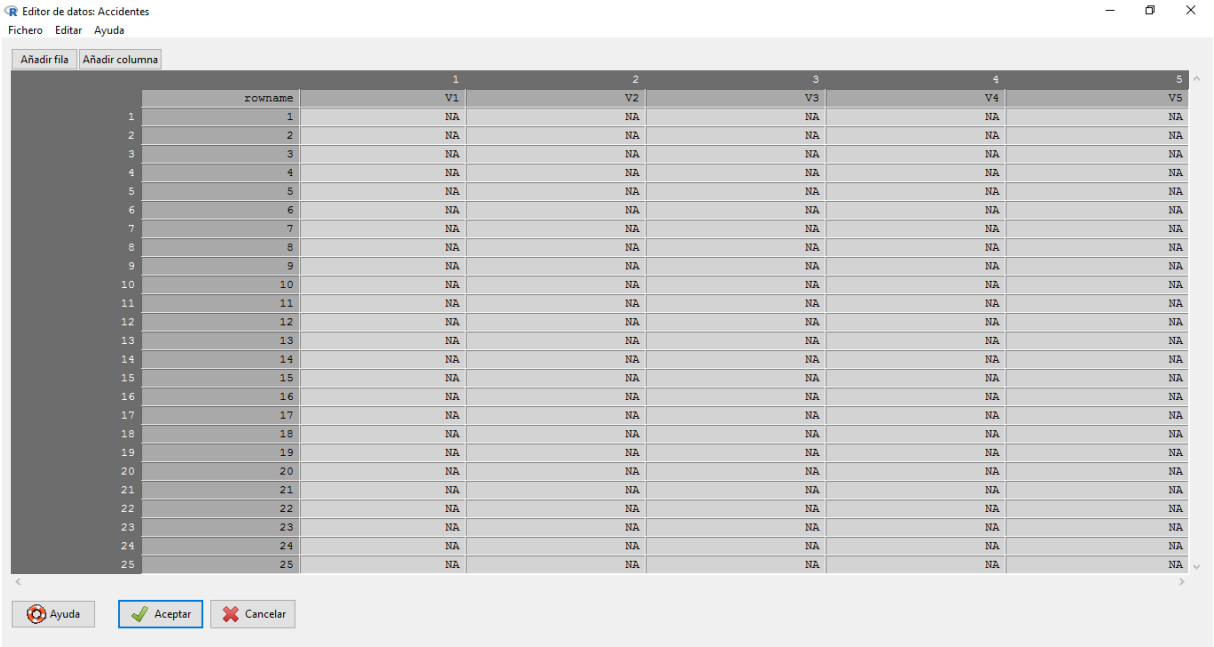
A continuación, se abrirá un cuadro de diálogo que solicita el nombre de la base de datos que se va a crear. Por defecto *R-Commander* asigna el nombre *Datos*, de manera que el usuario puede mantenerlo o escribir otro que considere más apropiado. El nombre de la base de datos puede ser cualquiera, siempre que comience por una letra y no contenga espacios ni símbolos. Para introducir la información correspondiente al estudio de accidentes por pinchazo se escribirá *Accidentes* como nombre de la base de datos, pulsando posteriormente el botón *Aceptar*.



El editor de datos se abrirá en una ventana independiente, mostrando en la cabecera el nombre de las dos primeras variables asignadas por defecto: *rowname* (número de registro) y *V1* (nombre de la primera variable).



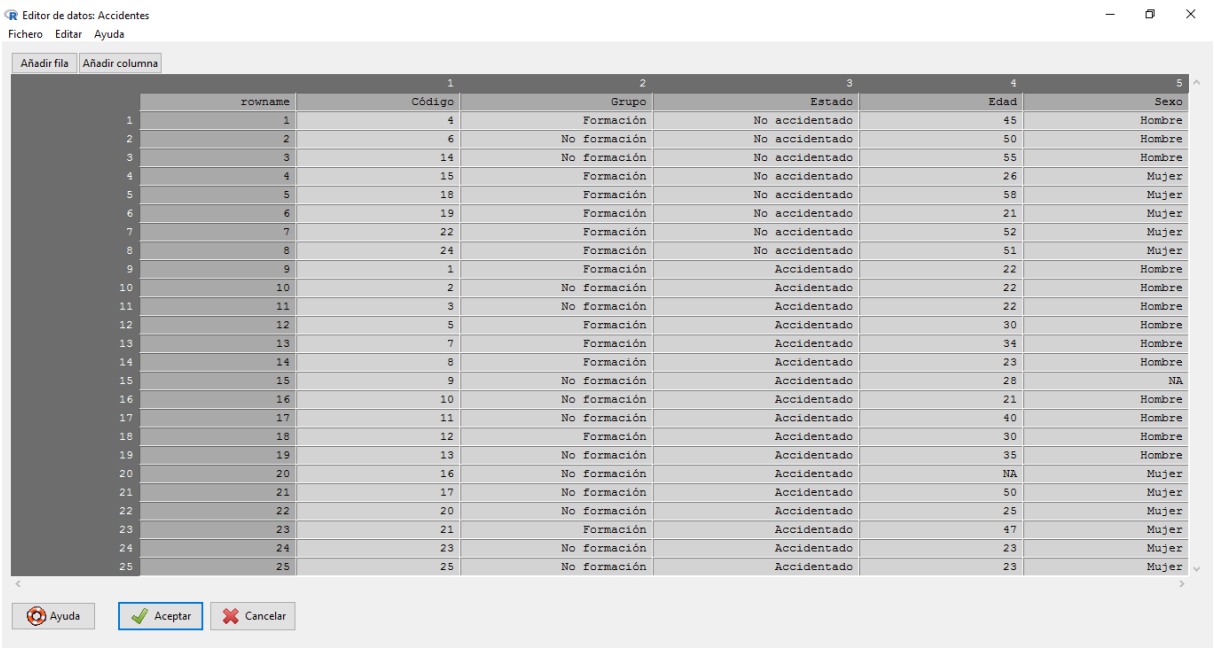
En este caso, la base de datos se compone de 25 registros (filas) y 5 variables (columnas), que deberán definirse pulsando 25 veces el botón *Añadir fila* y 5 veces el botón *Añadir columna*. Finalizado este proceso, la base de datos tendrá el siguiente formato:



	rowname	V1	V2	V3	V4	V5
1	1	NA	NA	NA	NA	NA
2	2	NA	NA	NA	NA	NA
3	3	NA	NA	NA	NA	NA
4	4	NA	NA	NA	NA	NA
5	5	NA	NA	NA	NA	NA
6	6	NA	NA	NA	NA	NA
7	7	NA	NA	NA	NA	NA
8	8	NA	NA	NA	NA	NA
9	9	NA	NA	NA	NA	NA
10	10	NA	NA	NA	NA	NA
11	11	NA	NA	NA	NA	NA
12	12	NA	NA	NA	NA	NA
13	13	NA	NA	NA	NA	NA
14	14	NA	NA	NA	NA	NA
15	15	NA	NA	NA	NA	NA
16	16	NA	NA	NA	NA	NA
17	17	NA	NA	NA	NA	NA
18	18	NA	NA	NA	NA	NA
19	19	NA	NA	NA	NA	NA
20	20	NA	NA	NA	NA	NA
21	21	NA	NA	NA	NA	NA
22	22	NA	NA	NA	NA	NA
23	23	NA	NA	NA	NA	NA
24	24	NA	NA	NA	NA	NA
25	25	NA	NA	NA	NA	NA

Para introducir el nombre de la primera variable habrá que hacer clic con el botón izquierdo del ratón sobre el texto *V1*. Esta acción activará la celda para poder borrar el texto *V1* y escribir en su lugar el nombre de la primera variable, en este caso *Código*. El nombre de la variable quedará guardado tras pulsar la tecla del cursor hacia la derecha, o haciendo clic con el ratón sobre la siguiente variable. El mismo procedimiento se repetirá para definir los nombres de las variables *Grupo*, *Estado*, *Edad* y *Sexo*. Como norma general, el nombre de las variables no puede comenzar con un valor numérico ni contener espacios o símbolos diferentes al punto (.) o guión bajo (_).

Una vez definidas las variables, los datos se introducen en las celdas de la base de datos teniendo en cuenta que cada fila se corresponde con la información de un sujeto. Para ello bastará con situar el cursor en la celda correspondiente con ayuda del ratón y escribir el valor de la variable correspondiente. La introducción del siguiente valor puede hacerse desplazando el cursor con las flechas del teclado o pulsando con el ratón sobre la siguiente celda. Habitualmente, *R-Commander* señala la celda en la que está situado el cursor con un fondo blanco. Tras finalizar la introducción de datos, se obtendrá una base de datos con el siguiente formato:


Editor de datos: Accidentes

Fichero Editar Ayuda

Añadir fila | Añadir columna

		1	2	3	4	5
	rowname	Código	Grupo	Estado	Edad	Sexo
1	1	4	Formación	No accidentado	45	Hombre
2	2	6	No formación	No accidentado	50	Hombre
3	3	14	No formación	No accidentado	55	Hombre
4	4	15	Formación	No accidentado	26	Mujer
5	5	18	Formación	No accidentado	58	Mujer
6	6	19	Formación	No accidentado	21	Mujer
7	7	22	Formación	No accidentado	52	Mujer
8	8	24	Formación	No accidentado	51	Mujer
9	9	1	Formación	Accidentado	22	Hombre
10	10	2	No formación	Accidentado	22	Hombre
11	11	3	No formación	Accidentado	22	Hombre
12	12	5	Formación	Accidentado	30	Hombre
13	13	7	Formación	Accidentado	34	Hombre
14	14	8	Formación	Accidentado	23	Hombre
15	15	9	No formación	Accidentado	28	NA
16	16	10	No formación	Accidentado	21	Hombre
17	17	11	No formación	Accidentado	40	Hombre
18	18	12	Formación	Accidentado	30	Hombre
19	19	13	No formación	Accidentado	35	Hombre
20	20	16	No formación	Accidentado	NA	Mujer
21	21	17	No formación	Accidentado	50	Mujer
22	22	20	No formación	Accidentado	25	Mujer
23	23	21	Formación	Accidentado	47	Mujer
24	24	23	No formación	Accidentado	23	Mujer
25	25	25	No formación	Accidentado	23	Mujer

Ayuda | Aceptar | Cancelar

Si la variable es cuantitativa, las celdas situadas en la columna deberán contener únicamente valores numéricos. Si la variable es cualitativa, el valor de la celda será el nombre de la categoría a la que pertenece el sujeto. Este texto deberá ir sin entrecomillar, utilizando siempre la misma combinación de letras mayúsculas y minúsculas, ya que *R-Commander* distingue entre ambos tipos de caracteres y tomará como categorías diferentes los textos "*No accidentado*" y "*no accidentado*". El nombre de cada categoría puede estar formado por varias palabras separadas por espacios y símbolos. Las celdas correspondientes a valores faltantes pueden quedar con el texto *NA*. Este símbolo corresponde a las iniciales inglesas del término *No Available* (no disponible).

El editor de datos presenta un menú en la parte superior con las opciones *Fichero*, *Editar* y *Ayuda*. La primera opción se utilizará para cerrar el editor de datos cuando la base de datos esté completa. La segunda para copiar, pegar o borrar la celda en la que esté situado el cursor o bien para añadir o borrar filas y columnas. Por último, la tercera opción ofrece ayuda sobre el editor.

Una vez finalizado el proceso de definición de variables e introducción de datos, el editor podrá cerrarse pulsando el botón *Aceptar*. Esta acción hará que la base de datos se guarde en la memoria del ordenador para proceder al análisis estadístico de datos. De hecho, *R-Commander* mostrará su nombre en color azul junto al texto "*Conjunto de datos"*, debajo del menú principal, indicando que esta base de datos es el conjunto de datos activo que utilizará para analizar.

IMPORTAR UNA BASE DE DATOS ELABORADA CON OTRO SOFTWARE

Cuando se manejan grandes cantidades de información y se desea realizar el análisis con *R-Commander*, lo habitual es diseñar y elaborar la base de datos utilizando un Sistema Gestor de Base de Datos (Microsoft Access o dBase), una hoja de cálculo (Excel) u otro software estadístico (SPSS, Minitab o STATA) e importar posteriormente los datos con *R-Commander* para su análisis.

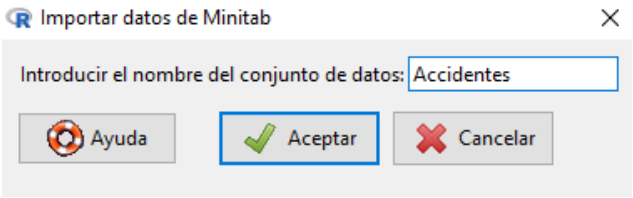
La captura de una base de datos externa puede hacerse desde la opción del menú principal

Datos – Importar datos

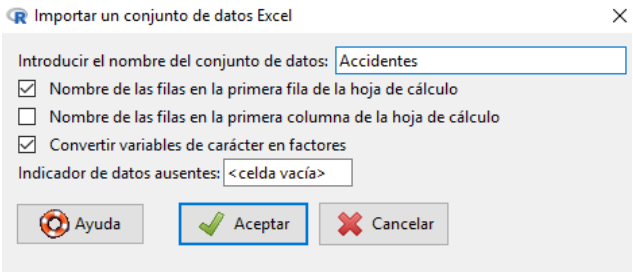
A continuación, se seleccionará el tipo de archivo que se desea importar y aparecerá un cuadro de diálogo en el que se especificarán las opciones de importación. Este cuadro siempre muestra, al menos, la opción "*Introducir el nombre del conjunto de datos"*. En el espacio reservado a la derecha de este texto se introducirá el nombre de la base de datos que se va a crear. Por defecto *R-Commander* asigna el nombre *Datos*, pero puede escribirse otro más apropiado que comience por una letra y no contenga espacios ni símbolos. Este nombre no es el archivo en el que está almacenada la base de datos, sino el nombre interno que usará *R-Commander* para trabajar con ella. Para capturar la información correspondiente al estudio de accidentes por pinchazo se escribirá *Accidentes*. El resto de opciones del cuadro de diálogo dependerá del tipo de archivo a importar, como se muestra a continuación.

Archivos Excel, SAS o Minitab

Los archivos procedentes del software estadístico Minitab o SAS no requieren más información que el nombre del conjunto de datos. Una vez escrito, bastará con pulsar el botón *Aceptar*.



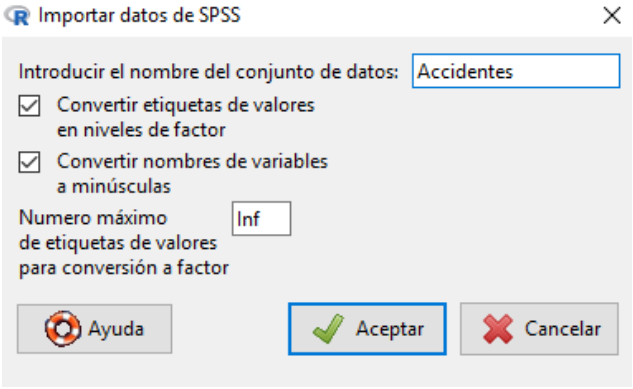
En los archivos Excel, la base de datos debe estar grabada en una hoja de cálculo con una estructura similar a la definida en el apartado *Estructura de una base de datos*. Es aconsejable que la primera fila de la hoja contenga el nombre de las variables. De esta forma se evitará tener que definirlas posteriormente en *R-Commander*.



No importa si las celdas de la hoja de cálculo o base de datos están definidas con formato texto o numérico. *R-Commander* siempre importará los números como variable numérica (cuantitativa) y el texto como variable carácter (cualitativa), para lo que es aconsejable mantener marcada la opción "*Convertir variables de carácter en factores*".

Archivos SPSS

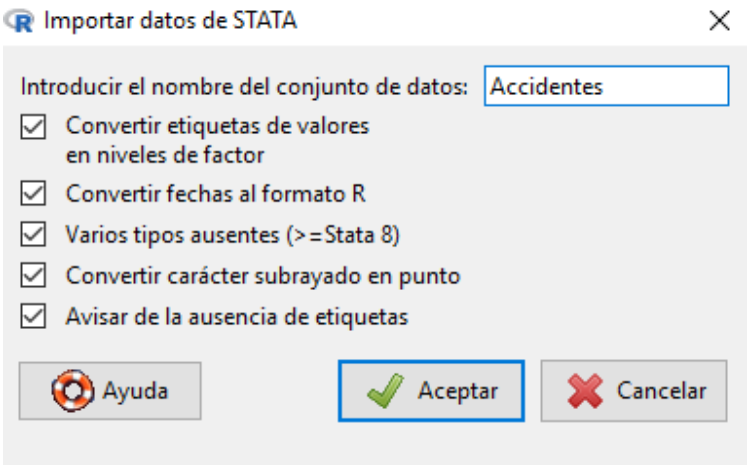
Si el archivo fue almacenado con el software estadístico SPSS, es importante activar la opción "*Convertir etiquetas de valores en niveles de factor"* para que *R-Commander* reconozca e importe el nombre de las categorías de cada variable cualitativa. En caso contrario, sólo capturará el valor numérico de cada categoría, sin su etiqueta.



La opción "*Número máximo de etiquetas de valores para conversión a factor"* hace referencia al máximo de categorías que puede tener una variable cualitativa para proceder a importar sus etiquetas. En principio, este valor suele dejarse en *Infinito*, como aparece en el cuadro de diálogo por defecto. Si este valor fuese 2, las etiquetas de las variables cualitativas con tres o más categorías no se importarían.

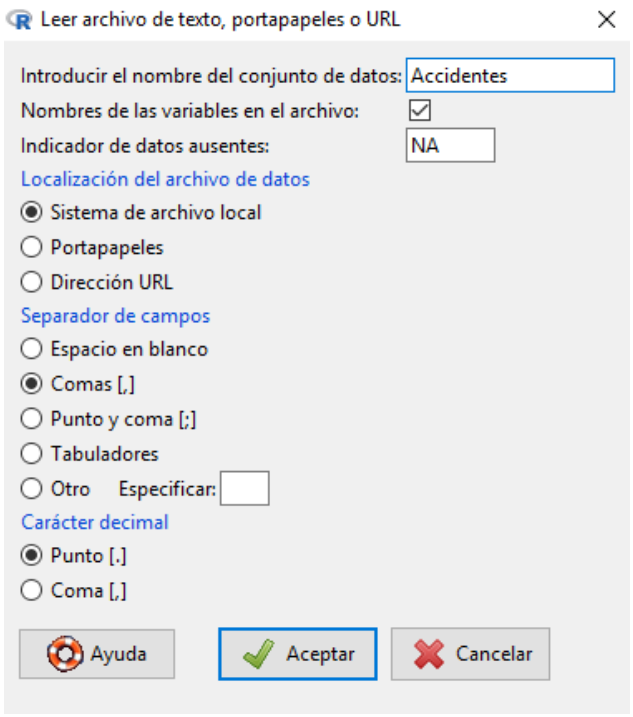
Archivos STATA

Al igual que en SPSS, cuando la base de datos está grabada con el software estadístico STATA es importante activar la opción "*Convertir etiquetas de valores en niveles de factor"*. De esta forma *R-Commander* reconocerá el nombre de las categorías de cada variable cualitativa cuando importe la información. El resto de opciones suelen dejarse activadas por defecto.



Archivos de texto

En ocasiones, la base de datos que se desea importar puede estar almacenada en un formato de archivo que *R-Commander* no reconoce directamente. En ese caso, la base de datos ha de ser capturada con el software que se utilizó para su diseño, exportarla en formato texto (*.txt*) y posteriormente importar este archivo con *R-Commander*.



Es aconsejable que la primera fila del archivo de texto contenga el nombre de las variables y activar la opción "*Nombre de las variables en el archivo*" del cuadro de diálogo para que *R-Commander* las reconozca. Además, el separador de campos deben ser comas y no espacios en blanco o tabuladores, especialmente cuando los valores de las variables cualitativas son textos que ya contienen espacios. Así, si el separador de campos fuese un espacio en blanco y el primer registro es un sujeto con los siguientes valores:

00004 Formación No accidentado 45 Hombre

R-Commander interpretaría "No accidentado" como dos valores de dos variables cualitativas diferentes: Por un lado "No" y por otro "accidentado" porque están separados por un espacio. En cambio, si el separador de campos fuese una coma y el primer registro estuviese definido de la siguiente forma

00004, Formación, No accidentado, 45, Hombre

R-Commander interpretaría que hay 5 variables y "No accidentado" es una categoría de la tercera variable.

Captura de la base de datos

Una vez definidas las opciones del cuadro de diálogo descritas anteriormente, dependiendo del tipo de archivo a importar, se pulsará el botón *Aceptar*. *Se* abrirá entonces una ventana en la que podrá localizarse la carpeta y el archivo que contiene la base de datos, denominado en este caso *Accidentes por pinchazo*. El nombre de este archivo puede ser cualquiera y contener espacios o símbolos. Una vez capturado, *R-Commander* almacenará su información con el nombre definido inicialmente en el cuadro de diálogo de importación. Este nombre aparecerá en color azul junto al texto "*Conjunto de datos"*, debajo del menú principal. Pulsando la opción "*Visualizar conjunto de datos"*, situada a la derecha del menú, se puede comprobar si la captura de la base de datos se ha realizado correctamente.

Las celdas que no contengan valores en la base de datos original se considerarán como valores perdidos. Estos casos serán identificados por *R-Commander* con el símbolo NA en las variables cuantitativas y <NA> en las cualitativas.

COMPLETAR INFORMACIÓN DE VARIABLES CUALITATIVAS

En la base de datos *Accidentes por pinchazo* los valores de las variables cuantitativas son numéricos y las categorías de las variables cualitativas se definen mediante caracteres de texto, como se mostró en los apartados anteriores. Aunque este suele ser el procedimiento habitual, ocasionalmente las bases de datos también se elaboran o importan en *R-Commander* utilizando únicamente valores numéricos tanto para las variables cuantitativas como para las cualitativas. La siguiente imagen muestra una situación de este tipo, donde las categorías de la variable *Grupo* están definidas con los valores 1 y 2, haciendo referencia a las categorías *Formación* y *No formación* respectivamente. De la misma forma, las categorías de la variable *Estado* están representadas por los valores 1 (*Accidentado*) y 2 (*No accidentado*) y las categorías de la variable *Sexo* por los valores 1 (*Hombre*) y 2 (*Mujer*):

Código	Grupo	Estado	Edad	Sexo
00004	1	2	45	1
00006	2	2	50	1
00014	2	2	55	1
00015	1	2	26	2
00018	1	2	58	2
00019	1	2	21	2
00022	1	2	52	2
00024	1	2	51	2
00001	1	1	22	1
00002	2	1	22	1
00003	2	1	22	1
00005	1	1	30	1
00007	1	1	34	1
00008	1	1	23	1
00009	2	1	28	
00010	2	1	21	1
00011	2	1	40	1
00012	1	1	30	1
00013	2	1	35	1
00016	2	1		2
00017	2	1	50	2
00020	2	1	25	2
00021	1	1	47	2
00023	2	1	23	2
00025	2	1	23	2

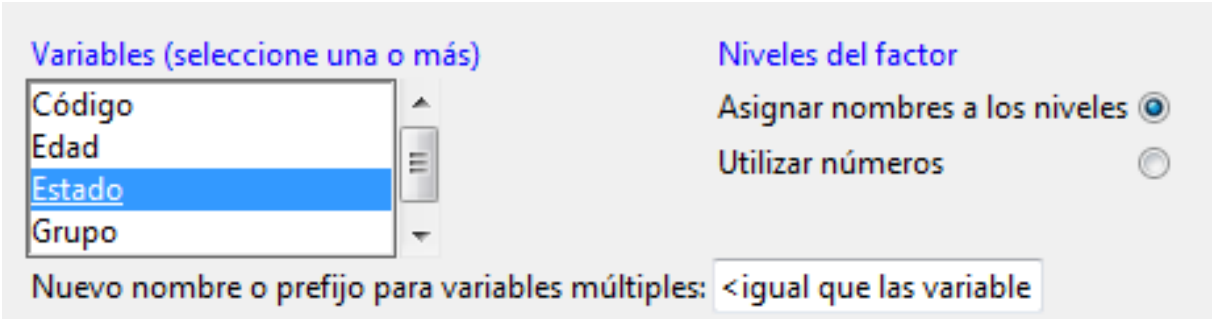
Base de datos elaborada sólo con información numérica.

Cuando se tiene una base de datos de este tipo, *R-Commander* interpretará que todas las variables son cuantitativas, puesto que sus valores son de tipo numérico.

Para evitar errores en la definición de variables y aplicar posteriormente las técnicas estadísticas apropiadas, es necesario especificar que las variables Grupo, Estado y Sexo son cualitativas y asignar una etiqueta de texto a cada una de sus categorías. Este proceso se realiza desde el menú principal pulsando la secuencia:

> *Datos - Modificar variables del conjunto de datos activo – Convertir variable numérica en factor*

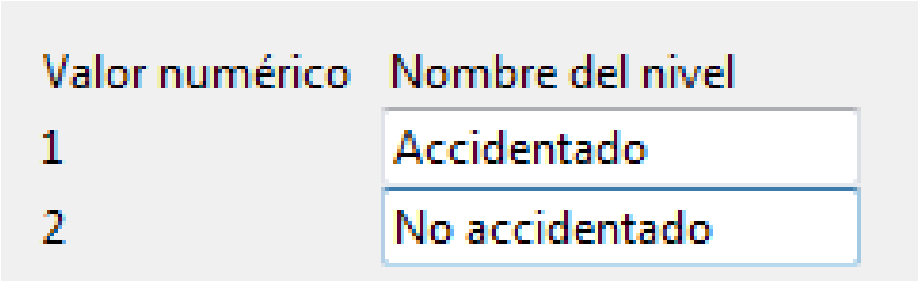
A continuación, se abrirá un cuadro de diálogo que muestra a la izquierda, en orden alfabético, el listado de variables que pueden ser transformadas en cualitativas. Haciendo clic con el ratón sobre *Estado* se marcará en azul, indicando que es la variable seleccionada.



A la derecha de la ventana, bajo el título "*Niveles del factor*", aparecen dos opciones para asignar un nombre a cada categoría de la variable *Estado*. La opción "*Asignar nombres a los niveles*" permitirá escribir una etiqueta de texto para cada categoría, mientras que la opción "*Utilizar números*" usará los valores numéricos de la variable (1 y 2) como nombres de sus categorías. La primera opción es la más recomendable y la utilizada por defecto por *R-Commander*.

Por último, en la opción "*Nuevo nombre o prefijo para variables múltiples*", situada en la parte inferior de la ventana, se puede especificar un nombre nuevo para la variable que incorporará ya los nombres de las categorías. Por ejemplo, se podría escribir en el recuadro blanco el nombre *Estado.etiquetas*. Esto permitirá mantener en la base de datos la variable *Estado* original, definida como cuantitativa, y añadir otra columna que contendrá la nueva variable *Estado.etiquetas* con una etiqueta para cada categoría. En general, esta opción no es muy recomendable, puesto que duplica variables y aumenta innecesariamente el tamaño de la base de datos. Por ello, a no ser que haya alguna razón especial, es conveniente dejar este espacio sin cumplimentar, en cuyo caso *R-Commander* incorporará directamente el nombre de las categorías a la variable *Estado* original, sin duplicarla.

Una vez definidas las opciones se pulsará el botón *Aceptar*. Si no se ha especificado un nuevo nombre para la variable, *R-Commander* mostrará un aviso en el que recuerda que la variable *Estado* ya existe y preguntará si se desea añadir el nombre de las categorías sobre ella. Una respuesta afirmativa dará paso a una nueva ventana en la que se podrá escribir el nombre de cada categoría: *Accidentado* para el valor numérico 1 y *No accidentado* para el valor numérico 2.



Si la variable tuviese más categorías, sus valores numéricos aparecerían ordenados uno debajo de otro para introducir sucesivamente los nombres. Tras pulsar el botón *Aceptar* la variable *Estado* quedará definida como cualitativa, incorporando las etiquetas que definen cada una de sus categorías.

El mismo procedimiento se repetirá para nombrar las categorías del resto de variables cualitativas.

Este proceso, iniciado en la ventana "*Convertir variables numéricas en factores*", permite seleccionar varias variables a la vez dejando pulsada la tecla *Control* (*Ctrl*) del teclado. De esta forma *R-Commander* solicitará los nombres de las categorías de cada variable de forma sucesiva, permitiendo ahorrar algunos pasos con respecto a tratar las variables de una en una. Sin embargo, cuando las variables seleccionadas tienen el mismo número de categorías, *R-Commander* asignará a todas ellas los nombres de las categorías definidas para la primera variable.

OPERACIONES USUALES CON BASES DE DATOS ACTIVAS

Una vez que la base de datos se encuentra activa en memoria, *R-Commander* ofrece varios procedimientos adicionales para gestionar su información, la mayoría de ellos localizados en el desplegable *Datos* del menú principal. A continuación, se describen los más utilizados antes de comenzar el análisis estadístico o durante el desarrollo del mismo.

Visualizar y editar la información de una base de datos

Debajo del menú principal de *R-Commander* hay dos botones: "*Visualizar conjunto de datos"* y "*Editar conjunto de datos"*. Pulsando sobre la primera opción se puede ver el contenido de la base de datos activa sin alterar su contenido. La segunda opción permite cambiar el nombre de las variables, modificar datos o incluir nuevos registros. En caso de utilizar esta última opción será necesario guardar la base de datos en formato *R-Commander* para poder recuperarla posteriormente en otras sesiones de trabajo.

Si la base de datos procede de un archivo importado de SPSS se podrá visualizar, pero no editar con *R-Commander*. Para incluir nuevos registros o modificar información será necesario hacerlo en SPSS, o mediante algún Sistema Gestor de Bases de Datos externo, y volver a importar la base de datos modificada.

Obtener nuevas variables a partir de las existentes: calcular, recodificar y segmentar

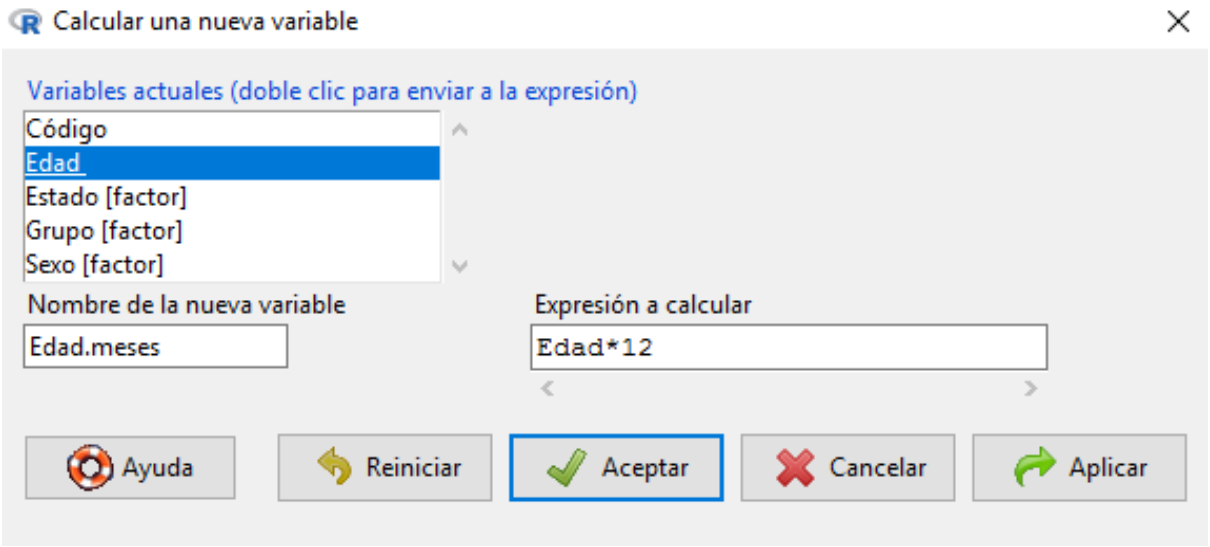
En ocasiones, el análisis de la información requiere modificar las unidades de medida de una variable cuantitativa, calcular índices mediante la combinación de diferentes mediciones o crear una nueva variable cualitativa que agrupe a los individuos en determinadas categorías. Estos y otros procedimientos pueden realizarse utilizando uno de los siguientes métodos:

Calcular una nueva variable

Permite generar nuevos valores a partir de la información de una o más variables. Así, a partir de la edad de los profesionales, expresada en años, podría calcularse una nueva variable denominada *Edad.meses* que contuviera la misma edad expresada en meses. Para ello, desde el menú principal se activará la secuencia:

Datos - Modificar variables del conjunto de datos activo – Calcular una nueva variable

El cuadro de diálogo abierto mostrará la siguiente apariencia, apareciendo en primer lugar el listado de variables originales.



Haciendo doble clic con el botón izquierdo del ratón sobre la variable *Edad*, ésta pasará al rectángulo blanco situado en la parte inferior derecha de la ventana, bajo el título "*Expresión a calcular*". La expresión para transformar la edad de años a meses es *Edad*12*, donde el asterisco equivale al signo de multiplicación.

En el rectángulo blanco situado a la izquierda de la ventana se escribirá el nombre de la nueva variable, *Edad.meses*. Este nombre puede contener cualquier combinación de letras mayúsculas, minúsculas, puntos (.) y guion bajo (_), pero no puede comenzar con un valor numérico ni contener cualquier otro símbolo diferente a los mencionados.

La nueva variable se añadirá automáticamente en la última columna de la base de datos tras pulsar el botón *Aceptar*.

La definición de la expresión a calcular puede usar todos los operadores aritméticos y funciones implementadas en el lenguaje *R*. Entre los más comunes están los operadores suma (+), resta (-), multiplicación (*), división (/) y elevación a una potencia (^), además de los recogidos en la tabla que aparece a continuación. De esta forma, si se hubiesen registrado las variables *Peso* (expresada en kilogramos) y *Altura* (expresada en metros), la expresión para calcular el Índice de Masa Corporal (*IMC*) a partir de ellas sería *Peso*/(*Altura*^2).

Operador	Símbolo	Expresión a calcular(*)
Suma	+	x+y
Resta	-	x–y
Multiplicación	*	x*y
División	/	x/y
Elevación a una potencia	^	x^y

Función	Nombre	Expresión a calcular(*)
Logaritmo		
neperiano	log	log(x)
en base 10	log10	log10(x)
Raíz cuadrada	sqrt	sqrt(x)

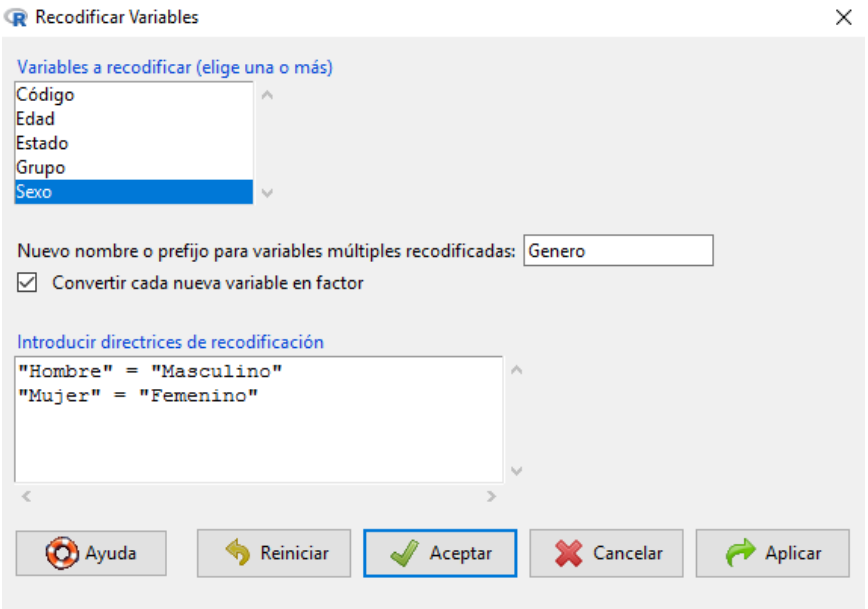
(*) *x e y pueden ser variables o valores numéricos*

a) Recodificar variables cualitativas y cuantitativas utilizando valores prefijados

Con este procedimiento es posible modificar los nombres de las categorías de una variable cualitativa o convertir una variable cuantitativa en cualitativa, agrupando a los individuos en las categorías que generen unos puntos de corte prefijados por el usuario. A modo de ejemplo, para sustituir las etiquetas "*Hombre*" y "*Mujer*" por "*Masculino*" y "*Femenino*" en la variable *Sexo* se realizará la siguiente secuencia del menú principal:

Datos - Modificar variables del conjunto de datos activo – Recodificar variables

En el listado de variables que muestra el cuadro de diálogo abierto se seleccionará la variable *Sexo*. Esta variable se marcará en azul tras hacer clic sobre ella con el botón izquierdo del ratón.



En el espacio situado a la derecha del título "*Nuevo nombre o prefijo para variables múltiples recodificadas*" se escribirá el nombre de la variable que contendrá las categorías del sexo con las nuevas etiquetas, en este caso *Género*. De esta forma se conservará en la base de datos la variable *Sexo* original, con categorías *Hombre* y *Mujer*, y se creará otra variable *Género* con categorías *Masculino* y *Femenino*. Si este rectángulo se deja vacío, los nombres de las categorías originales de *Sexo* se sustituirán por los nuevos y no se creará una variable adicional. Esta última opción es recomendable cuando no se desea seguir utilizando las etiquetas originales, evitando la duplicidad de variables en la base de datos.

Bajo el título "*Introducir directrices de recodificación*" se especificarán las transformaciones que *R-Commander* deberá realizar para modificar los valores de la variable. Estas transformaciones tienen siempre el mismo formato de escritura: A la izquierda el valor de la variable original, a la derecha el nuevo valor de la variable y ambos valores separados por el signo igual. Cuando los valores de la variable sean texto, como es el caso de los nombres de las categorías, éstos deberán ir entrecomillados. Así, las expresiones mostradas en la ventana anterior significan que la etiqueta "*Hombre*" será sustituida por "*Masculino*" y la etiqueta "*Mujer*" por "*Femenino*".

Habitualmente, la opción "*Convertir cada nueva variable en factor*" estará siempre activada, puesto que la nueva variable que se genera es cualitativa.

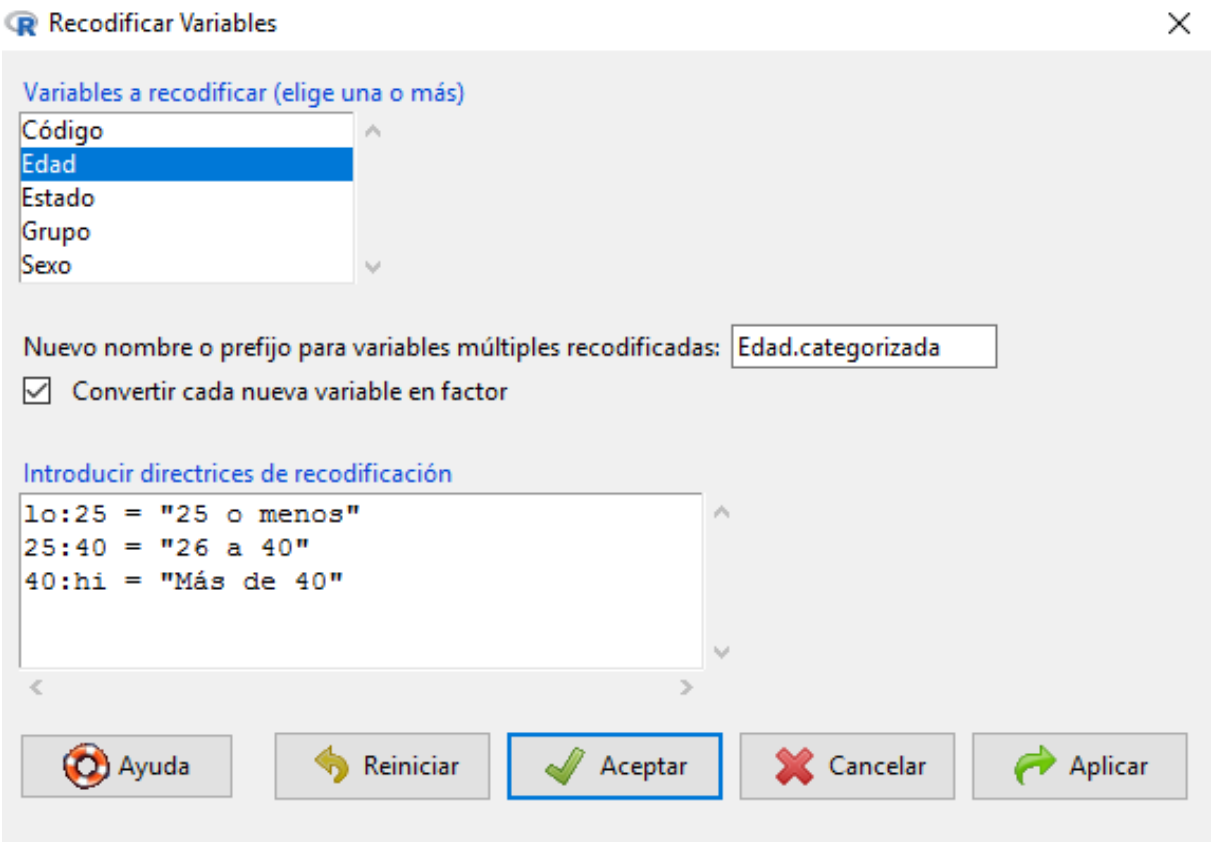
Tras pulsar el botón *Aceptar*, la nueva variable *Género* se añadirá en la última columna de la base de datos. En caso de no haber especificado un nuevo nombre de variable, los nombres de las categorías de la variable *Sexo* se habrán modificado automáticamente sin duplicar la variable.

Otra situación frecuente en la recodificación de variables es la transformación de una variable cuantitativa en otra cualitativa, creando dos o más categorías que agrupen a los individuos. A modo de ejemplo, a partir de la variable *Edad* se podrían crear tres intervalos que clasifiquen a los profesionales en las siguientes categorías: 25 años o menos, entre 26 y 40 años y Más de 40 años. De esta forma, dos puntos de corte preestablecidos en 25 y 40 años generarán tres intervalos o categorías diferentes.

El procedimiento a seguir es similar al caso anterior, partiendo de la siguiente secuencia del menú principal:

Datos - Modificar variables del conjunto de datos activo – Recodificar variables

Tras seleccionar la variable *Edad* en el cuadro de diálogo, se escribirá *Edad.categorizada* como nuevo nombre de la variable en el espacio correspondiente. Esta nueva variable será cualitativa con tres categorías, por tanto, la opción "*Convertir cada nueva variable en factor*" deberá estar activada.



Las directrices de recodificación son ahora un poco diferentes a las anteriores, como muestra la imagen superior. Puesto que la variable a recodificar es cuantitativa, la parte izquierda de la expresión ya no son valores individuales sino intervalos o rangos de valores. La parte situada a la derecha del signo igual es el nombre o etiqueta de cada categoría, que al ser texto deberá ir entrecomillada.

Las letras *lo* constituyen la abreviación de la palabra inglesa *lower* (el más bajo), mientras que *hi* es la abreviación de *higher* (el más alto). Así, la expresión *lo:25="25 o menos"* significa que los valores de la variable *Edad* comprendidos entre el valor más bajo de la base de datos y los 25 años formarán una categoría denominada "*25 o menos*". En este intervalo se incluye el valor 25.

La expresión *25:40="26 a 40"* indica que todos los profesionales con edad superior a 25 años e inferior o igual a 40 años formarán parte de la categoría "*26 a 40*". Aunque esté presente en la expresión, el valor 25 no se incluirá en este intervalo. *R-Commander* lo excluirá automáticamente al detectar que ya forma parte del primer intervalo. En este caso, una expresión equivalente para definir esta categoría sería *26:40="26 a 40"*, ya que la edad está recogida mediante números enteros. Sin embargo, si hubiera decimales, el valor 25.36 no quedaría recogido ni en el primer intervalo ni en el segundo. Para evitar errores de este tipo es aconsejable definir siempre el intervalo mediante la expresión *25:40="26 a 40"*.

Por último, *40:hi="Más de 40"* expresa que todos los profesionales con edad superior a 40 años constituirán la categoría "*Más de 40*". Como antes, *R-Commander* excluirá automáticamente el valor 40 de este intervalo al detectar que ya forma parte del intervalo anterior.

Pulsando *Aceptar*, la nueva variable *Edad.categorizada* se añadirá en la última columna de la base de datos. La siguiente imagen muestra el resultado obtenido tras accionar el botón "*Visualizar conjunto de datos*" bajo la barra de menú principal.

Accidentes

	Código	Grupo	Estado	Edad	Sexo	Edad.categorizada
1	4	Formación	No accidentado	45	Hombre	Más de 40
2	6	No formación	No accidentado	50	Hombre	Más de 40
3	14	No formación	No accidentado	55	Hombre	Más de 40
4	15	Formación	No accidentado	26	Mujer	26 a 40
5	18	Formación	No accidentado	58	Mujer	Más de 40
6	19	Formación	No accidentado	21	Mujer	25 o menos
7	22	Formación	No accidentado	52	Mujer	Más de 40
8	24	Formación	No accidentado	51	Mujer	Más de 40
9	1	Formación	Accidentado	22	Hombre	25 o menos
10	2	No formación	Accidentado	22	Hombre	25 o menos
11	3	No formación	Accidentado	22	Hombre	25 o menos
12	5	Formación	Accidentado	30	Hombre	26 a 40
13	7	Formación	Accidentado	34	Hombre	26 a 40
14	8	Formación	Accidentado	23	Hombre	25 o menos
15	9	No formación	Accidentado	28	<NA>	26 a 40
16	10	No formación	Accidentado	21	Hombre	25 o menos
17	11	No formación	Accidentado	40	Hombre	26 a 40
18	12	Formación	Accidentado	30	Hombre	26 a 40
19	13	No formación	Accidentado	35	Hombre	26 a 40
20	16	No formación	Accidentado	NA	Mujer	<NA>
21	17	No formación	Accidentado	50	Mujer	Más de 40
22	20	No formación	Accidentado	25	Mujer	25 o menos
23	21	Formación	Accidentado	47	Mujer	Más de 40
24	23	No formación	Accidentado	23	Mujer	25 o menos
25	25	No formación	Accidentado	23	Mujer	25 o menos

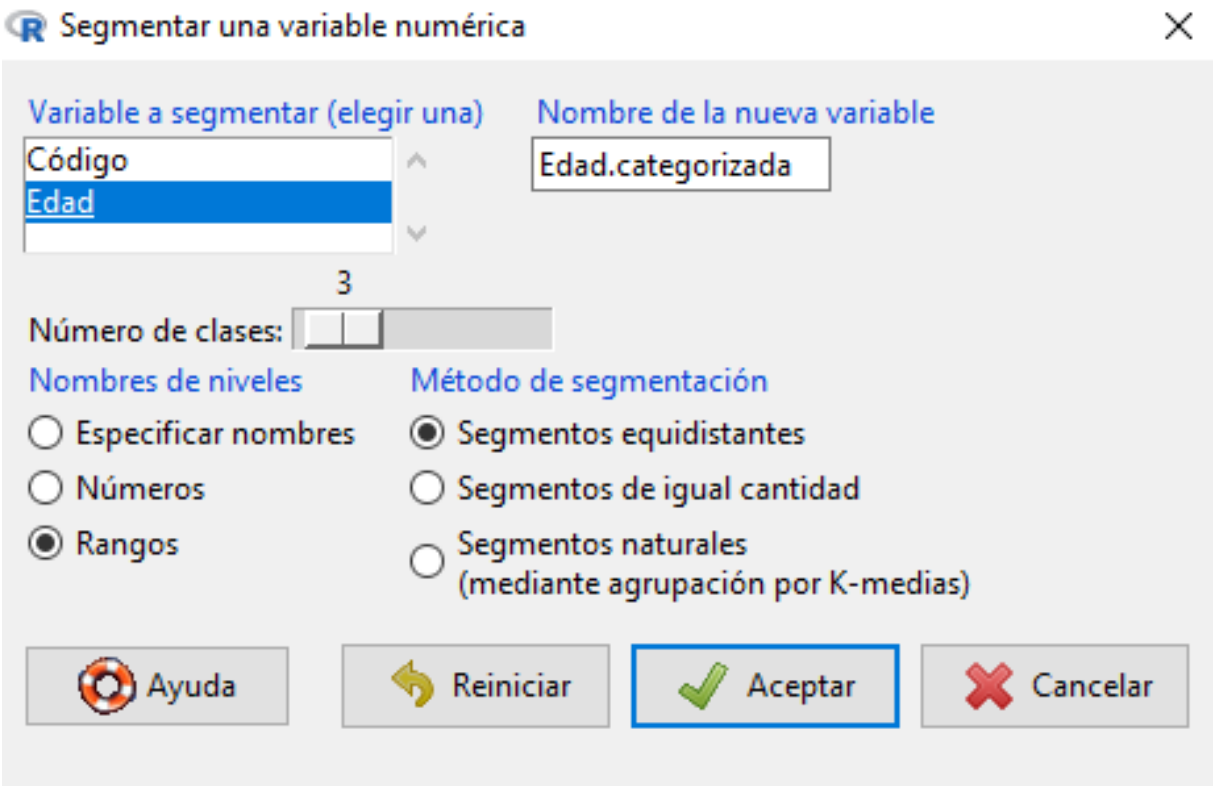
En caso de haber dejado en blanco el espacio "*Nuevo nombre o prefijo para variables múltiples recodificadas*", los valores originales de la variable *Edad* serían sustituidos por las nuevas categorías. Esta opción no es muy recomendable, ya que impedirá trabajar con la variable original en posteriores sesiones de trabajo.

b) Segmentar variables cuantitativas mediante puntos de corte automáticos

La recodificación de variables vista en el apartado anterior permite transformar una variable cuantitativa en otra cualitativa, agrupando a los individuos en categorías que generen unos puntos de corte prefijados por el usuario. Cuando no se dispone de información sobre los puntos de corte más adecuados, *R-Commander* permite realizar un procedimiento de segmentación utilizando puntos de corte automáticos, no preestablecidos previamente. Para ello, desde el menú principal se activará la secuencia:

Datos - Modificar variables del conjunto de datos activo – Segmentar variable numérica

El cuadro de diálogo abierto mostrará sólo las variables cuantitativas de la base de datos, ya que este procedimiento únicamente es válido para valores numéricos.



Para categorizar o segmentar la variable *Edad* en tres grupos, utilizando puntos de corte automáticos, se seleccionará del listado de variables haciendo clic sobre ella con el botón izquierdo del ratón. *R-Commander* la marcará en azul, pudiendo escribir a continuación el nombre de la nueva variable en la parte superior derecha de la ventana.

El botón "*Número de clases*" permite definir el número de categorías de la nueva variable *Edad.categorizada*, en este caso tres.

Como método de segmentación, *R-Commander* ofrece los siguientes:

Segmentos equidistantes: Permite realizar una partición de la variable en intervalos de igual longitud. Es el método más sencillo para categorizar una variable cuantitativa, de manera que el segmento a dividir estará dado por la diferencia entre el valor mayor y el valor menor de la variable. En este caso, la edad menor es 21 años y la mayor 58, por lo que la amplitud de cada uno de los tres intervalos será (58 - 21) / 3 = 12.33 años. Así, el primer grupo de profesionales estará formado por aquellos con edades comprendidas entre 21 y 33.3 años, el segundo grupo entre 33.3 y 45.7 años y el tercer grupo entre 45.7 y 58.

Segmentos de igual cantidad: Realiza una partición de la variable de forma que en cada intervalo haya el mismo número de sujetos.

Segmentos naturales (mediante agrupación por K-medias): Es un algoritmo más complejo que divide a la variable en los intervalos especificados bajo la condición de que los sujetos de cada grupo tengan valores parecidos, minimizando la distancia entre cada uno de ellos y el valor medio del intervalo al que pertenecen.

Por último, las opciones del bloque "*Nombres de niveles*" permitirán poner nombre a cada uno de los intervalos o categorías generadas. La opción "*Especificar nombres*" se utilizará para escribir el texto que el usuario desee, mientras que las opciones "*Números*" y "*Rangos*" harán que *R-Commander* asigne automáticamente las etiquetas de las categorías, utilizando respectivamente números consecutivos o el mismo rango de valores del intervalo. Tras pulsar el botón *Aceptar* y visualizar la base de datos, el método de segmentos equidistantes y el etiquetado mediante rangos mostrará el siguiente resultado:

Accidentes

	Código	Grupo	Estado	Edad	Sexo	Edad.categorizada
1	4	Formación	No accidentado	45	Hombre	(33.3,45.7]
2	6	No formación	No accidentado	50	Hombre	(45.7,58]
3	14	No formación	No accidentado	55	Hombre	(45.7,58]
4	15	Formación	No accidentado	26	Mujer	(21,33.3]
5	18	Formación	No accidentado	58	Mujer	(45.7,58]
6	19	Formación	No accidentado	21	Mujer	(21,33.3]
7	22	Formación	No accidentado	52	Mujer	(45.7,58]
8	24	Formación	No accidentado	51	Mujer	(45.7,58]
9	1	Formación	Accidentado	22	Hombre	(21,33.3]
10	2	No formación	Accidentado	22	Hombre	(21,33.3]
11	3	No formación	Accidentado	22	Hombre	(21,33.3]
12	5	Formación	Accidentado	30	Hombre	(21,33.3]
13	7	Formación	Accidentado	34	Hombre	(33.3,45.7]
14	8	Formación	Accidentado	23	Hombre	(21,33.3]
15	9	No formación	Accidentado	28	<NA>	(21,33.3]
16	10	No formación	Accidentado	21	Hombre	(21,33.3]
17	11	No formación	Accidentado	40	Hombre	(33.3,45.7]
18	12	Formación	Accidentado	30	Hombre	(21,33.3]
19	13	No formación	Accidentado	35	Hombre	(33.3,45.7]
20	16	No formación	Accidentado	NA	Mujer	<NA>
21	17	No formación	Accidentado	50	Mujer	(45.7,58]
22	20	No formación	Accidentado	25	Mujer	(21,33.3]
23	21	Formación	Accidentado	47	Mujer	(45.7,58]
24	23	No formación	Accidentado	23	Mujer	(21,33.3]
25	25	No formación	Accidentado	23	Mujer	(21,33.3]

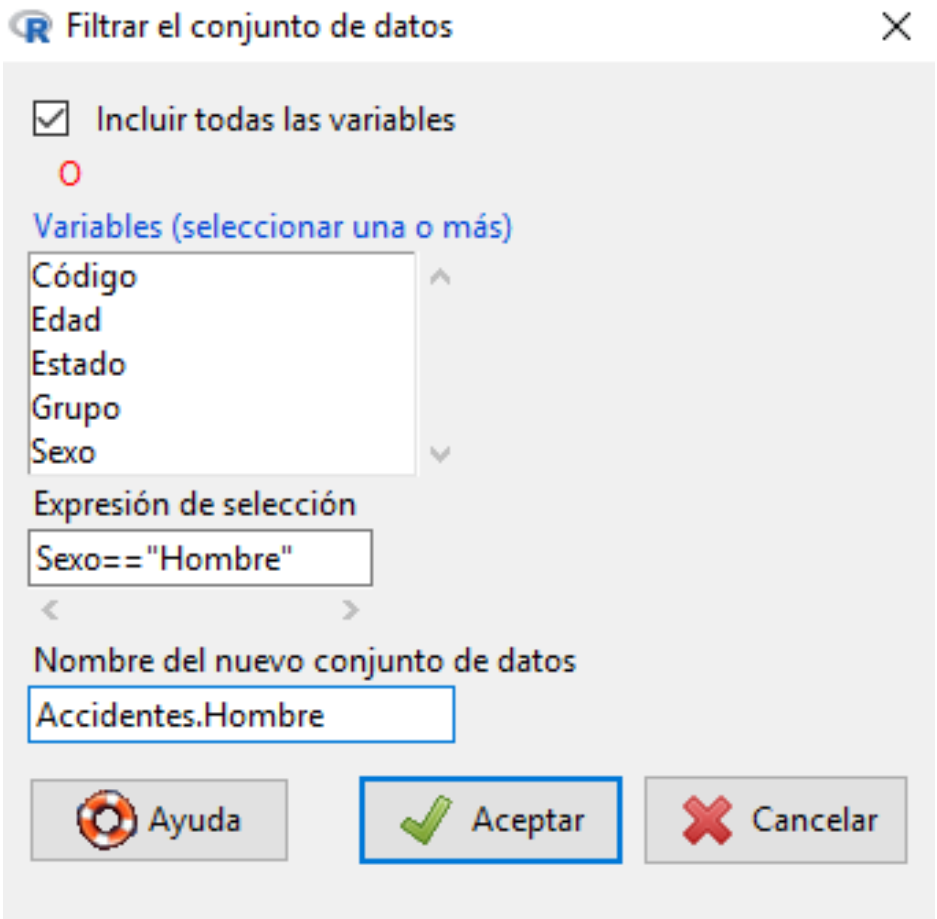
Seleccionar registros y variables

En ocasiones es necesario filtrar la base de datos para elegir determinadas variables o seleccionar sólo aquellos casos o registros que verifiquen una determinada condición. Este procedimiento puede hacerse desde el menú principal de *R-Commander* siguiendo la secuencia:

Datos – Conjunto de datos activo – Filtrar el conjunto de datos activo

Las opciones mostradas en el cuadro de diálogo permitirán generar una nueva base de datos que contenga sólo las variables y los registros especificados. Por defecto, la opción "*Incluir todas las variables*" se encuentra activada. Sin embargo, es posible desactivarla y elegir sólo aquellas variables que se deseen trasladar a la nueva base de datos. Para ello bastará con marcarlas usando el botón izquierdo del ratón a la vez que se pulsa la tecla *Control* (*Ctrl*) del teclado.

En el cuadro "*Expresión de selección*" se insertará la condición que deben cumplir los registros de la base de datos para ser seleccionados. Así, para filtrar por la variable *Sexo* eligiendo sólo a los hombres se escribirá *Sexo=="Hombre"*.



Obsérvese que la condición de igualdad se expresa con el doble signo == y no con = como suele ser habitual. Además, puesto que *Sexo* es una variable cualitativa, la categoría especificada como filtro ha de ir entre comillas. Es muy importante que las expresiones de selección respeten el lenguaje *R-Commander* para que el filtro se realice correctamente. Estas son las expresiones y operadores lógicos más frecuentes:

Expresiones lógicas	Descripción
<	Menor que
<=	Menor o igual que
>	Mayor que
>=	Mayor o igual que
==	Igual a
! =	Distinto a

Operadores lógicos	Descripción
&	Y
|	O
!	No

De esta forma, para seleccionar a los sujetos con edades comprendidas entre 25 y 40 años, ambas inclusive, la expresión de selección será *25<=Edad & Edad<=40*.

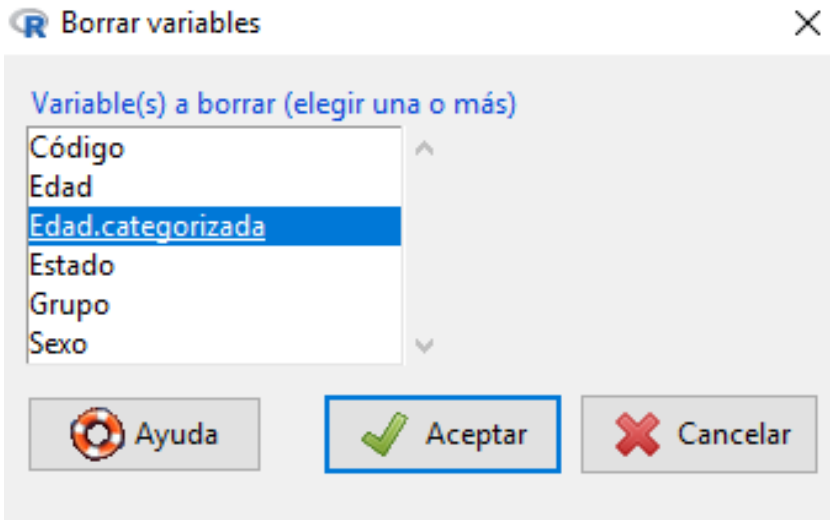
Por último, la opción "*Nombre del nuevo conjunto de datos*" permitirá almacenar el filtro en una nueva base de datos, que pasará a ser la base de datos activa.

Eliminar variables y registros

R-Commander permite borrar variables o registros de la base de datos. Estos procedimientos pueden realizarse desde el menú principal, de manera que para eliminar una variable de la base de datos se pulsará la secuencia:

Datos - Modificar variables del conjunto de datos activo – Eliminar variables del conjunto de datos

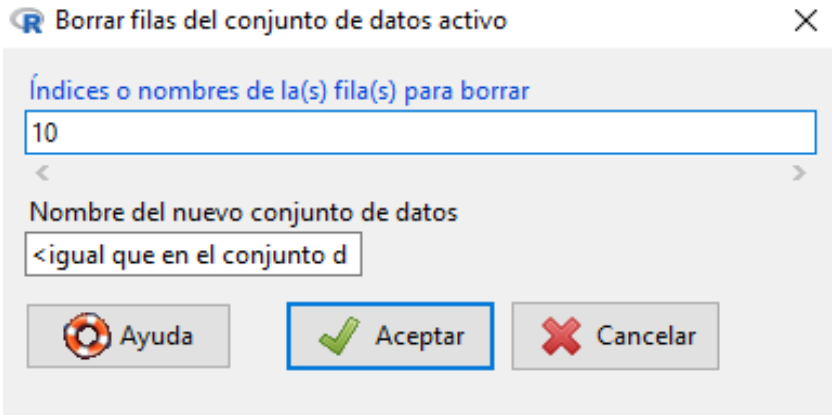
Para eliminar una única variable se pulsará sobre ella en el cuadro de diálogo abierto y a continuación el botón *Aceptar*. Es posible borrar varias variables a la vez dejando pulsada la tecla *Control* (*Ctrl*) del teclado mientras se seleccionan del listado todas las variables que se desean eliminar.



Para eliminar uno o varios registros de la base de datos se pulsará la secuencia:

Datos – Conjunto de datos activo – Borrar fila(s) del conjunto de datos activo

En el cuadro de diálogo abierto se especificará el número de la fila que se desea eliminar, justo debajo del título "*Índices o nombres de la(s) fila(s) para borrar*".



Para borrar sólo la décima fila se escribirá el valor 10, mientras que para borrar todas las filas comprendidas entre la 10 y la 22 se escribirá 10:22.

La nueva base de datos puede guardarse con otro nombre cumplimentando el espacio "*Nombre del nuevo conjunto de datos*". De esta forma se mantendrá intacta la base de datos original en la memoria de *R-Commander* y se creará una copia de ella en la que se eliminarán los registros. Si no se especifica ningún nombre, los registros serán borrados directamente en la base de datos activa en memoria.

Guardar la base de datos activa en un archivo *R-Commander*

Cuando se elabora o importa una base de datos, se crean nuevas variables o se modifican datos, *R-Commander* guarda la nueva información en memoria, pero no la almacena físicamente en el disco duro del ordenador. Esto supone que al cerrar una sesión de trabajo y salir del programa se perderá toda la información junto con los cambios realizados, siendo necesario volver a introducir, importar o modificar los datos en la siguiente sesión.

Para evitar este problema, es aconsejable guardar la base de datos en un archivo *R-Commander*, de manera que sea posible recuperar la información en sesiones posteriores. Para ello se seguirá la siguiente secuencia desde el menú principal:

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

En la ventana abierta a continuación se deberá seleccionar la carpeta en la que se quiere guardar la base de datos, especificar el nombre del archivo y pulsar el botón *Guardar*. El archivo tendrá extensión *.rda*, un tipo de formato que sólo podrá leerse posteriormente con *R-Commander*.

Abrir una base de datos en formato *R-Commander*

Al comenzar una sesión de trabajo será necesario cargar en memoria la base de datos que se desea analizar. Si ésta ya fue elaborada o importada en una sesión de trabajo anterior y se guardó posteriormente como archivo *R-Commander*, será posible recuperarla pulsando la siguiente secuencia del menú principal:

Datos – Cargar conjunto de datos

Cuando se abra la ventana de selección de archivos bastará con buscar la base de datos en la carpeta correspondiente y hacer doble clic sobre ella. Para facilitar la búsqueda es conveniente seleccionar la opción "*Archivos de datos de R (*.rda, *.Rda, *.RDA)"*, situada en la esquina inferior derecha de la ventana, en lugar de la opción *Todos los archivos (*.*)* que aparece por defecto. De esta forma se mostrarán sólo las bases de datos previamente grabadas en formato *R-Commander*.

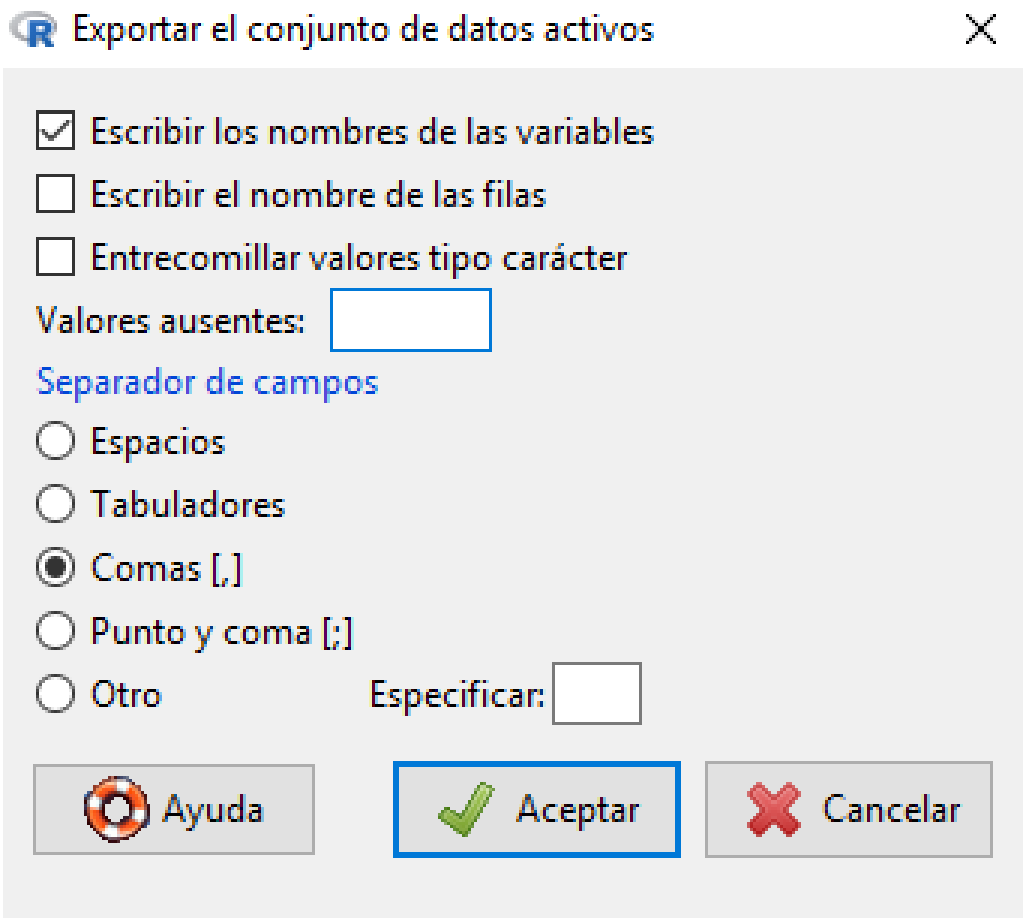
Una vez capturada la base de datos, *R-Commander* mostrará su nombre en color azul junto al texto "*Conjunto de datos"*, debajo del menú principal. Pulsando la opción "*Visualizar conjunto de datos"*, situada a la derecha del menú, se puede comprobar si la captura de la base de datos se ha realizado correctamente.

Exportar la base de datos activa a un archivo con formato texto

Las bases de datos guardadas en archivos *R-Commander* (con extensión *.rda*) sólo pueden abrirse con este software. Para compartir la información con otros usuarios que no utilicen *R-Commander* o trabajar con otros programas estadísticos será necesario exportar la información a un archivo de texto, formato universal que puede leer cualquier software. Este proceso puede realizarse desde el menú principal, pulsando la secuencia:

Datos – Conjunto de datos activo – Exportar el conjunto de datos activo

A continuación, se abrirá una ventana con diferentes opciones que podrán ser activadas o desactivadas dependiendo del formato con el que se desee exportar la información.



En general, es aconsejable seguir estas indicaciones:

Escribir los nombres de las variables: Activado. Esta opción registrará el nombre de las variables en el archivo de texto, facilitando la comprensión de la información almacenada.

Escribir el nombre de las filas: Desactivado. Si se deja activada, esta opción añadirá una variable adicional con números correlativos. En principio no suele ser útil y aumenta innecesariamente el volumen de la base de datos, por lo que es preferible desactivarla.

Entrecomillar valores tipo carácter: Desactivado. De esta forma se facilitará la importación de datos desde otro software que no utilice las comillas para identificar valores de tipo carácter.

Valores ausentes: Dejar en blanco, sin escribir el valor por defecto NA.

Separador de campos: Usar comas. Si hay variables cualitativas en las que el nombre de alguna categoría tenga espacios, no es aconsejable usar a su vez el espacio como separador de campos.

Tras pulsar el botón *Aceptar* se deberá seleccionar la carpeta en la que se quiere guardar la base de datos, especificar el nombre del archivo y pulsar el botón *Guardar*. El archivo deberá tener extensión *.txt*, *.TXT*, *.dat*, *.DAT*, *.csv* o *.CSV*, formato de texto universal compatible con cualquier software.

4

ANÁLISIS DESCRIPTIVO UNIVARIANTE

El objetivo del Análisis Descriptivo es resumir la información recogida en la base de datos, describir las características del grupo estudiado y detectar posibles anomalías que hayan podido producirse durante el registro de la información. Suele ser el primer análisis estadístico de cualquier investigación, ya que sus resultados son esenciales para conocer los datos y planificar análisis más complejos. A veces, un análisis descriptivo adecuado cubre, por sí solo, el objetivo principal de algunos estudios.

Los contenidos de este capítulo exponen las técnicas más frecuentes para la descripción de cada una de las variables que componen una base de datos, usando para ello el caso práctico *Accidentes por pinchazo en profesionales de enfermería*.

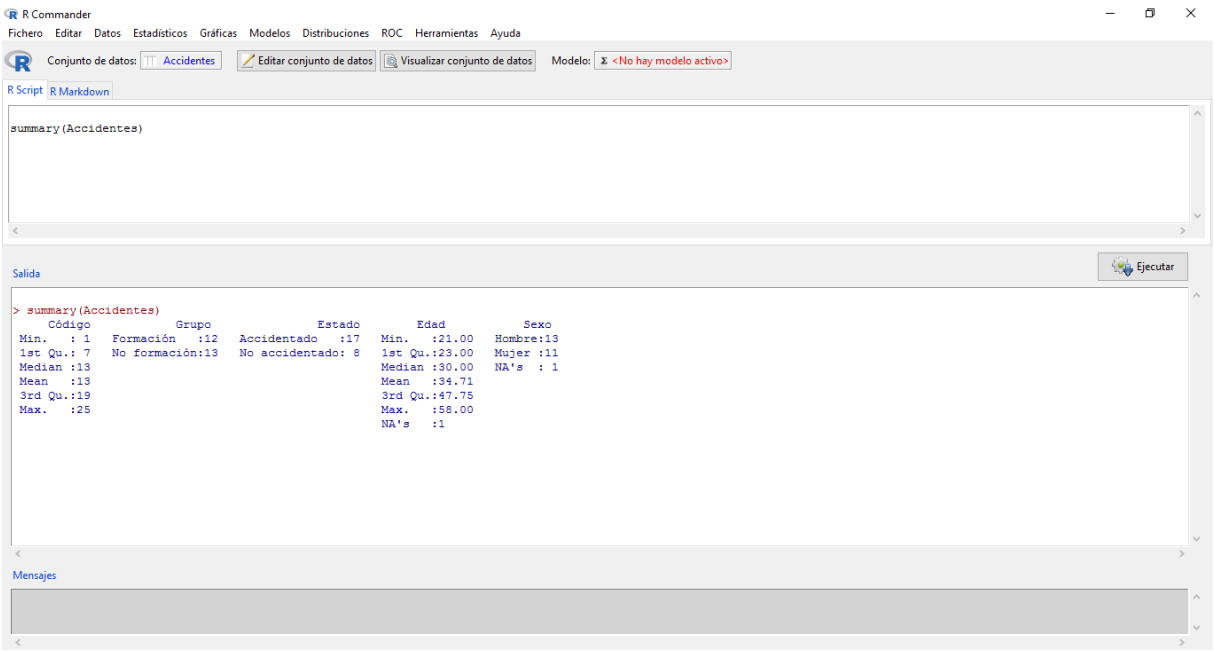
DESCRIPCIÓN INICIAL DE VARIABLES

Las técnicas que se utilizan para describir variables cualitativas son diferentes a las utilizadas para la descripción de variables cuantitativas. Usualmente, una variable cualitativa se describe a través de una tabla de frecuencias, mostrando el número de sujetos que constituye cada categoría y su porcentaje con respecto al total de sujetos. Por contra, la descripción de una variable cuantitativa se realiza mediante un resumen numérico, que habitualmente incorpora los siguientes valores: mínimo, máximo, media y desviación típica.

Parte de esta información se puede obtener con *R-Commander* en un análisis exploratorio inicial, realizando la siguiente secuencia desde el menú principal:

Estadísticos – Resúmenes – Conjunto de datos activo

La ventana de resultados mostrará el nombre de cada variable de la base de datos y debajo del él un recuento del número de sujetos por categoría, si la variable es cualitativa, o un resumen numérico, dado por los valores mínimo, primer cuartil, mediana, media, tercer cuartil y máximo, si la variable es cuantitativa. La última información de cada variable es el número de sujetos con valores perdidos, identificados con la etiqueta NA’s.



Los resultados obtenidos muestran una distribución de sujetos aproximadamente equilibrada entre las categorías de cada variable cualitativa, exceptuando el estado al final del seguimiento, donde la mayoría de profesionales se encuentra en el grupo *Accidentado*.

La edad oscila entre 21 y 58 años, con una edad media de 34.7 años. La media es un representante del grupo, un valor central en torno al que se sitúa la edad de los profesionales. El valor mínimo y máximo, además de describir el rango de la variable, son indicadores de posibles valores extremos, introducidos a veces por errores durante el registro de la información.

En general, los cuartiles no suelen incorporarse a un resumen descriptivo básico, aunque serán útiles para comprender el significado y la utilidad de un gráfico de caja, descrito más adelante. Estos parámetros dividen a la variable en cuatro partes iguales, de manera que, una vez ordenados los datos de menor a mayor, cada intervalo contiene al 25% de los valores registrados. En este caso, el primer cuartil se sitúa en 23 años, indicando que el 25% de los profesionales tienen una edad inferior a 23 años. La mediana o segundo cuartil, localizada en 30 años, indica que el 50% de los sujetos tiene menos de 30 años. Por último, el tercer cuartil, situado en 47.75 años, señala que el 75% de los profesionales tiene una edad inferior a 47.75 años.

DESCRIPCIÓN DE VARIABLES CUALITATIVAS

Tabla de frecuencias

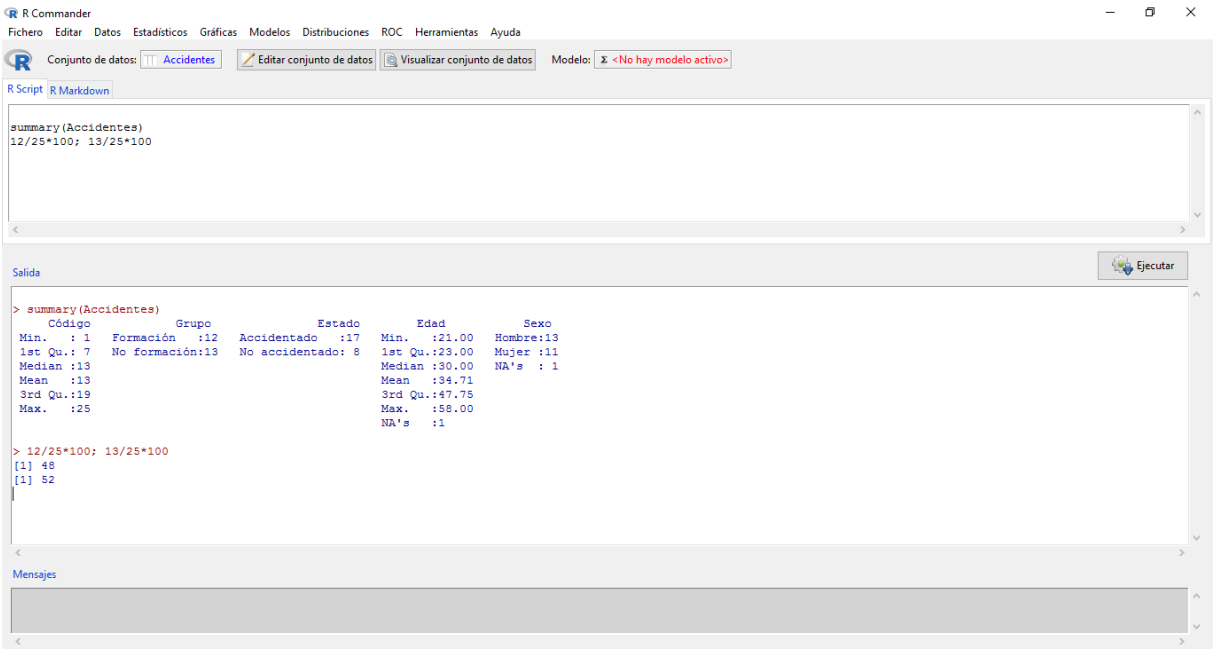
La descripción de variables realizada en el apartado anterior muestra el número de sujetos que componen las categorías de cada variable cualitativa. Esta información deberá transcribirse a una tabla de tres columnas, confeccionada con un procesador de textos, para obtener la siguiente tabla de frecuencias:

Variable	Número de sujetos	Porcentaje de sujetos
Grupo		
Formación	12	48%
No formación	13	52%
Estado		
Accidentado	17	68%
No accidentado	8	32%
Sexo		
Hombre	13	52%
Mujer	11	44%
Valores perdidos	1	4%

Puesto que la salida anterior de *R-Commander* no ofrece el porcentaje de sujetos correspondiente a cada categoría, será necesario obtenerlos a través de una calculadora, una hoja de cálculo o utilizando la ventana de instrucciones de la propia interfaz. Así, el porcentaje de profesionales que recibieron formación será (12/25) x 100=48, mientras que el porcentaje de personas que no la recibieron es (13/25) x 100=52. Ambos porcentajes pueden calcularse escribiendo en la ventana de instrucciones de *R-Commander* la siguiente línea:

```
12/25*100; 13/25*100
```

Dejando el cursor colocado en la misma línea, justo después del último 100, se pulsará el botón *Ejecutar*, situado en la parte inferior derecha de la ventana de instrucciones. La ventana de resultados mostrará entonces los porcentajes correspondientes, que habrán de transcribirse a la tabla de frecuencias del procesador de textos.



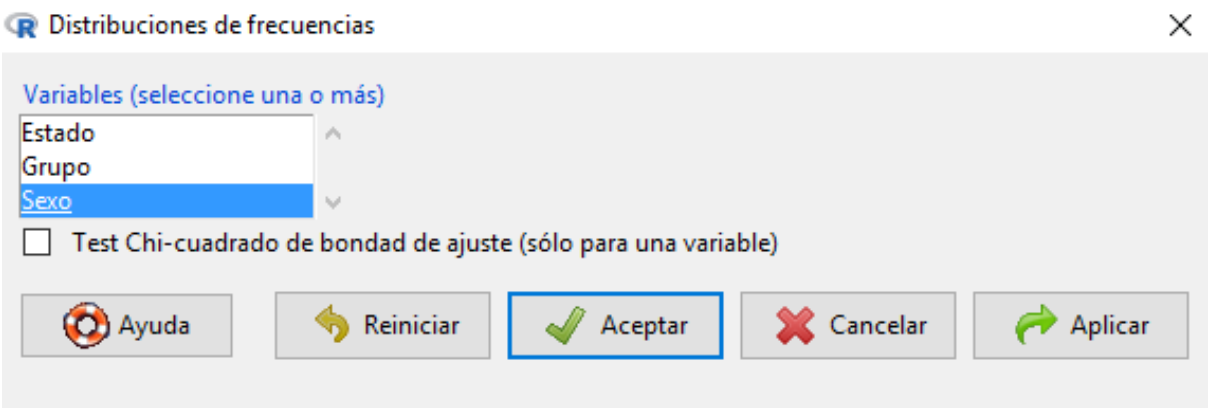
El mismo procedimiento se seguirá para obtener el resto de porcentajes, escribiendo sucesivamente las siguientes líneas y pulsando el botón *Ejecutar* al final de cada una de ellas:

```
17/25*100; 8/25*100
13/25*100; 11/25*100; 1/25*100
```

Un resultado similar puede obtenerse pulsando la siguiente secuencia desde el menú principal de *R-Commander*:

Estadístico – Resúmenes – Distribución de frecuencias

A continuación, aparecerá un cuadro de diálogo en el que podrá seleccionarse la variable cualitativa que se desea describir, en este caso *Sexo*.



Pulsando sobre el botón *Aceptar* se mostrará la siguiente información en la ventana de resultados:

```
> .Table <- table(Accidentes$Sexo)
> .Table  # counts for Sexo

Hombre  Mujer
    13     11

> round(100*.Table/sum(.Table), 2)  # percentages for Sexo

Hombre  Mujer
 54.17  45.83

> remove(.Table)
```

En rojo aparecerán las instrucciones que *R-Commander* utiliza para contar el número de sujetos de cada categoría y calcular el porcentaje correspondiente. En azul los resultados, que tendrán que transcribirse a la tabla de frecuencias del procesador de textos.

Este procedimiento ofrece automáticamente las frecuencias absolutas (sujetos) y relativas (porcentajes). Sin embargo, elimina del cálculo los valores faltantes. Así, el total de sujetos no es 25 sino 24 por haber un valor perdido en la variable *Sexo*. No es posible, por tanto, obtener automáticamente el porcentaje de valores perdidos para esta variable, información a veces muy valiosa para describir el grado de cumplimentación de la variable y la calidad de la fuente de información.

Los resultados del primer procedimiento, en el que los porcentajes se calcularon escribiendo líneas de texto, coincidirán con los resultados de este procedimiento automático cuando la variable no tenga valores perdidos, en cuyo caso éste último puede ser preferible por la rapidez en la obtención de información.

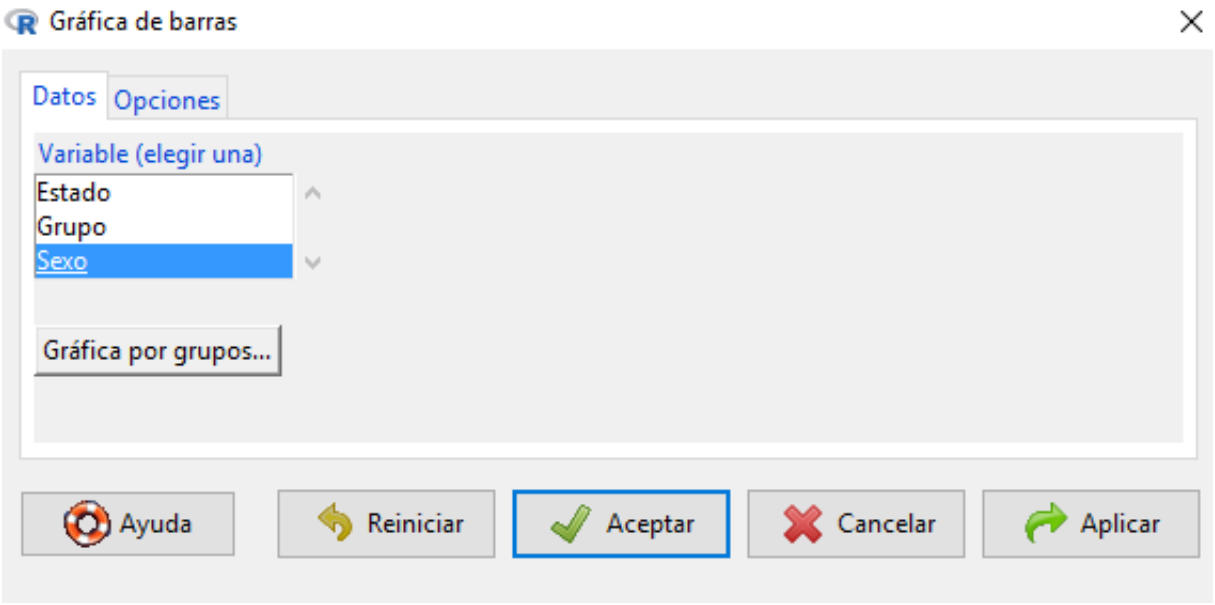
Gráfico de barras

R-Commander permite representar en un gráfico de barras la misma información de una tabla de frecuencias. Para ello, desde el menú principal se seguirá la secuencia:

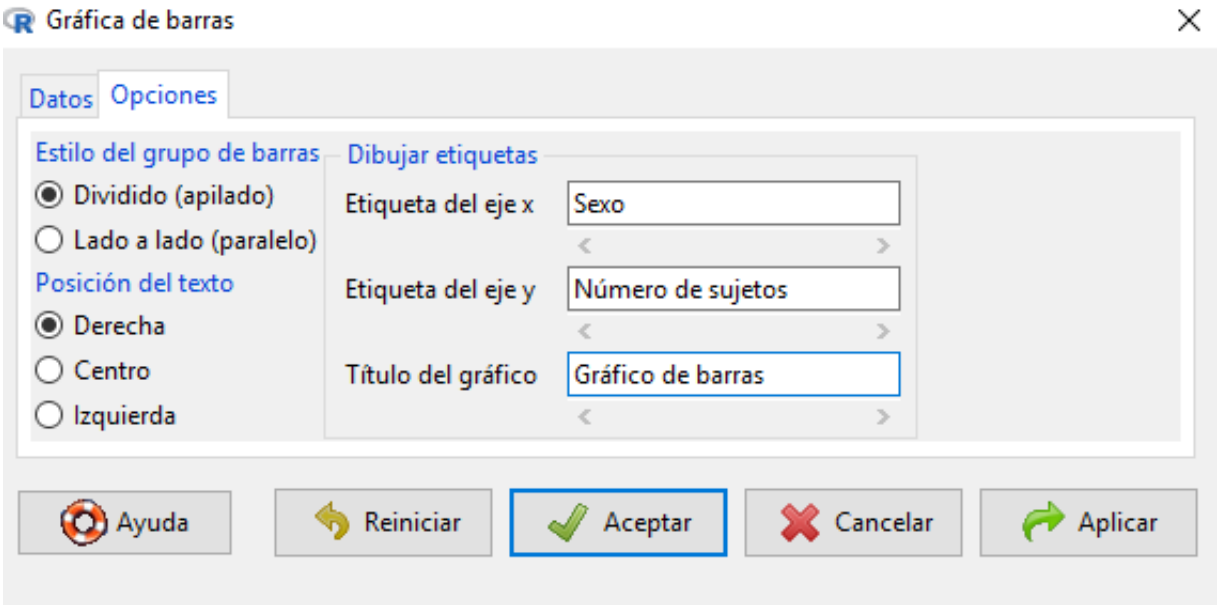
Gráficas – Gráfica de barras

A continuación, aparecerá un cuadro de diálogo que permitirá seleccionar la variable a representar. En esta ventana sólo se mostrarán las variables cualitativas, definidas como factor

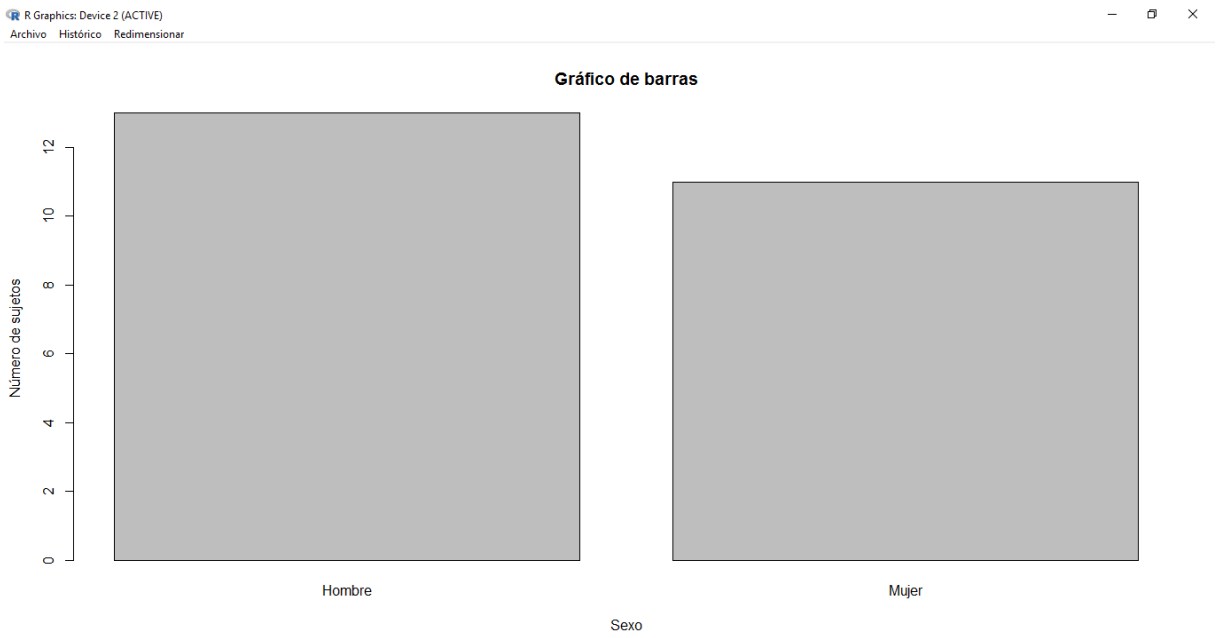
o de tipo carácter en *R-Commander*, ya que este gráfico no tiene sentido para variables cuantitativas.



Tras marcar en azul la variable correspondiente, en este caso *Sexo*, la pestaña *Opciones* permite incluir el título de los ejes del gráfico y algunos parámetros opcionales más.



Al pulsar el botón *Aceptar*, aparecerá una nueva ventana que contendrá el gráfico.



En el diagrama de barras, el eje horizontal muestra todas las categorías de la variable cualitativa y la altura de la barra representa el número de sujetos que componen cada una de ellas.

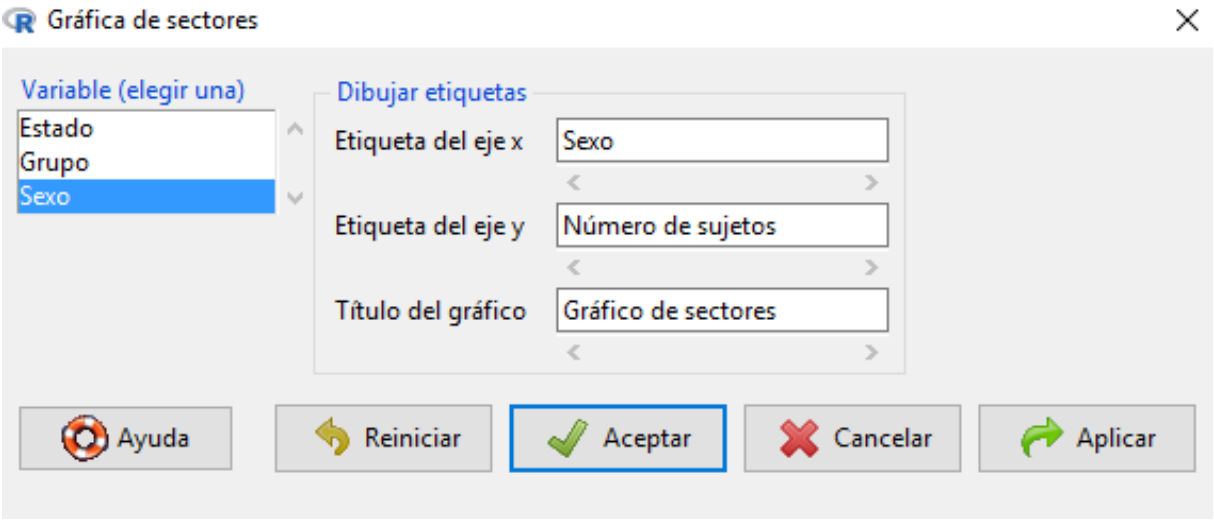
Algunas cuestiones relacionadas con la edición de gráficos no pueden realizarse en *R-Commander* sin recurrir a comandos, por lo que su uso requerirá conocer más detalles sobre las instrucciones y parámetros del software *R*. Una alternativa es usar la información de la tabla de frecuencias para crear el gráfico directamente en el procesador de textos, hoja de cálculo o programa de presentación que se esté utilizando para elaborar el informe.

Diagrama de sectores

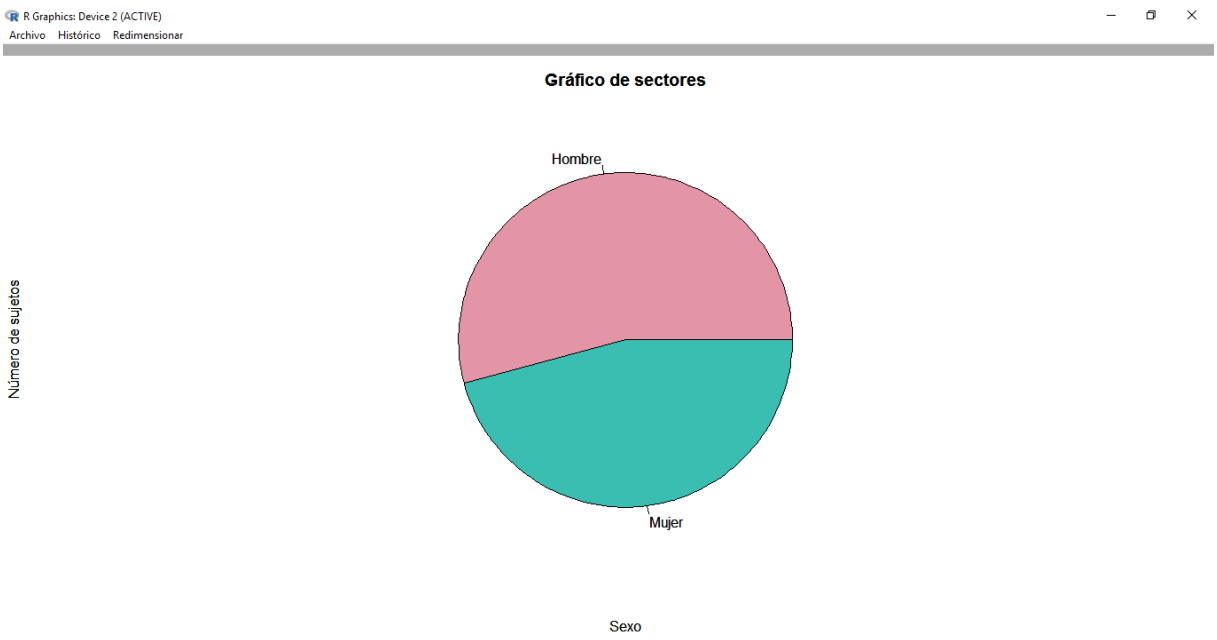
Una alternativa equivalente al diagrama de barras es el diagrama de sectores. Esta representación gráfica divide un círculo en tantas secciones como categorías tenga la variable cualitativa, siendo el tamaño de cada una de ellas proporcional al número de sujetos que contiene. Para realizarlo, desde el menú principal de *R-Commander* se pulsará:

Gráficas – Gráfica de sectores

El cuadro de diálogo que aparecerá a continuación es similar al descrito para el diagrama de barras. En él se seleccionará la variable a representar marcándola en azul.



Tras pulsar el botón *Aceptar* la ventana gráfica mostrará el diagrama de sectores, sustituyendo el gráfico que hubiese anteriormente.



La última línea de la ventana de instrucciones de *R-Commander* contiene el comando interno utilizado para realizar el gráfico. En este caso, el comando es:

with(Accidentes, pie(table(Sexo), labels=levels(Sexo), xlab="Sexo", ylab="Número de sujetos", main="Gráfico de sectores", col=rainbow_hcl(length(levels(Sexo)))))

que podrá modificarse de la siguiente forma para que aparezca, a modo de ejemplo, el color rojo y azul claro para los sectores:

with(Accidentes, pie(table(Sexo), labels=levels(Sexo), xlab="Sexo", ylab="Número de sujetos", main="Gráfico de sectores", col=c("red", "lightblue")))

Si el comando ocupara dos líneas de texto dentro de la ventana de instrucciones, para ejecutarlo será necesario seleccionar ambas líneas con el ratón y posteriormente pulsar el botón *Ejecutar*. Las líneas seleccionadas quedarán marcadas en azul, como muestra la imagen anterior.

Como ocurría con el gráfico de barras, una alternativa a la sintaxis de *R-Commander* es usar la información de la tabla de frecuencias para crear el diagrama de sectores directamente en el procesador de textos, hoja de cálculo o programa de presentación que se esté utilizando para elaborar el documento.

DESCRIPCIÓN DE VARIABLES CUANTITATIVAS

Resúmenes numéricos

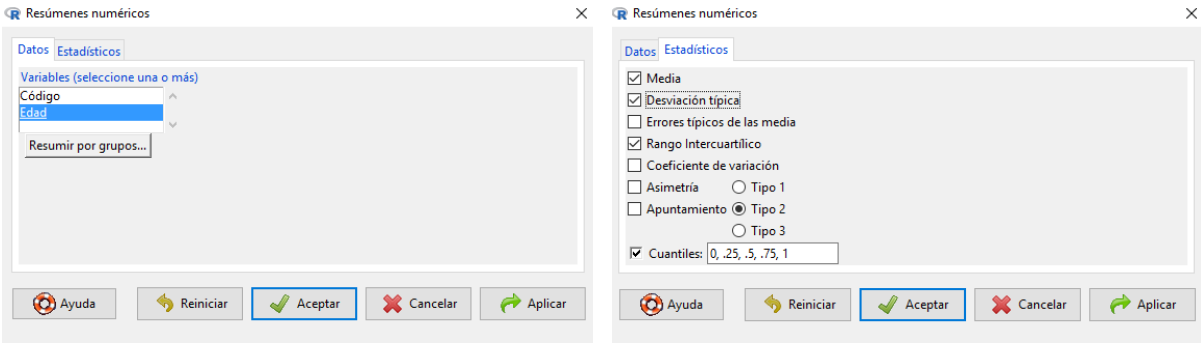
La ventana de resultados obtenida en el apartado *"Descripción inicial de variables"* muestra algunos parámetros necesarios para describir la variable *Edad*, única variable cuantitativa de la base de datos. Esta información deberá transcribirse a una tabla de seis columnas, confeccionada con un procesador de textos, para conseguir una parte de la siguiente tabla:

Variable	Sujetos	Mínimo	Máximo	Media	Desviación típica
Edad	24	21	58	34.71	12.85

Puesto que la salida anterior de *R-Commander* no ofrece la desviación típica, será necesario obtenerla desde el menú principal activando la secuencia:

Estadísticos – Resúmenes – Resúmenes numéricos

En el cuadro de diálogo que aparece a continuación se seleccionará la variable de interés, en este caso *Edad*, y en el apartado *Estadísticos* se activará la opción *Desviación típica*.



El resto de opciones puede quedar activado o desactivado, dependiendo de las necesidades del análisis y teniendo en cuenta que parte de la información ofrecida se obtuvo en la descripción inicial de variables.

Tras pulsar el botón *Aceptar*, la desviación típica de la edad aparecerá en la ventana de resultados de *R-Commander*, pudiendo transcribirla a la tabla del procesador de textos para completar la información. En este caso su valor es 12.85 años, ofreciendo una medida de la dispersión de los valores individuales con respecto a la media del grupo. En general, la desviación típica estima la separación entre los valores individuales y la media del grupo. Cuanto más pequeña sea, más parecidos serán los sujetos entre sí, de manera que una desviación típica igual a cero indicaría que todos los individuos tienen la misma edad.

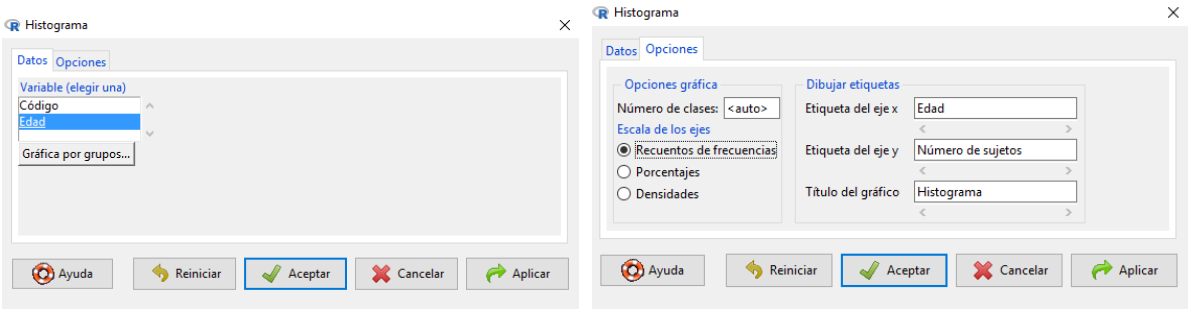
Si en la base de datos hubiese más variables cuantitativas, su información se incorporaría en filas adicionales de la tabla descriptiva anterior, siguiendo el mismo procedimiento que el mostrado para la variable edad.

Histograma

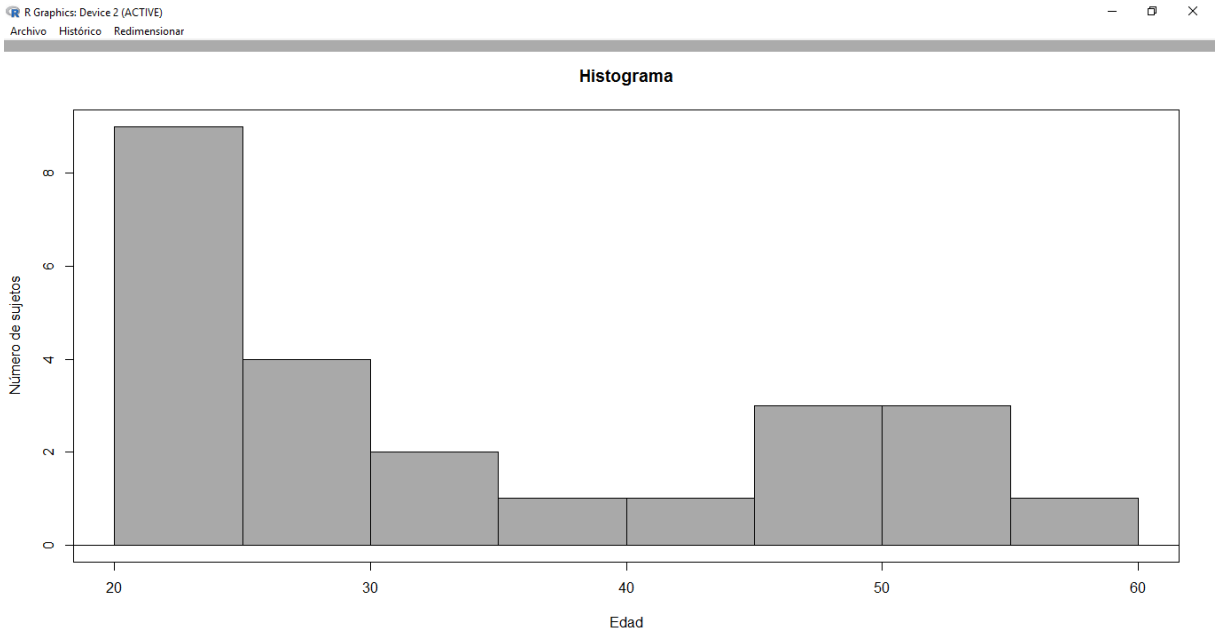
Uno de los gráficos más utilizados para representar variables cuantitativas es el histograma. Su representación más habitual se realiza mediante dos ejes: Uno horizontal en el que se representan las categorías de la variable segmentada en intervalos iguales y otro vertical que, mediante barras, muestra el número de sujetos que contiene cada categoría. *R-Commander* calcula la longitud de los intervalos mediante algoritmos automáticos, de manera que todos tengan igual amplitud y definan barras de igual anchura. La secuencia para realizar este gráfico desde el menú principal es la siguiente:

Gráficas – Histograma

En el cuadro de diálogo se seleccionará la variable a representar, en este caso *Edad*, y en la pestaña *Opciones* la escala deseada para el eje vertical, cuyos valores representan habitualmente el número o el porcentaje de sujetos en cada intervalo.



Tras pulsar el botón *Aceptar*, el histograma aparecerá en la ventana gráfica.



El gráfico muestra una distribución de valores asimétrica, sesgada a la derecha, con tres frecuencias máximas localizadas en los intervalos de edad 20-25, 45-50 y 50-55. Este tipo de distribuciones se denomina *multimodal* y suele aparecer cuando están mezclados datos que proceden de distintos grupos o poblaciones.

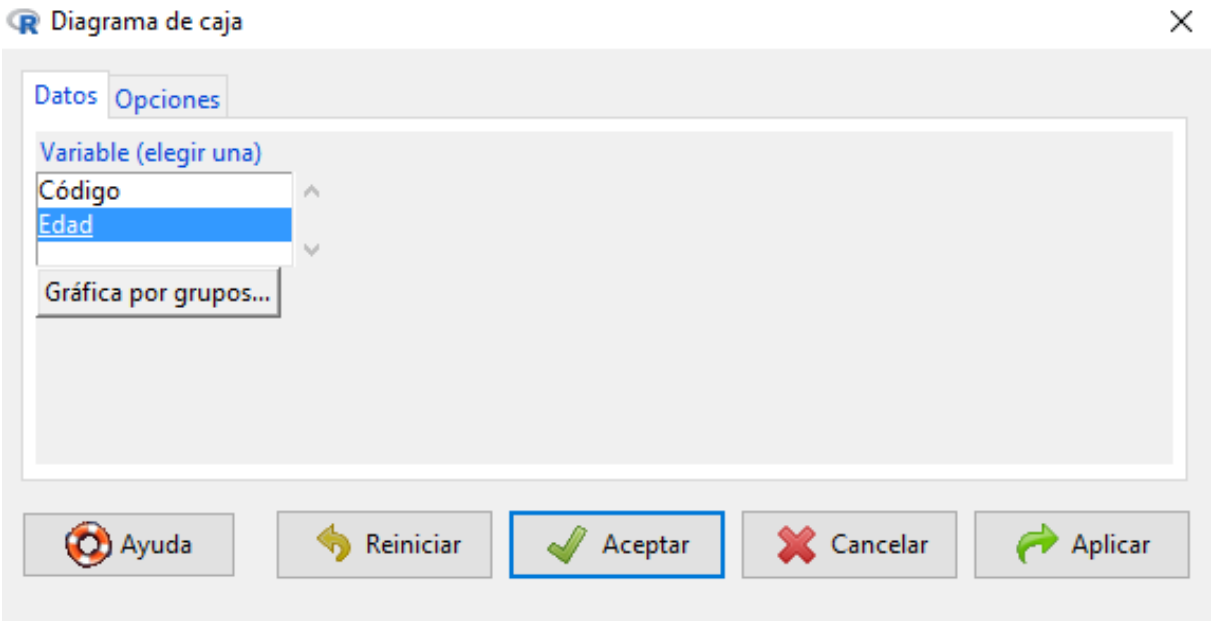
Gráfico de caja

Otra de las representaciones gráficas utilizada para la descripción de variables cuantitativas es el gráfico de caja, también denominado *box-plot* o *box-and-whisker plot*. Su construcción se basa en los cuartiles de la variable, siendo un gráfico muy útil para visualizar la dispersión de los datos, conocer la simetría de su distribución e identificar casos raros o atípicos, es decir, valores que se diferencian notablemente del resto de los valores del grupo.

La siguiente secuencia permite realizar este tipo de gráfico desde el menú principal de *R-Commander*:

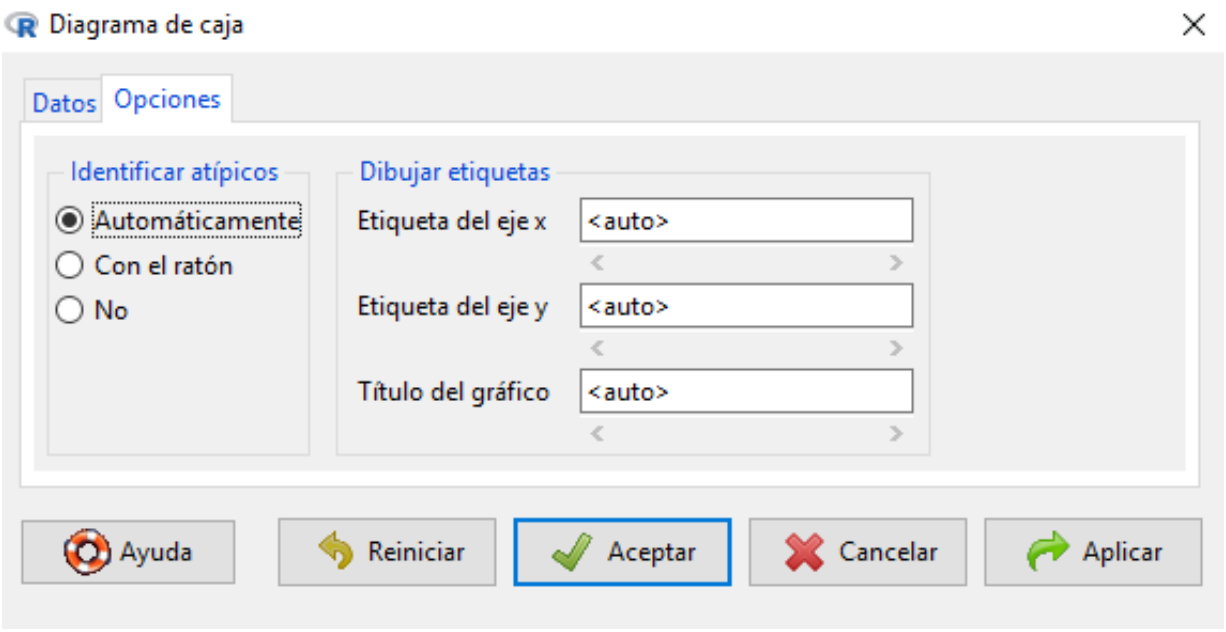
Gráficas – Diagrama de caja

El cuadro de diálogo mostrará las variables cuantitativas que pueden representarse, entre las que se seleccionará *Edad*.

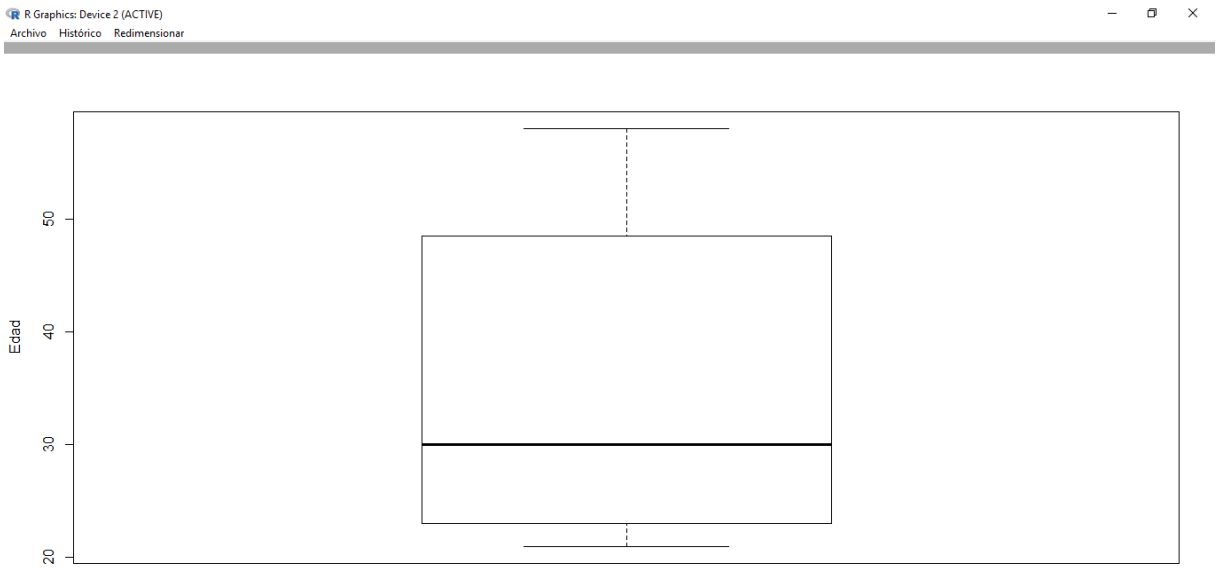


En la pestaña *Opciones* se encuentra activada por defecto la opción "*Identificar atípicos automáticamente*", que ofrece información sobre los casos raros o atípicos, marcados con un círculo por *R-Commander*. También es posible seleccionar "*Identificar atípicos con el ratón*", de manera que al hacer clic con el ratón sobre cada uno de estos sujetos aparecerá el número de fila que ocupa dentro de la base de datos. Estos procedimientos pueden ser útiles para identificar errores en el registro de la información o identificar valores extremos dentro de un grupo.

Al igual que en los gráficos de barras, sectores e histograma, la pestaña *Opciones* ofrece también la posibilidad de modificar el nombre de los ejes y el título del gráfico que aparece por defecto.



Tras pulsar el botón *Aceptar*, el gráfico de caja se mostrará en la ventana gráfica de *R-Commander*.



Los lados inferior y superior del rectángulo se sitúan a la altura del primer y tercer cuartil respectivamente. La línea central en negrita es la mediana. Todos los valores de la edad comprendidos entre las marcas dibujadas al final de las líneas punteadas se considerarán no atípicos. Si en la base de datos hubiese valores de la edad por debajo de la marca inferior o por encima de la marca superior quedarían señalados con un círculo, indicando que se trata de casos raros. En ocasiones, estas marcas se denominan límites de admisibilidad.

En este caso no hay valores atípicos para la edad. La mediana está desplazada respecto al centro del rectángulo y su distancia al límite superior es mayor que la distancia al límite inferior, lo que indica que la distribución de los valores es asimétrica. El sesgo a la derecha sugiere mayor heterogeneidad entre los sujetos que tienen una edad superior a la mediana.

PRESENTACIÓN DE RESULTADOS

La información que muestra la tabla de frecuencias, el gráfico de barras y el diagrama de sectores es equivalente. Por este motivo sería redundante ofrecer los tres resultados en un informe o presentación de diapositivas. Además de ocupar demasiado espacio cuando se analizan varias variables dificultaría el resumen de la información en un mensaje claro y conciso. Igual ocurre con el histograma y el gráfico de cajas para variables cuantitativas.

La forma de presentar los resultados descriptivos dependerá del objetivo que se desee alcanzar. Aunque no existen normas preestablecidas, habitualmente un artículo científico suele incluir sólo tablas de frecuencia y resúmenes numéricos y únicamente de forma excepcional algún gráfico que permita destacar información relevante no recogida en las tablas. Por el contrario, una presentación oral, en la que el oyente no dispone de tiempo para procesar grandes cantidades de información numérica, es preferible incorporar gráficos que permitan recibir el mensaje de forma visual, rápida y concisa.

La selección de resultados descriptivos relevantes y su disposición en tablas o gráficos autoexplicativos, claros y sencillos facilitará enormemente la lectura y comprensión del mensaje transmitido. La presentación habitual mediante tablas es la siguiente:

Descripción de variables cualitativas

Variable	Número de sujetos	Porcentaje de sujetos
Grupo		
Formación	12	48%
No formación	13	52%
Estado		
Accidentado	17	68%
No accidentado	8	32%
Sexo		
Hombre	13	52%
Mujer	11	44%
Valores perdidos	1	4%
:		

Nota: Se incorporarán a la primera columna de la tabla tantas variables como sea necesario, siguiendo la misma estructura para cumplimentar su información.

Descripción de variables cuantitativas

Variable	Sujetos	Mínimo	Máximo	Media	Desviación típica
Edad	24	18	58	34.38	13.24
:					

Nota: Se incorporarán a la primera columna de la tabla tantas variables como sea necesario, siguiendo la misma estructura para cumplimentar su información.

5

ANÁLISIS DESCRIPTIVO BIVARIANTE

El análisis estadístico de la información no se circunscribe únicamente a la descripción de las características de una población o de un grupo de sujetos. Uno de los principales objetivos de muchas investigaciones es estudiar la relación entre dos variables, observando cómo cambian los valores de una de ellas cuando se modifican los de la otra. En este contexto surge el concepto de variable dependiente y variable independiente, cuya definición previa es fundamental para abordar este tipo de análisis.

La variable independiente es aquella que a priori se considera como la causa, o una de las posibles causas, del efecto estudiado, cuyos valores constituyen la variable dependiente. En la investigación experimental, el investigador manipula la variable independiente para observar el cambio que se produce en la variable dependiente, de manera que conociendo los valores de la primera se podría predecir el comportamiento de esta última. Por ello, la variable independiente también se conoce como predictora, explicativa, exposición o causa, en cuyo caso la variable dependiente suele recibir respectivamente el nombre de respuesta, explicada, enfermedad o efecto.

El carácter cualitativo o cuantitativo de una variable es intrínseco a ella. Sin embargo, su cualidad de independiente o dependiente dependerá del objetivo del estudio. Así, en una investigación sobre los factores relacionados con el peso del recién nacido, la variable peso al nacer será la variable dependiente o resultado final. Por contra, en un estudio sobre factores de riesgo relacionados con la mortalidad neonatal, el peso al nacer actuará ahora como variable independiente o predictora de la mortalidad, que será la dependiente.

El análisis bivariante describe la relación entre dos variables, donde habitualmente una de ellas actúa como independiente y otra como dependiente. Este análisis engloba varias técnicas estadísticas, cuyo uso particular dependerá del carácter cualitativo o cuantitativo de las variables analizadas. En este capítulo se exponen los métodos más utilizados para este propósito.

VARIABLE DEPENDIENTE CUALITATIVA

Siempre que la variable dependiente sea cualitativa, la técnica estadística más utilizada para describir su relación con otras variables, cualitativas o cuantitativas, es la tabla de contingencia. En su forma más sencilla, esta tabla presenta una doble entrada, donde las categorías de la variable independiente (exposición o causa) se disponen habitualmente en las filas y las categorías de la variable dependiente (enfermedad o efecto) en las columnas. Si la variable independiente fuese cuantitativa se segmentará en dos o más grupos para conseguir una tabla con el siguiente formato:

		Dependiente		
		Enfermedad	No enfermedad	
Independiente	Exposición	a	b	a+b
	No exposición	c	d	c+d
		a+c	b+d	a+b+c+d

Las celdas de la tabla representan el número de sujetos que tienen una determinada característica. Así, hay *a* personas expuestas y enfermas, *b* expuestas y no enfermas, *c* no expuestas y enfermas y *d* no expuestas ni enfermas.

Junto a estos números absolutos, en los estudios de cohortes y transversales es útil calcular los que se denomina porcentaje por filas, es decir, la proporción de personas enfermas tanto en el grupo de expuestos como en el de no expuestos. De esta forma, [a/(a+b)] *x* 100 será el porcentaje de enfermos entre las personas que estuvieron expuestas y [c/(c+d)] *x* 100 el porcentaje de enfermos entre las no expuestas. En un estudio de cohortes, estos valores pueden interpretarse como la incidencia acumulada de la enfermedad en cada uno de los grupos de exposición y su cociente es la razón de incidencias o *Riesgo Relativo* (*RR*). En un estudio transversal ambos porcentajes serán la prevalencia de la variable dependiente en cada grupo de la variable independiente y su cociente la *Razón de Prevalencias* (*RP*). Si exposición y enfermedad no están relacionadas, la incidencia o la prevalencia serán similares en cada grupo de exposición.

En los estudios de casos y controles, suele obtenerse el porcentaje por columnas, describiendo de forma separada las características de las personas enfermas (casos) y las características de las personas no enfermas (controles). En los casos, la proporción de sujetos expuestos será [a/(a+c)] *x* 100, mientras que en los controles este porcentaje será [b/(b+d)] *x* 100. Si la variable independiente no está relacionada con la dependiente, la proporción de sujetos expuestos será similar en el grupo de los casos y en el grupo de los controles.

Cualquiera que sea el tipo de diseño, el producto cruzado (a *x* d)/(c *x* b) se denomina *Odds Ratio* (*OR*), traducido como razón de ventajas o razón de oportunidades. Es una medida de asociación que generalmente representa la oportunidad de enfermar de una persona expuesta con respecto a otra no expuesta. Esta definición será válida siempre que la variable independiente esté situada en las filas, la dependiente en columnas y las personas *enfermas-expuestas* en la primera celda de la tabla. De no ser así, la interpretación del producto cruzado anterior será diferente. Igualmente, para que la interpretación de los porcentajes por filas o columnas coincida con el definido anteriormente, la variable independiente ha de estar en las filas y la dependiente en las columnas.

El caso práctico *Accidentes por pinchazo en profesionales de enfermería* se diseñó para investigar los factores relacionados con este tipo de accidentes. Aquí, la variable dependiente es *Estado*, variable cualitativa con dos categorías que recoge si el profesional tuvo algún accidente por pinchazo al final del seguimiento. Las variables *Grupo*, *Sexo* y *Edad* son variables independientes, las dos primeras cualitativas y la última cuantitativa. Puesto que la variable dependiente es cualitativa, la técnica estadística que se utilizará para describir su relación con el resto de variables será la tabla de contingencia. Los siguientes apartados se basan en este caso práctico para mostrar cómo hacerlo con *R-Commander*.

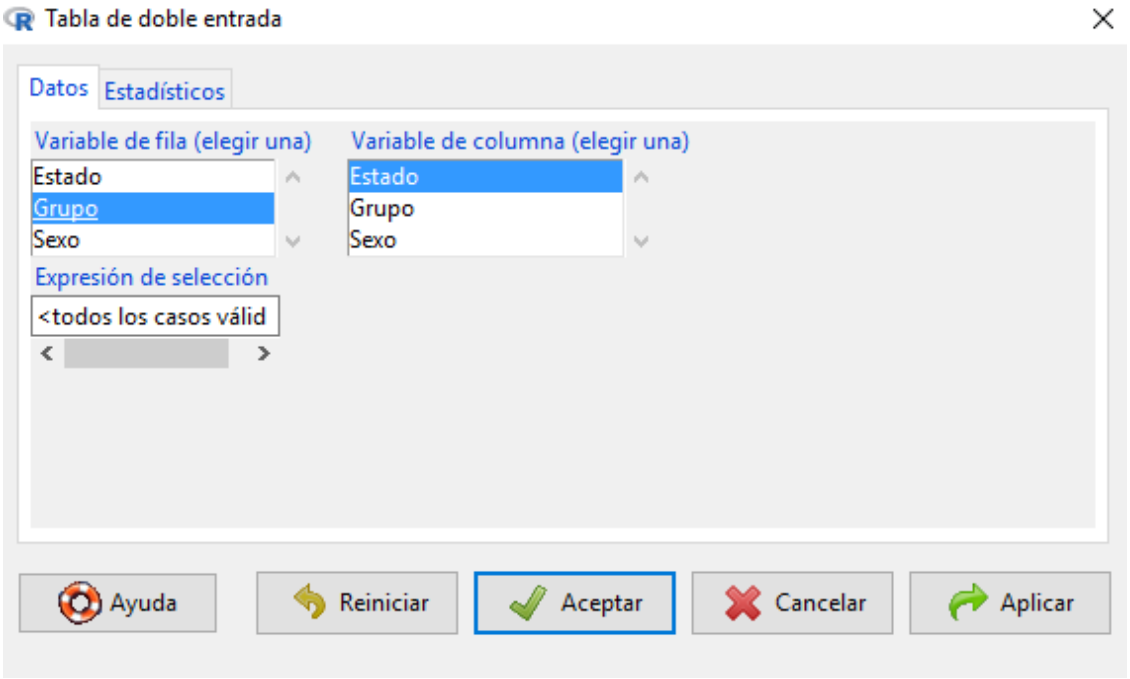
Tabla de contingencia con variable independiente cualitativa

El principal objetivo del estudio de accidentes por pinchazo era evaluar la eficacia del programa de formación sobre la disminución de accidentes. Para comprobarlo bastará con hacer un recuento de accidentes en el grupo de profesionales que recibió formación específica y en el que no la recibió, de manera que si el programa fuese eficaz se esperaría encontrar un porcentaje de accidentes menor en el grupo que recibió formación.

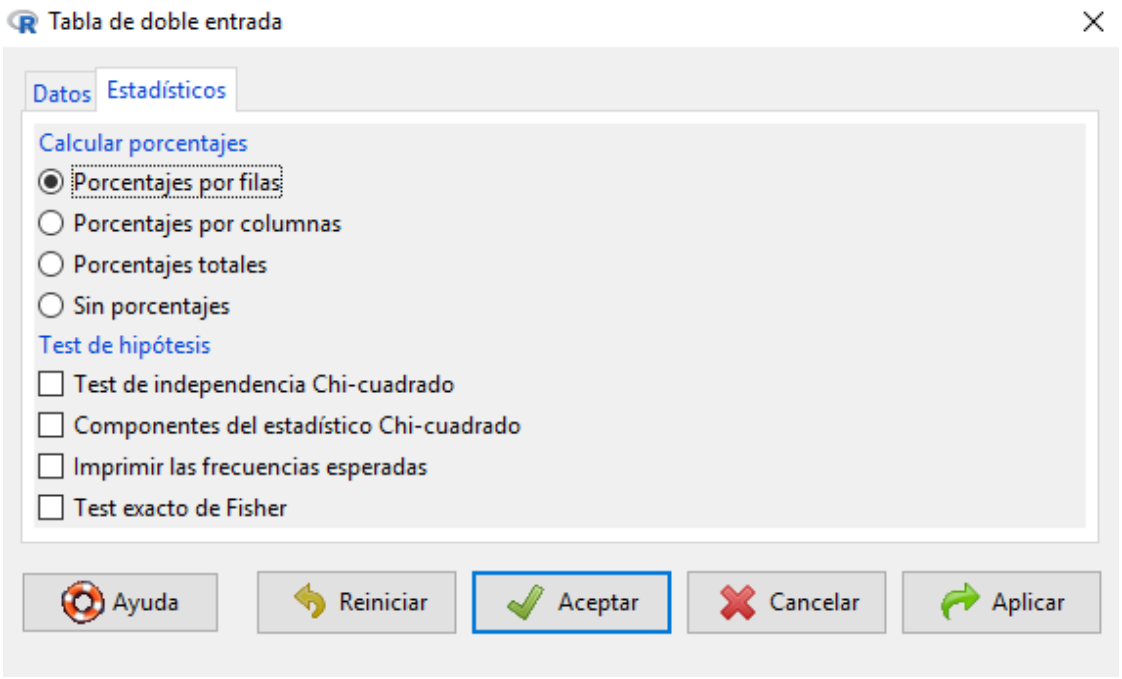
En este caso la variable dependiente es el estado del profesional al final del seguimiento (*Accidentado - No accidentado*) y la independiente el grupo al que pertenece (*Formación – No formación*). Por ser un análisis de dos variables la tabla de contingencia será de doble entrada, pudiendo realizarse desde el menú principal de *R-Commander* siguiendo esta secuencia:

Estadísticos – Tablas de contingencia – Tabla de doble entrada

Tras ejecutarla aparecerá un cuadro de diálogo que muestra dos grupos de variables cualitativas. A la izquierda, bajo el título "*Variable de fila*", se elegirá la variable independiente (*Grupo*) y a la derecha, bajo el título "*Variable de columna*", la dependiente (*Estado*).

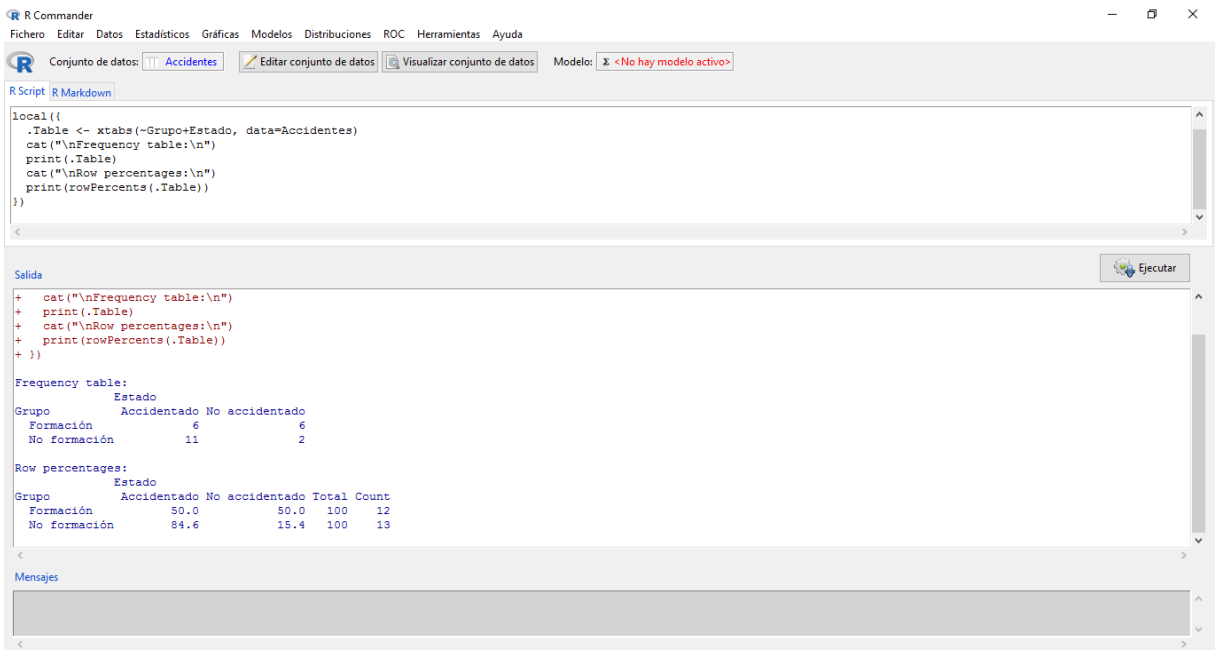


En la pestaña *Estadísticos*, la opción "*Calcular porcentajes*" permitirá completar la tabla calculando para cada celda la proporción de sujetos con respecto al total de individuos de su fila, de su columna o del global de la base de datos.



El diseño de este estudio es de seguimiento, por lo que el porcentaje activado será por filas para obtener la incidencia de accidentes en el grupo que recibió formación y en el grupo que no la recibió. Las opciones presentadas bajo el título "*Test de hipótesis*" corresponden a métodos de inferencia estadística que no serán abordados en este capítulo.

Tras pulsar el botón *Aceptar* se mostrará la siguiente información en la ventana de resultados de *R-Commander*:



Estos resultados deberán transcribirse a un procesador de textos para confeccionar una tabla como la mostrada a continuación que contenga, de momento, la siguiente información:

Variables	Estado Accidentado	No accidentado	RR(*)
Grupo			
Formación	6 (50.0%)	6 (50.0%)	0.59
No formación	11 (84.6%)	2 (15.4%)	1

(*) *R-Commander* no ofrece el riesgo relativo (*RR*) en la salida de resultados

Entre los profesionales que recibieron formación, el 50% se accidentó. En el grupo que no recibió información el porcentaje de accidentes fue del 84.6%. Con esta información, la accidentabilidad es inferior en el grupo de profesionales que recibió información. De hecho, el *RR* calculado mediante el cociente 50.0/84.6=0.59 indica que el riesgo de accidente de las personas que recibieron formación es inferior al de los profesionales que no recibieron formación, ya que su valor es menor que 1. En concreto, el riesgo de accidente es un 41% inferior en las personas con formación con respecto a las personas sin formación. Esta última

categoría, con respecto a la que se realiza la comparación, se denomina *categoría de referencia* y suele señalarse con el valor 1 en la tabla de resultados anterior.

En lugar del riesgo relativo, en un estudio de cohortes también puede utilizarse la *OR* como medida de asociación. Calculada mediante el producto cruzado (6 x 2) / (11 x 6) = 0.18, su valor inferior a 1 sugiere un efecto protector de la formación sobre los accidentes por pinchazo, con magnitud diferente al riesgo relativo.

Las tablas de contingencia realizadas con *R-Commander* no ofrecen el valor del *RR* ni el de la *OR*, obtenida como producto cruzado. Para calcularlos sin recurrir a comandos será necesario escribir las expresiones numéricas en la ventana de instrucciones y pulsar posteriormente el botón *Ejecutar*, utilizar una calculadora externa o emplear una calculadora estadística como *OpenEpi* (*www.openepi.com*).

Tabla de contingencia con variable independiente cuantitativa

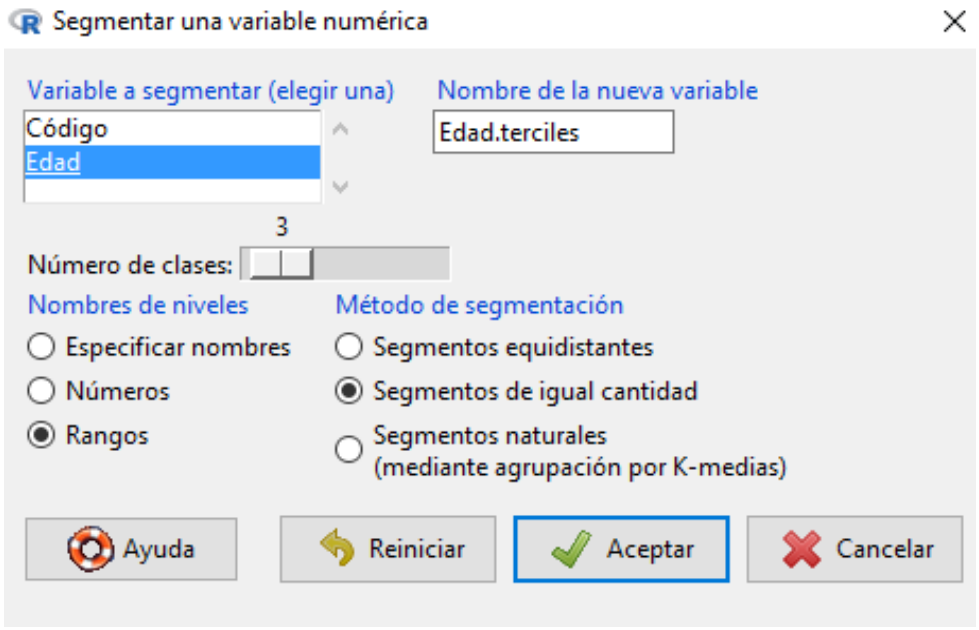
Otro de los objetivos del estudio de accidentes por pinchazo era estudiar la relación de las variables *Sexo* y *Edad* con el *Estado* de los profesionales al final del seguimiento, siendo ésta última la variable dependiente. Puesto que *Sexo* es una variable independiente cualitativa, su relación con *Estado* se describirá mediante una tabla de contingencia, utilizando el mismo procedimiento del apartado anterior. Sin embargo, la relación entre *Edad* y *Estado* requerirá previamente segmentar la variable cuantitativa *Edad* en dos o más grupos para convertirla en cualitativa y poder realizar una tabla de contingencia.

El número de categorías a efectuar y los puntos de corte usados para segmentar una variable independiente cuantitativa dependerán de las hipótesis del estudio. Si no hubiera una hipótesis de partida clara se recurrirá a criterios clínicos o epidemiológicos, tomando las categorías y puntos de corte consensuados en la literatura científica internacional. Por último, si tampoco se dispone de criterios epidemiológicos estandarizados se recurrirá a criterios estadísticos, recodificando la variable o segmentándola en intervalos con el mismo número de sujetos, equidistantes o naturales como se describió en el Capítulo 3, dentro del apartado *Obtener nuevas variables a partir de las existentes: Calcular, recodificar y segmentar*.

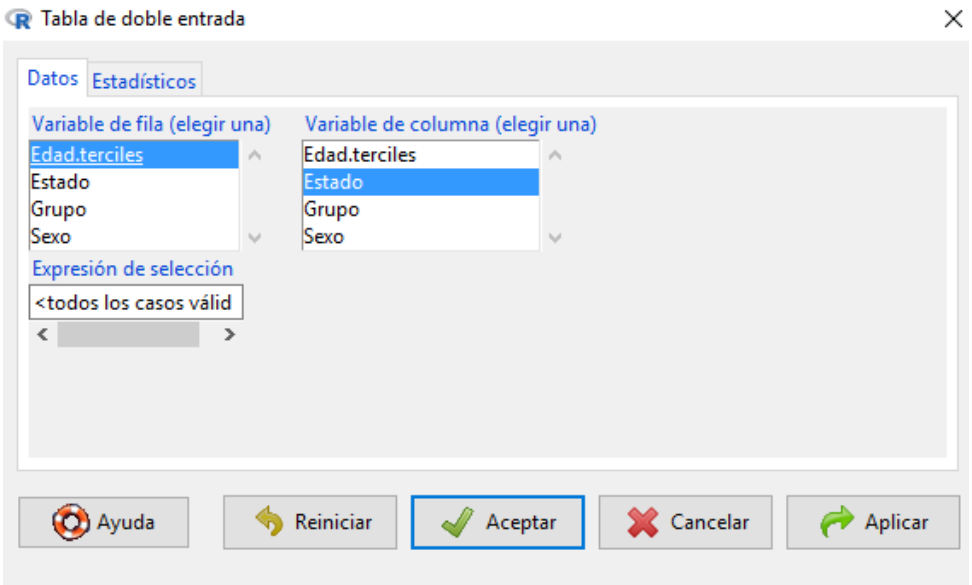
En este caso no existen hipótesis de partida ni criterios epidemiológicos que puedan ser utilizados para segmentar la variable *Edad*. Por ello, a modo de ejemplo, se dividirá en tres grupos de igual tamaño utilizando los percentiles 33 y 66 como puntos de corte, de manera que cada intervalo contenga al 33% de los profesionales. Para realizar este procedimiento, desde el menú principal se activará la secuencia:

Datos - Modificar variables del conjunto de datos activo – Segmentar variable numérica

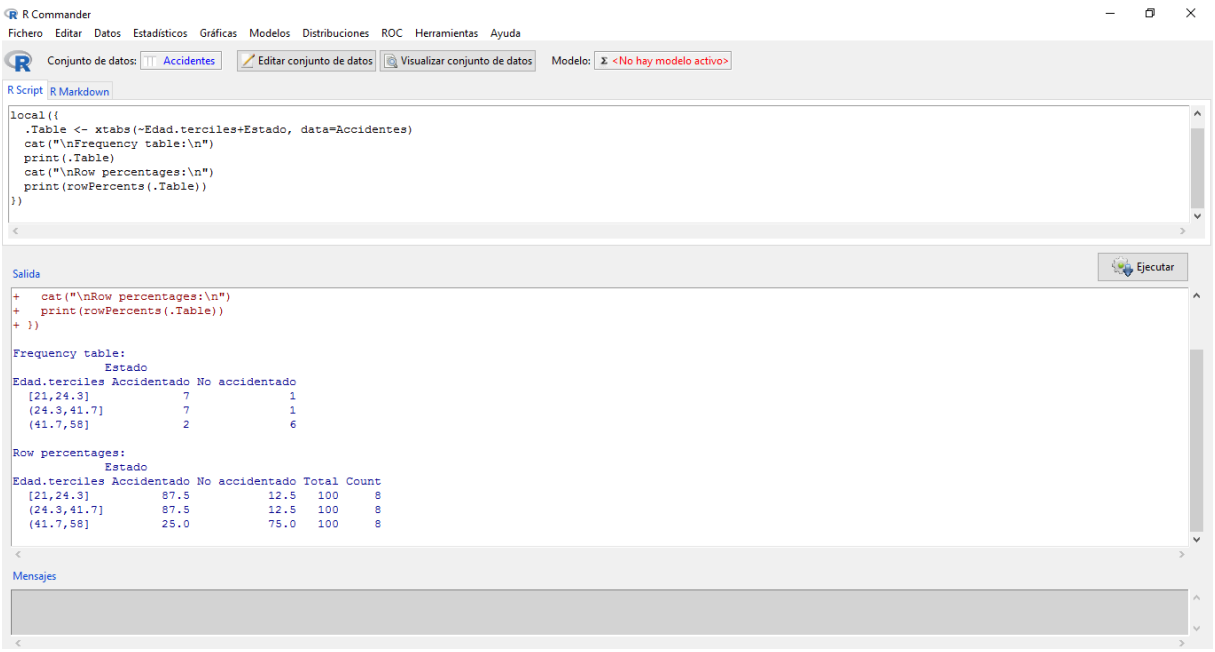
En la ventana emergente se seleccionará la variable *Edad*, marcando como opciones 3 clases, *segmentos de igual cantidad* como método de segmentación, *rangos* como nombres de niveles y *Edad.terciles* como nombre de la nueva variable.



Una vez efectuada la segmentación se realizará una tabla de contingencia de doble entrada situando la variable independiente *Edad.terciles* en las filas y la variable dependiente *Estado* en las columnas, a través de *Estadísticos – Tablas de contingencia – Tabla de doble entrada*.



Tras pulsar el botón *Aceptar*, *R-Commander* mostrará los resultados de la tabla de contingencia.



La transcripción de la ventana de resultados a un procesador de textos permitirá elaborar una tabla con la siguiente información:

Variables	Estado Accidentado	No accidentado	RR(*)
Edad			
24.3 o menos	7 (87.5%)	1 (12.5%)	3.5
24.4 - 41.7	7 (87.5%)	1 (12.5%)	3.5
41.8 o más	2 (25.0%)	6 (75.0%)	1

(*) *R-Commander* no ofrece el riesgo relativo (RR) en la salida de resultados

La proporción de accidentados disminuye con la edad, de manera que tomando como referencia el grupo de 41.8 o más años se tiene que, con respecto a éste, el riesgo de accidente es 3.5 veces superior en cualquiera de los grupos más jóvenes. Como antes, este *RR* se obtiene al dividir 87.5 / 25.0 en la ventana de instrucciones de *R-Commander* o utilizando una calculadora. De la misma forma, también es posible calcular la *OR* de cada grupo de edad con respecto a la categoría *41.8 o más*, que en este caso sería (7 x 6) / (2 x 1) tanto para el grupo *24.3 o menos* como para el grupo *24.4 – 41.7* años.

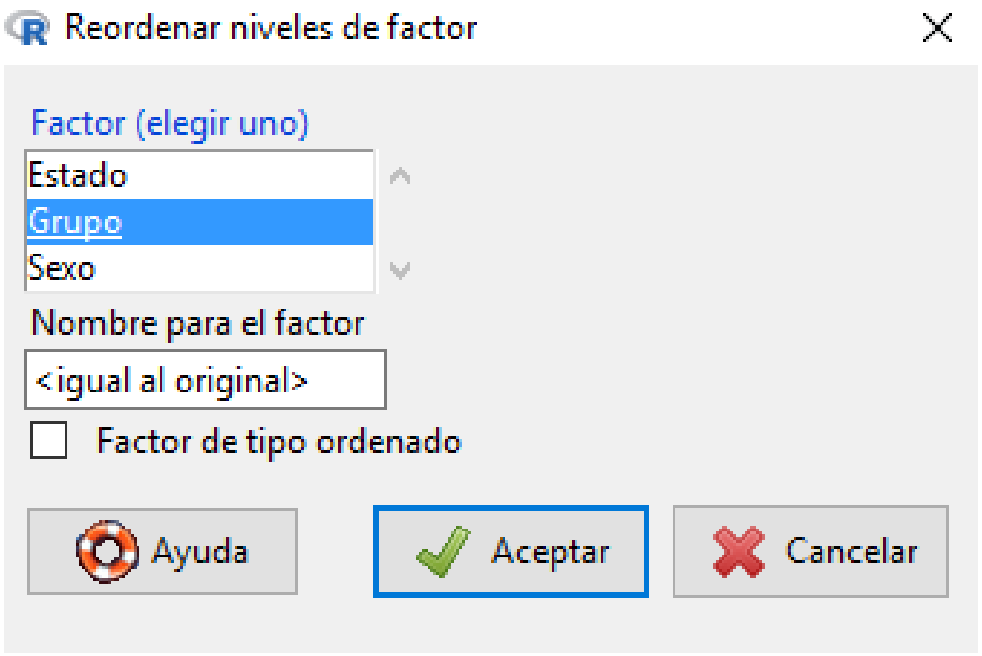
Reordenar las categorías en una tabla de contingencia

Cuando en el primer apartado se estudió la relación entre *Grupo* y *Estado*, la primera celda de la tabla de contingencia estaba dada por las categorías *Accidentado-Formación*. Por ello, la *OR* calculada mediante el tradicional producto cruzado es la oportunidad de *Accidente* de un profesional *Formado* con respecto a otro *No formado*. Esta última categoría, con respecto a la que se realiza la comparación, siempre es la categoría de referencia y aparece con el valor 1 en la tabla de resultados.

Aunque no es necesario, a veces es preferible que la tabla de contingencia aparezca configurada con el par *Enfermo-Expuesto* en la primera celda, de manera que la *OR* generada sea la oportunidad de enfermar de las personas expuestas con respecto a las no expuestas. En este caso, los expuestos son los profesionales que no recibieron formación, por lo que las categorías de la variable independiente han de reordenarse para que las filas de la tabla de contingencia aparezcan intercambiadas, mostrando en primer lugar a los profesionales no formados (expuestos) y debajo a los formados (no expuestos). Este procedimiento puede realizarse desde el menú principal siguiendo la secuencia:

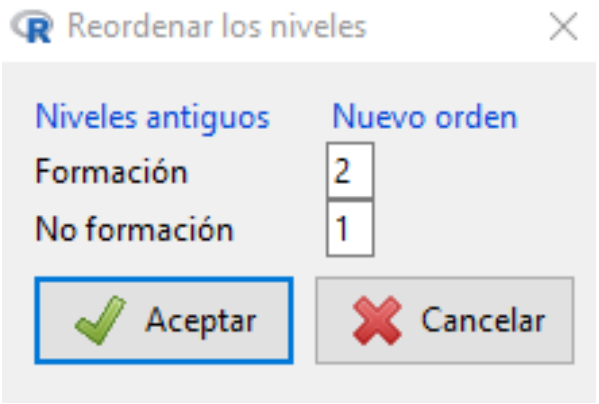
Datos - Modificar variables del conjunto de datos activo – Reordenar niveles de factor

Aparecerá un primer cuadro de diálogo que contiene las variables cualitativas de la base de datos. En este caso se seleccionará la variable *Grupo*, marcándola con el botón izquierdo del ratón.



R-Commander ofrece la opción de guardar con un nuevo nombre la variable reordenada, utilizando para ello el espacio bajo el título *"Nombre para el factor"*. En general, esto no será necesario ya que siempre se podrá reasignar el orden original de las categorías volviendo a realizar este mismo procedimiento. La opción *"Factor de tipo ordenado"* se marcará cuando la variable cualitativa sea ordinal. Puesto que *Grupo* es nominal, esta opción quedará desactivada.

Tras pulsar el botón *Aceptar* se mostrará un aviso recordando que la variable *Grupo* ya existe y se va a sobreescribir con la nueva reordenación de categorías. Una vez confirmada la acción *R-Commander* abrirá una ventana con dos columnas denominadas respectivamente "*Niveles antiguos"* y "*Nuevo orden"*. Bajo la primera columna aparecerán las categorías de la variable cualitativa en el orden original: En primer lugar, *Formación*, con el número 1 asignado a la derecha. En segundo lugar, *No formación*, con el número 2. Para reordenar estas categorías bastará con escribir la nueva numeración en los espacios de la columna "*Nuevo orden*", asignando el valor 1 a *No formación* y 2 a *Formación*.



Pulsando el botón *Aceptar*, *R-Commander* cambiará internamente el orden de las categorías de la variable *Grupo*, de manera que al rehacer la tabla de contingencia para estudiar su relación con *Estado* se obtendrá lo siguiente en la ventana de resultados:

```
> .Table <- xtabs(~Grupo+Estado, data=Accidentes)
> .Table

              Estado
Grupo          Accidentado No accidentado
  No formación          11              2
  Formación              6              6

> rowPercents(.Table) # Row Percentages

              Estado
Grupo          Accidentado No accidentado Total Count
  No formación        84.6           15.4   100    13
  Formación           50.0           50.0   100    12

> remove(.Table)
```

Ahora, la categoría *No formación* aparece en la primera fila de la tabla. El resumen de información transcrito a la tabla de un procesador de textos es el siguiente:

	Estado		
Variables	**Accidentado**	**No accidentado**	**RR(*)**
Grupo			
No formación	11 (84.6%)	2 (15.4%)	1.69
Formación	6 (50.0%)	6 (50.0%)	1

(*) *R-Commander* no ofrece el riesgo relativo (RR) en la salida de resultados

Como antes, el 84.6% de los profesionales que no recibieron formación se accidentó, mientras que la proporción de accidentes fue del 50% en el grupo que recibió formación. El cociente entre ambos es el riesgo relativo, cuyo valor muestra que el riesgo de accidente es 1.69 veces superior en los profesionales que no recibieron formación con respecto a aquellos que la recibieron. Como es usual, el 1 insertado en la columna *RR* para la categoría *Formación* señala la categoría de referencia.

De la misma forma, puesto que la primera celda de la tabla está formada ahora por el par de categorías *Accidentado-No formación*, la *OR* calculada mediante el cociente (11x6)/(6x2)=5.5 indicaría que la oportunidad de *Accidente* de las personas *No formadas* es 5.5 veces superior con respecto a los profesionales que recibieron formación.

Si se desea mantener la reordenación de categorías para utilizarla en futuras sesiones de trabajo será necesario guardar la base de datos en formato *R-Commander*. De no ser así, al cerrar el programa se perderán todos los cambios efectuados.

Presentación de resultados

Dos de los objetivos del caso *práctico Accidentes por pinchazo en profesionales de enfermería* eran, por un lado, evaluar la eficacia del programa de formación en la disminución de los accidentes. Por otro, estudiar la relación de la edad y el sexo con el estado de los profesionales al final del seguimiento.

Para responder a estos objetivos de forma clara y comprensible es necesario resumir los resultados del análisis de datos en una tabla o en un gráfico sencillo, incluyendo únicamente la información necesaria. Cuando el método estadístico se basa en tablas de contingencia, la forma usual de hacerlo es elaborando una tabla que contenga la siguiente información para cada una de las variables independientes:

Variables	Estado Accidentado	No accidentado	RR
Grupo			
No formación	11 (84.6%)	2 (15.4%)	1.69
Formación	6 (50.0%)	6 (50.0%)	1
Edad			
24 o menos	7 (87.5%)	1 (12.5%)	3.5
25-41	7 (87.5%)	1 (12.5%)	3.5
42 o más	2 (25.0%)	6 (75.0%)	1
Sexo			
Hombre	10 (76.9%)	3 (23.1%)	1.41
Mujer	6 (54.5%)	5 (45.5%)	1

Así, la persona que lea el documento sabrá con un simple golpe de vista que los profesionales no formados tienen más riesgo de accidente que los formados, los más jóvenes más que los mayores y los hombres más que las mujeres.

VARIABLE DEPENDIENTE CUANTITATIVA

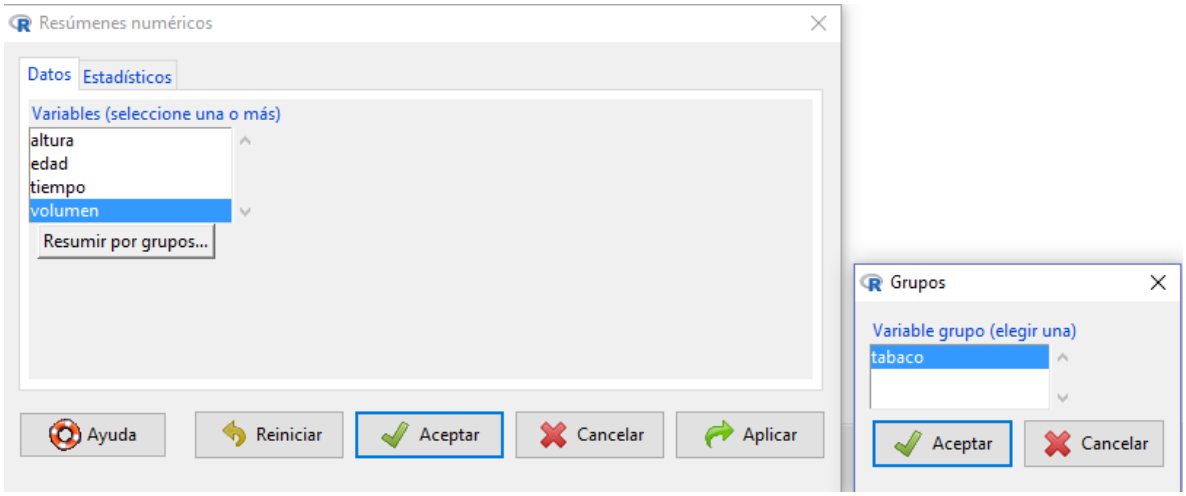
El caso práctico *Volumen espiratorio* es un estudio transversal diseñado para investigar los factores relacionados con el volumen espiratorio de personas que trabajan en la mina. La variable dependiente es *Volumen*, variable cuantitativa medida en mililitros por segundo. La variable *Tabaco* es una variable independiente cualitativa, mientras que las *variables Tiempo, Edad y Altura* son variables independientes cuantitativas. En este apartado se mostrarán los métodos estadísticos apropiados para describir la relación entre una variable dependiente cuantitativa y el resto de características.

Comparación de los grupos definidos por una variable independiente cualitativa

La variable *Tabaco* es cualitativa, con categorías *Nunca fumó*, *Exfumador* y *Fuma actualmente*. Si el tabaco tuviese relación con el volumen espiratorio, se esperaría encontrar una diferencia clínicamente importante en el volumen espiratorio de los tres grupos. Para comprobarlo se comparará de la media del volumen espiratorio de un grupo con otro a través de la siguiente secuencia del menú principal:

Estadísticos – Resúmenes – Resúmenes numéricos

En el cuadro de diálogo abierto se seleccionará la variable dependiente, en este caso *Volumen*, y en el botón "*Resumir por grupos*" la variable independiente *Tabaco*.



Puesto que para utilizar esta técnica la variable dependiente ha de ser cuantitativa, en la pantalla inicial no aparece *Tabaco* como posible elección. De igual forma, puesto que la variable independiente tiene que ser cualitativa, en el listado de variables para resumir por grupos sólo aparece *Tabaco* y no el resto de variables, que son cuantitativas.

Como en el análisis descriptivo univariante, los parámetros necesarios serán los que activa *R-Commander* por defecto: media, desviación típica y cuantiles.

Tras pulsar el botón *Aceptar* en ambos cuadros de diálogo, la ventana de resultados mostrará la siguiente salida:

```
                     mean        sd   0%    25%  50%    75% 100%  n
Nunca fumó       3977.857  933.0032 2350 3280.0 3930 4747.5 5480 14
Exfumador        4148.571 1037.3249 1720 3890.0 4190 4930.0 5900 21
Fuma actualmente 3736.667  914.6173 1770 3207.5 3585 4355.0 5780 48
```

La descripción de cada grupo se transcribirá a la tabla de un procesador de textos con la siguiente información:

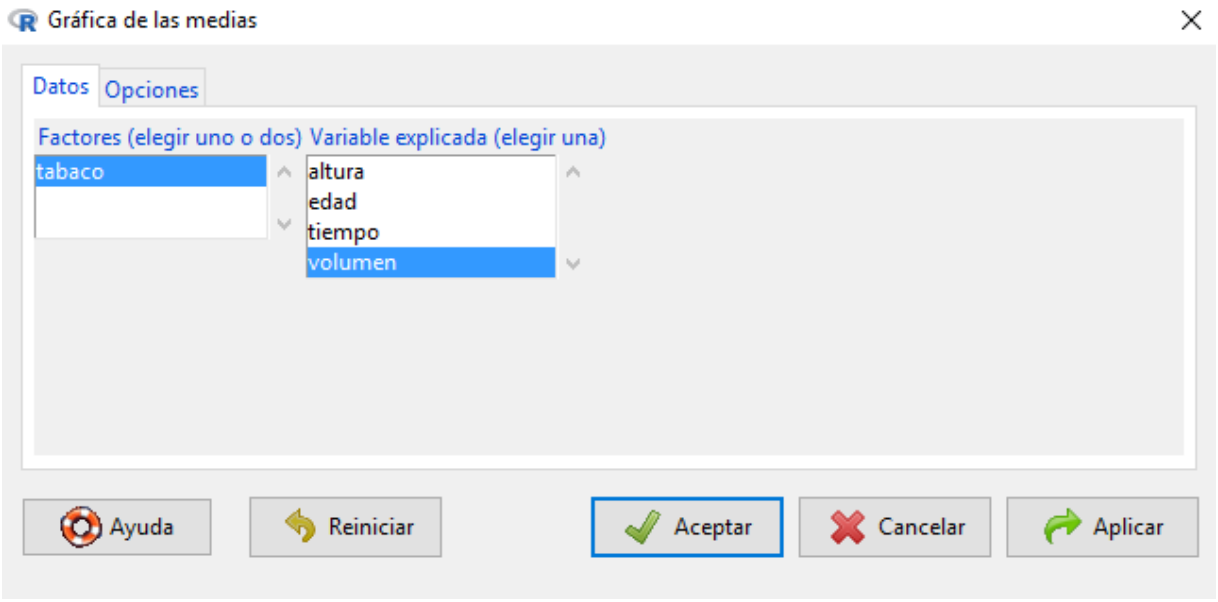
Variables	Sujetos	Mínimo	Máximo	Media	Desviación típica
Tabaco					
Nunca fumó	14	2350	5480	3977.86	933.00
Exfumador	21	1720	5900	4148.57	1037.32
Fuma actualmente	48	1770	5780	3736.67	914.62

El grupo que por término medio presenta mayor volumen espiratorio es el de exfumadores, seguido por los trabajadores que nunca fumaron. El grupo de fumadores actuales es el que muestra el menor volumen espiratorio medio.

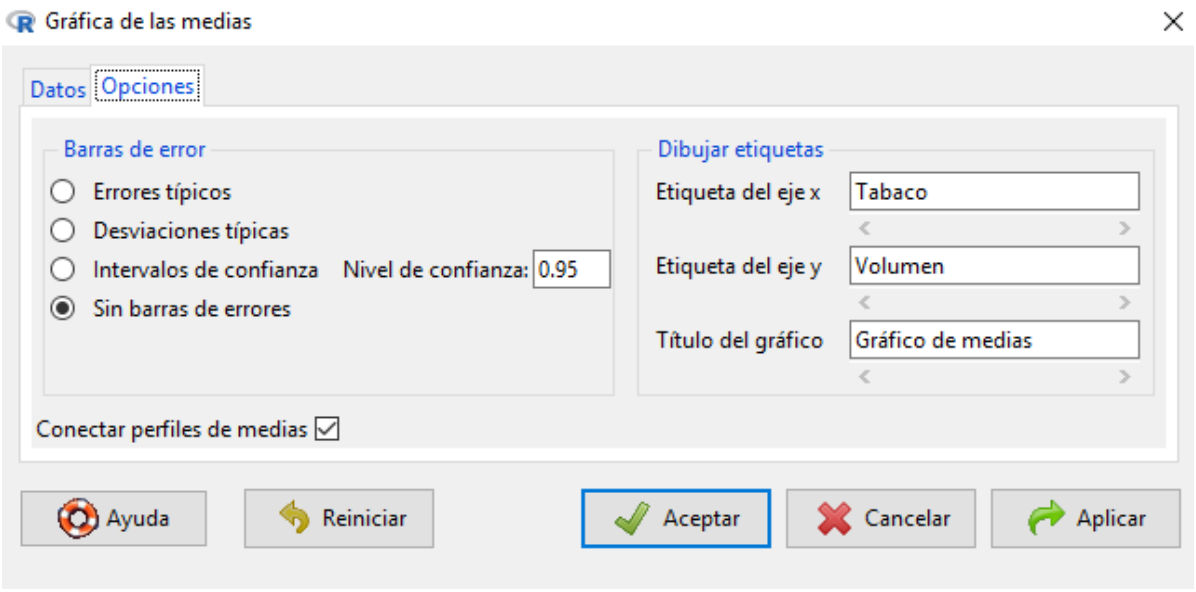
A continuación, se muestran diferentes gráficos que pueden complementar el resumen numérico anterior y son útiles para estudiar la relación entre una variable dependiente cuantitativa y una variable independiente cualitativa.

a) Gráfica de las medias

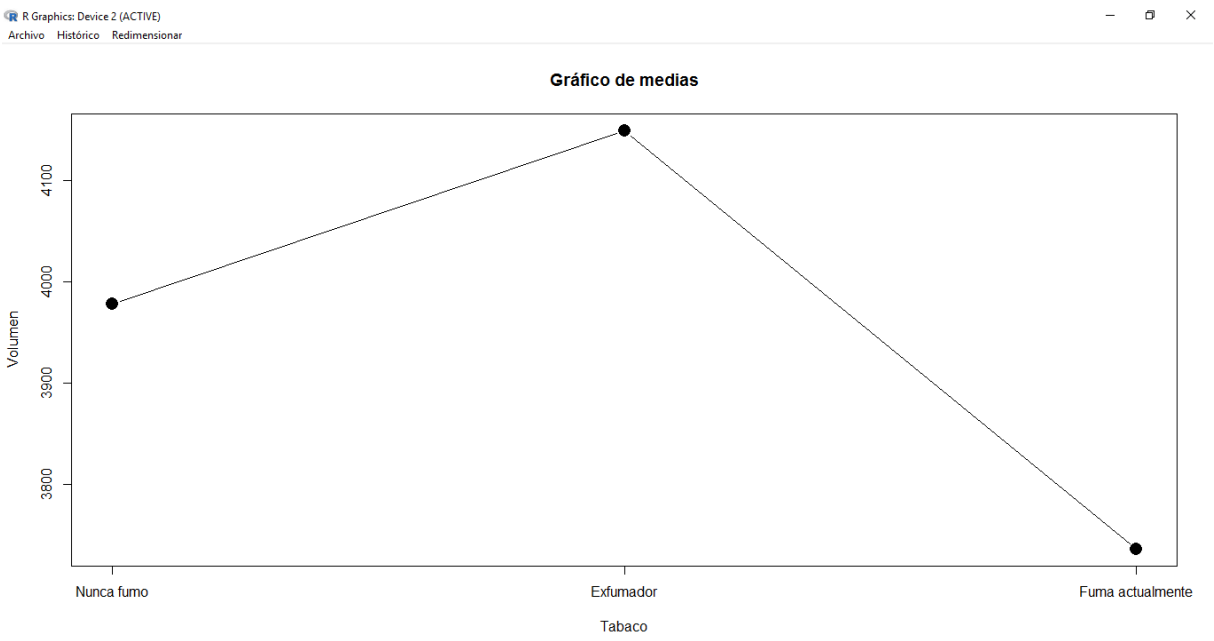
La media de cada uno de los grupos puede representarse en un gráfico con dos ejes, uno horizontal en el que se muestran las categorías de la variable independiente y otro vertical en el que se representan los valores de la variable dependiente. Para cada categoría se dibujará un punto de altura igual al valor medio de la variable dependiente en ese grupo. Este gráfico se realiza activando la secuencia *Gráficas - Gráfica de las medias* desde el menú principal. En el cuadro de diálogo abierto se seleccionará la variable independiente a la izquierda, en la columna *Factores*, y la dependiente a la derecha, en la columna *Variable explicada*.



En la pestaña *Opciones* se pueden definir los títulos de los ejes y seleccionar acciones sobre barras de error, que en este caso se desactivarán marcando "*sin barras de errores*".



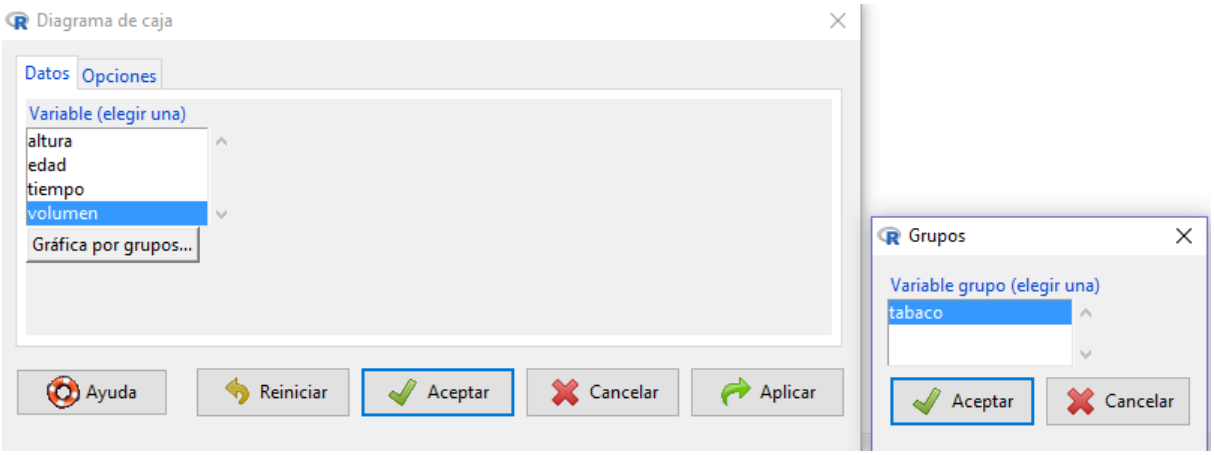
Tras pulsar el botón Aceptar, se obtendrá el gráfico buscado:



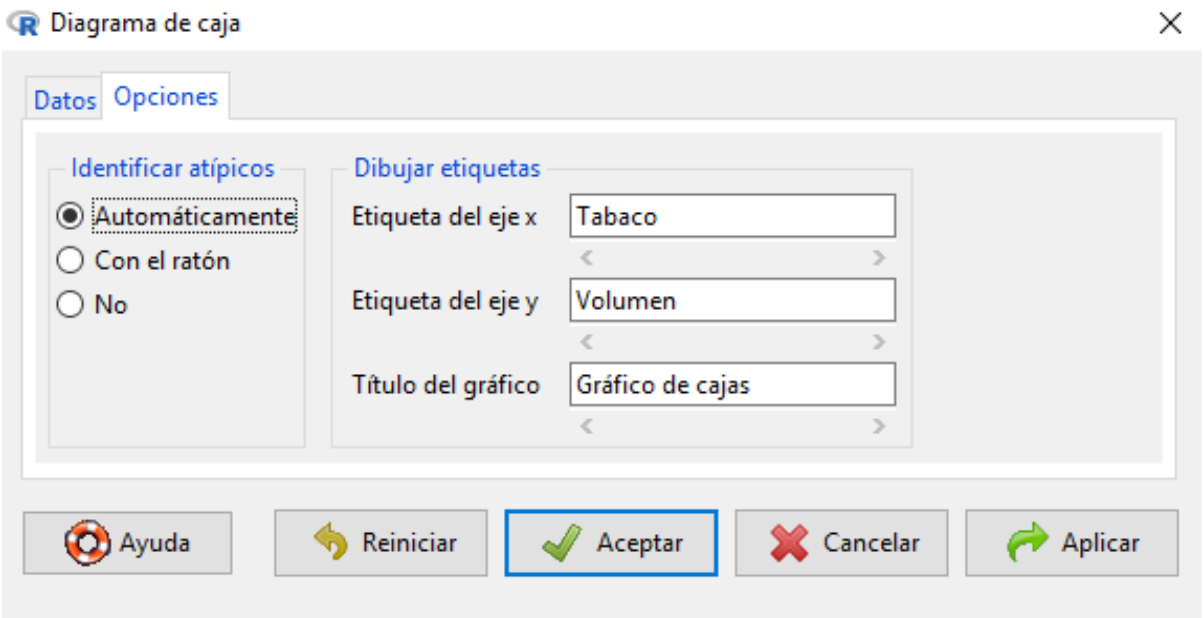
b) Gráfico de caja

Un gráfico más utilizado para comparar grupos es el diagrama de caja, donde se representa la mediana de la variable dependiente en lugar de la media. El acceso se realiza desde el menú principal mediante la secuencia *Gráficas – Diagrama de caja*. El cuadro de diálogo abierto es

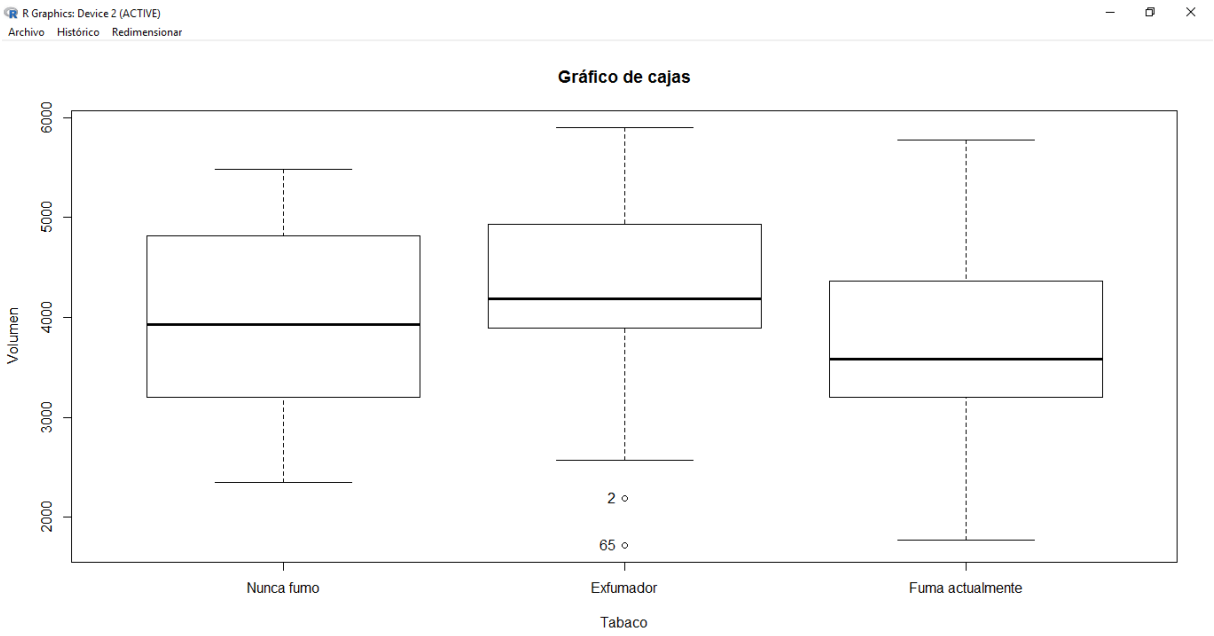
muy parecido al del procedimiento *Resúmenes numéricos* para comparar medias. En él se seleccionará la variable dependiente *Volumen*. A continuación, se pulsará el botón *Gráfica por grupos* y se elegirá *Tabaco* como variable de agrupación.



La pestaña *Opciones* permite modificar el título de los ejes y otras opciones habituales ya comentadas para los gráficos de caja.

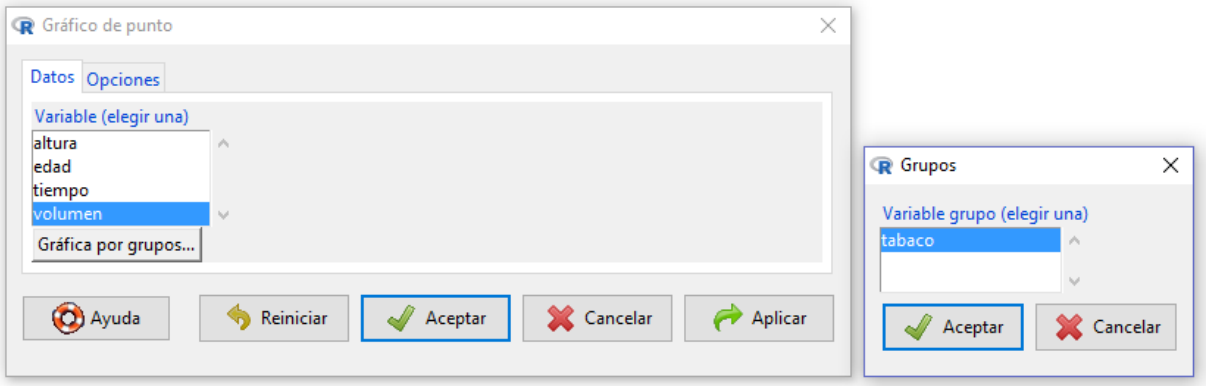


El gráfico resultante muestra dos valores atípicos en el grupo de exfumadores, identificados con el número de registro 2 y 65 respectivamente. A través de la mediana, se observa que el grupo con menor volumen espiratorio es el de fumadores actuales.



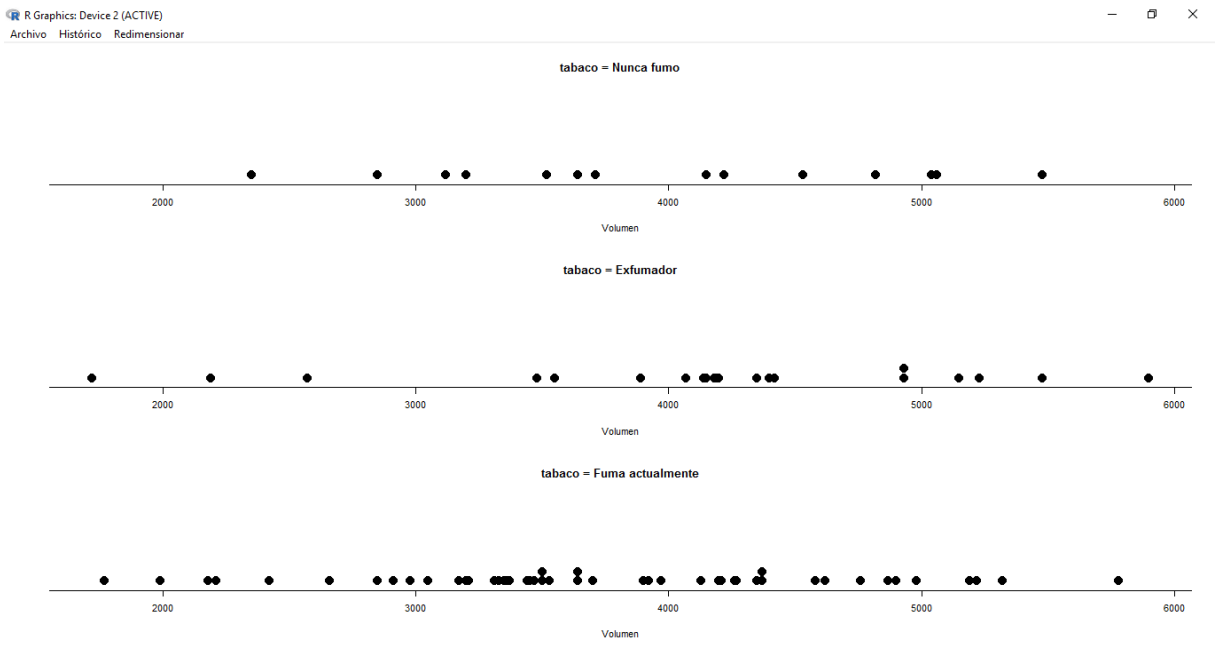
c) Diagrama de puntos

Por último, el procedimiento *Gráficas - Diagrama de puntos* realiza un gráfico donde, para cada categoría de la variable independiente, situada en el eje horizontal, se dibujará el valor de la variable dependiente para todos los sujetos del grupo.

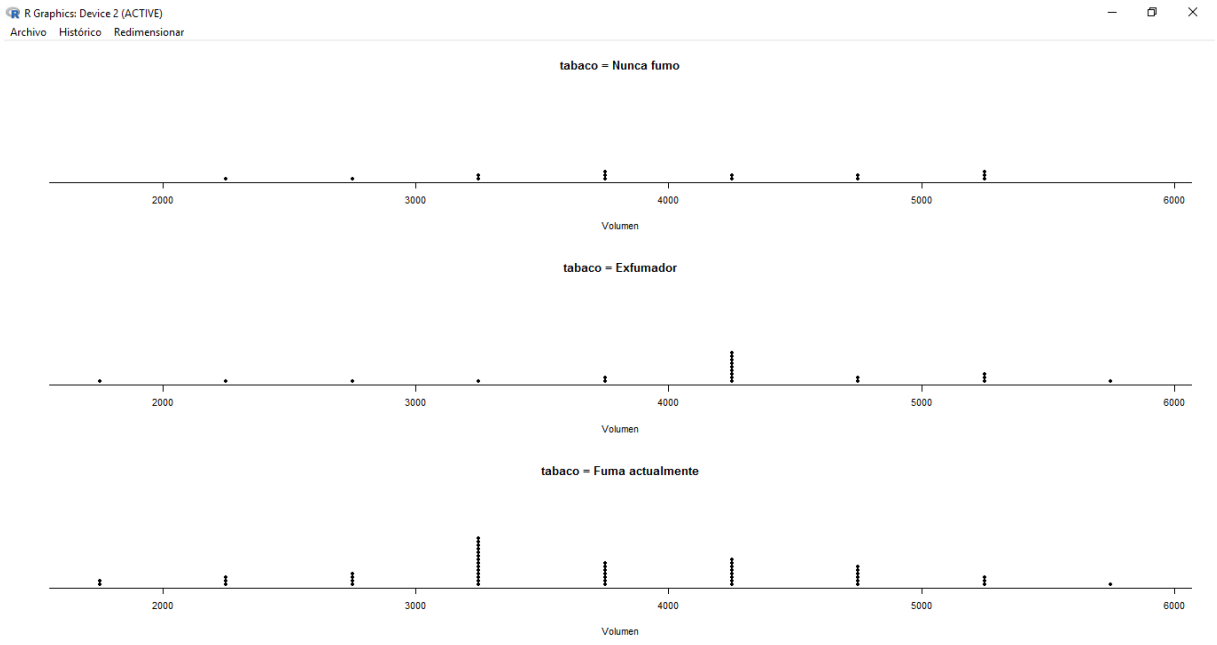


La pestaña *Opciones* ofrece la posibilidad de poner títulos al gráfico y activar la opción gráfica "*Variable binaria*", que realizará un gráfico apilando los puntos con un efecto similar al histograma.

Gráfica con opción "*Variable binaria*" desactivada:



Gráfica con opción "*Variable binaria*" activada:



Este procedimiento no se suele utilizar con demasiada frecuencia, siendo el diagrama de cajas el más interesante para comparar grupos de forma gráfica.

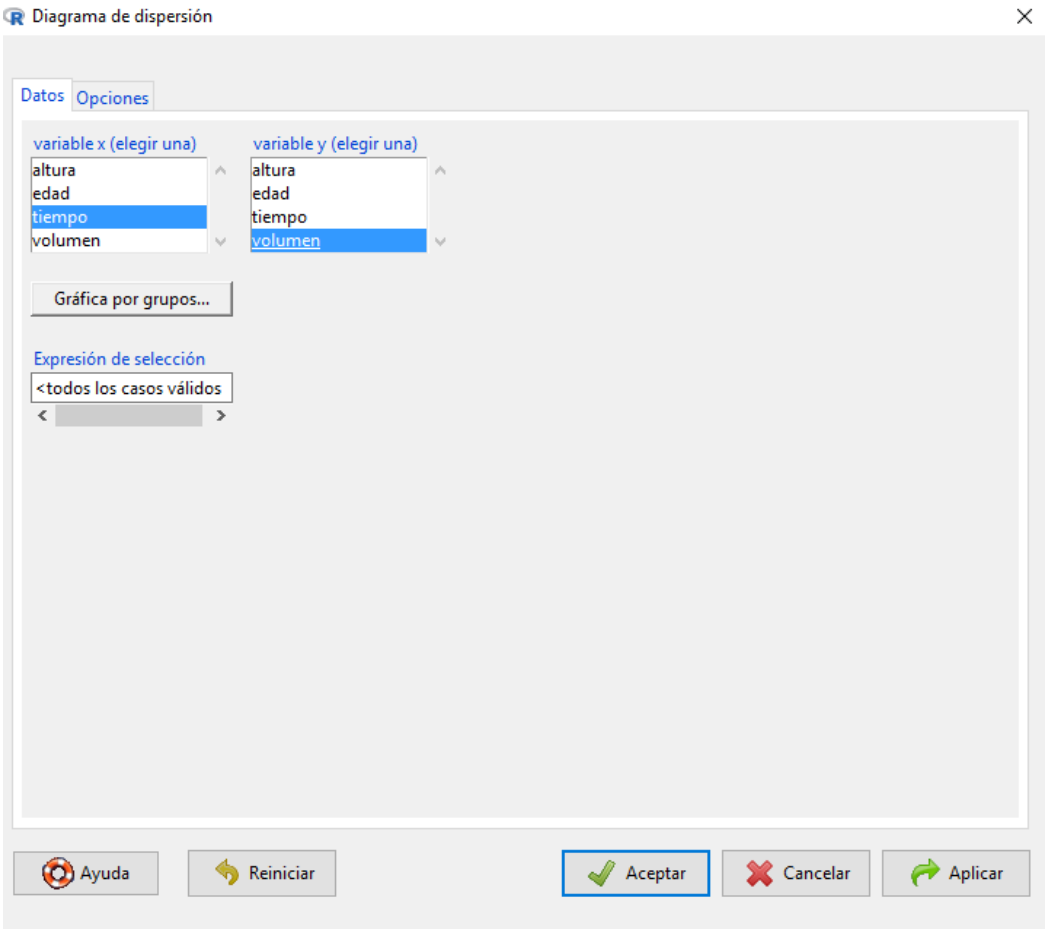
Diagrama de dispersión con variable independiente cuantitativa

Una de las hipótesis del estudio *Volumen espiratorio* era que el tiempo de exposición al polvo de la mina estaba relacionado con el volumen espiratorio, de manera que éste sería menor en los trabajadores expuestos durante más años. En este caso tanto la variable independiente *Tiempo* como la dependiente *Volumen* son cuantitativas, siendo el diagrama de dispersión la técnica apropiada para estudiar su relación.

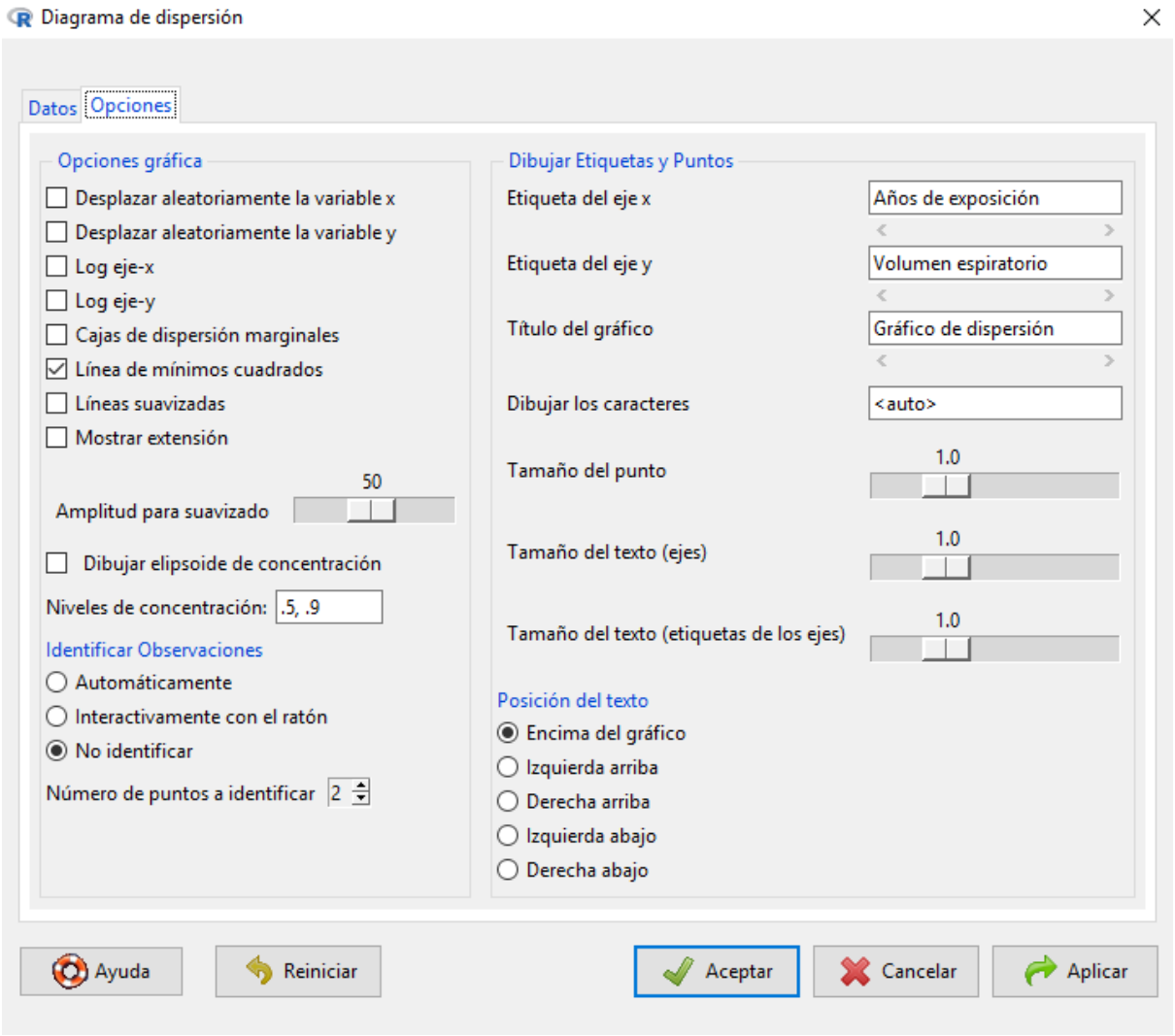
Este diagrama es un gráfico con dos ejes en el que se representan los valores la variable independiente y dependiente en el eje horizontal -X- y vertical -Y- respectivamente. Así, para cada sujeto se dibujará un punto en el plano con coordenadas dadas por el tiempo que lleva expuesto y su volumen espiratorio. Este gráfico se realiza desde el menú principal con la secuencia:

Gráficas – Diagrama de dispersión

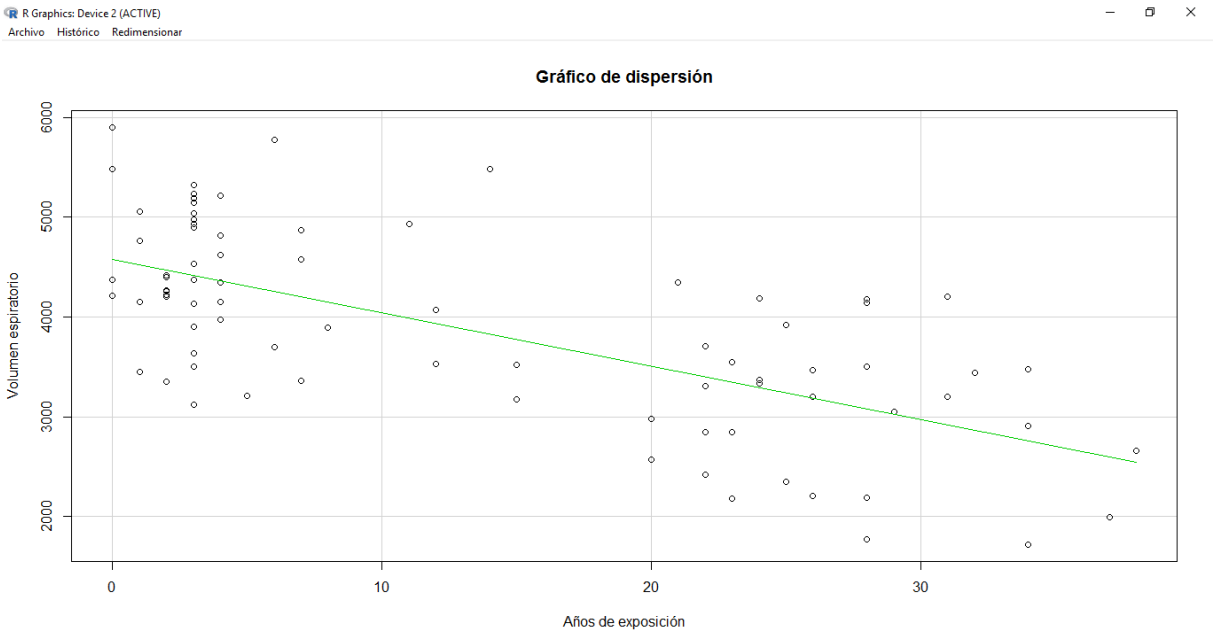
En el cuadro de diálogo se seleccionará la variable independiente en el listado de la izquierda, correspondiente al eje horizontal X, y la dependiente en el listado de la derecha, correspondiente al eje vertical Y.



A continuación, en la pestaña *Opciones*, se activará la opción *Línea de mínimos cuadrados*, que dibujará la recta que mejor representa la tendencia de los puntos. En este gráfico, *R-Commander* permite etiquetar con un título ambos ejes. Así, se escribirá "*Años de exposición*" debajo de "*Etiqueta del eje x*" y "*Volumen espiratorio*" debajo de "*Etiqueta del eje y*". El resto de opciones pueden ser útiles para personalizar el gráfico.

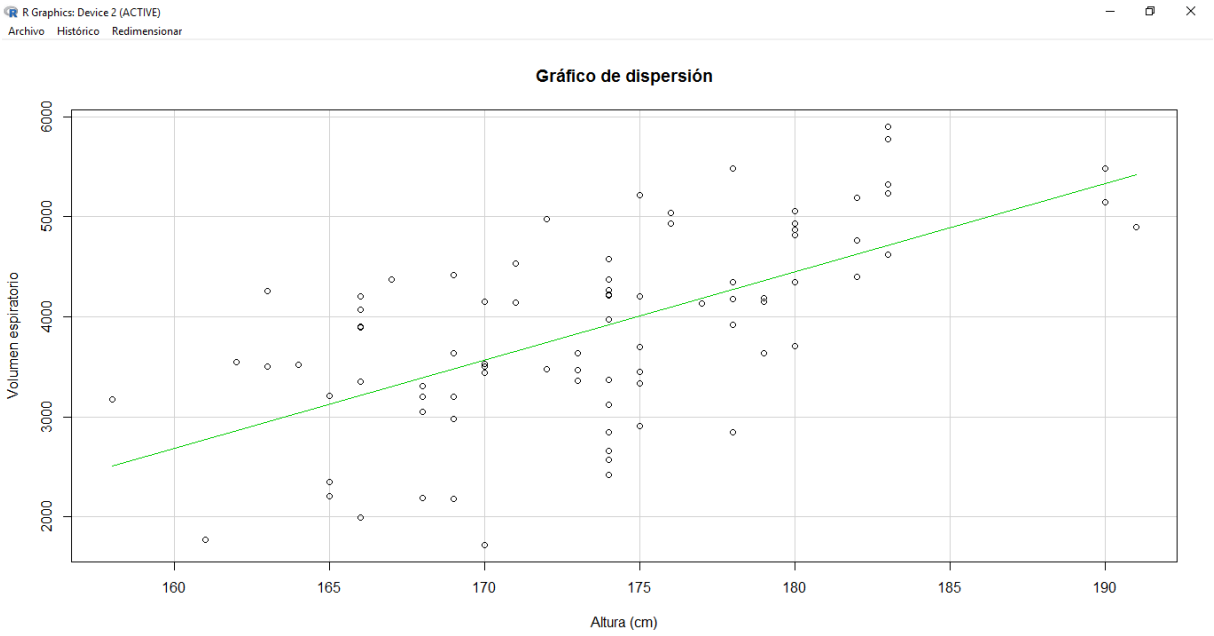


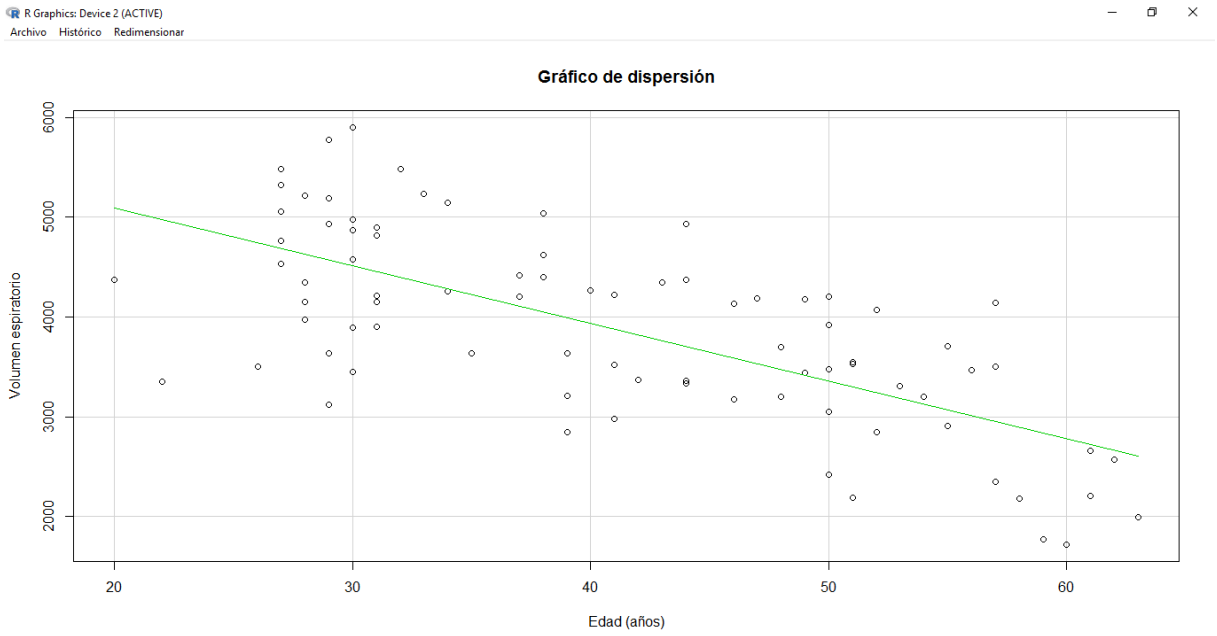
Tras pulsar el botón *Aceptar* aparecerá el gráfico de dispersión, también denominado *nube de puntos*.



Los puntos están distribuidos aproximadamente alrededor de una recta decreciente, por lo que la relación entre el tiempo de exposición y el volumen espiratorio es lineal indirecta. A medida que aumentan los años de exposición disminuye el volumen espiratorio.

El mismo tipo de gráfico puede realizarse para estudiar la relación de la altura y la edad con el volumen espiratorio, obteniendo lo siguiente:



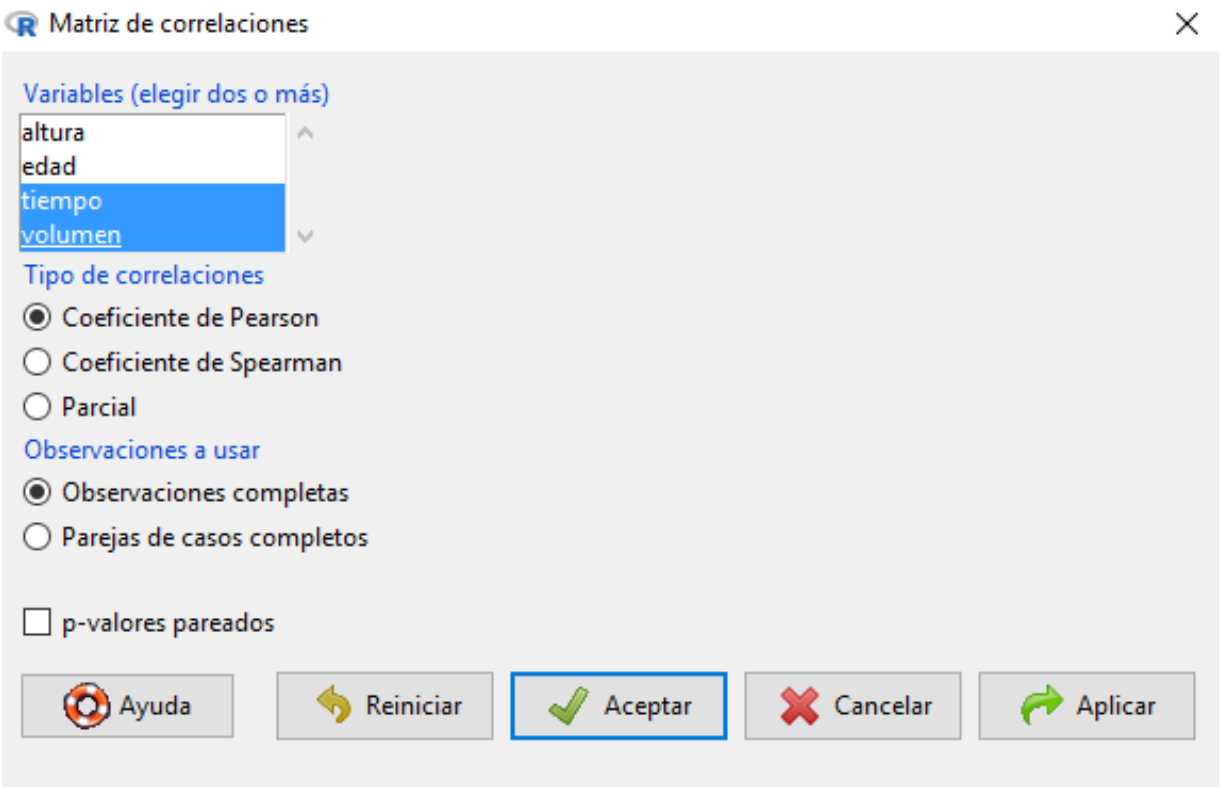


La relación entre la altura y el volumen espiratorio es lineal directa, de manera que el volumen es mayor en los sujetos de mayor altura. En cambio, la edad y el volumen espiratorio muestran una relación lineal indirecta, donde éste disminuye con la edad.

Una vez comprobada que la relación entre las variables independiente y dependiente es lineal, el coeficiente de correlación lineal de Pearson puede medir la fuerza de asociación entre ambas. Su valor absoluto está en un gradiente comprendido entre 0 y 1, donde 0 corresponde a la ausencia de relación lineal y 1 a una relación lineal perfecta. En esta última, todos los puntos estarán situados sobre la línea recta. El signo del coeficiente de correlación lineal de Pearson será negativo en una relación lineal indirecta y positivo en una directa. Para obtenerlo en *R-Commander* se seguirá la secuencia:

Estadísticos – Resúmenes – Matriz de correlaciones

En la ventana emergente se marcarán con el ratón la variable independiente y la dependiente dejando pulsada la tecla *Control* (*Ctrl*) del teclado. A continuación, en "*Tipo de correlaciones*" se seleccionará la opción "*Coeficiente de Pearson*".



Tras pulsar el botón aceptar se obtienen los siguientes valores en la ventana de resultados:

	tiempo	**volumen**
tiempo	1.00	-0.66
volumen	-0.66	1.00

El coeficiente de correlación lineal de Pearson entre *Tiempo* y *Volumen* es -0.66, con signo negativo por ser una relación lineal indirecta. Aunque no existe un consenso generalizado, la asociación suele considerarse débil cuando el valor absoluto del coeficiente de correlación sea inferior a 0.40, media cuando esté entre 0.40 y 0.80 y fuerte cuando sea superior a 0.80.

El mismo procedimiento puede seguirse para estimar el coeficiente de correlación lineal de Pearson entre *Altura* y *Volumen* y entre *Edad* y *Volumen*.

Presentación de resultados

Cuando se estudia la relación entre una variable dependiente cuantitativa (*Volumen*) y otra independiente cualitativa (*Tabaco*), los resultados suelen presentarse en una única tabla que muestra un resumen numérico de la variable dependiente para cada una de las categorías de la variable independiente. Los parámetros estadísticos habituales son el número de sujetos,

mínimo, máximo, media y desviación típica. En el caso práctico *Volumen espiratorio* la única variable independiente cualitativa es *Tabaco*. Si hubiese habido más variables independientes, los resultados aparecerían secuencialmente en una tabla como esta:

Variables	Sujetos	Mínimo	Máximo	Media	Desviación típica
Tabaco					
Nunca fumó	14	2350	5480	3977.86	933.00
Exfumador	21	1720	5900	4148.57	1037.32
Fuma actualmente	48	1770	5780	3736.67	914.62
Variable independiente 2					
Categoría A					
Categoría B					
:					
Variable independiente 3					
Categoría A					
Categoría B					
:					

A veces, el espacio asignado a documentos escritos, como artículos científicos o informes, es limitado. Por ello, los gráficos de cajas no suelen incorporarse cuando el número de variables independientes es elevado. Sin embargo, son útiles en presentaciones orales para destacar características de alguna de las variables.

Para mostrar los resultados del análisis bivariante de variables cuantitativas se utiliza el diagrama de dispersión junto al coeficiente de correlación lineal, siempre que la relación sea lineal. Si hay más de una variable independiente, la disposición de los gráficos se realiza de manera que el espacio quede lo más aprovechado posible. La disposición en cuadrículas de este tipo suele ser una opción frecuente:

Gráfico 1	Gráfico 2
Gráfico 3	Gráfico 4

Los coeficientes de correlación lineal pueden incorporarse al pie de cada gráfico o bien agruparse en una tabla similar a esta:

Correlación lineal del volumen espiratorio con el resto de variables independientes

Variable independiente	Coeficiente de correlación lineal de Pearson
Tiempo	-0.66
Altura	0.62
Edad	-0.69

COMENTARIOS ADICIONALES

Relaciones entre variables cualesquiera

Todo lo expuesto anteriormente está basado en la relación entre una variable dependiente y otra independiente, ya que es el tipo de asociación que se persigue en la mayoría de los objetivos de una investigación. Sin embargo, los mismos procedimientos sirven para describir la relación entre dos variables cualesquiera. Así, en el caso práctico *Accidentes por pinchazo en profesionales de enfermería*, las variables *Sexo* y *Edad* son independientes, pero también es posible describir su relación comparando la edad media de hombres y mujeres. De la misma forma, en el estudio *Volumen espiratorio en profesionales de la minería* se podría describir la relación entre las variables independientes *Tiempo* y *Edad* utilizando un diagrama de dispersión. Será el propio grupo de trabajo el que establezca en cada momento el objetivo del estudio y el interés por analizar determinadas relaciones, justificando siempre su decisión con un marco teórico previo.

Limitaciones del análisis descriptivo bivariante

La relación entre dos variables puede estar distorsionada por un tercer factor de confusión que el análisis descriptivo bivariante no puede controlar. Si esto ocurriera, la medida de asociación entre la variable dependiente e independiente podría estar sesgada, mostrando un efecto que realmente no existe o revelando una asociación real cuya magnitud podría estar atenuada o aumentada.[10] Por ello es necesario avanzar un poco más en el análisis de datos, utilizando modelos de regresión multivariante antes de llegar a una conclusión plausible sobre el problema de investigación.

[10] J. de Irala *et al*. ¿Qué es una variable de confusión? *Medicina Clínica (Barcelona)* 2001; 117: 337-385.

6

CASOS PRÁCTICOS

Los análisis estadísticos realizados en esta monografía están basados en varios casos prácticos cuyo contenido se describe a continuación.

ACCIDENTES POR PINCHAZO EN PROFESIONALES DE ENFERMERÍA

Los accidentes por pinchazo con aguja hipodérmica son un problema de salud importante en enfermería, tanto por el riesgo de contagio por VIH y otras enfermedades infecciosas como por las consecuencias psicológicas que conlleva. Las actividades formativas en medidas de prevención pueden contribuir a la reducción de este tipo de accidentes, aunque no todas han mostrado su utilidad. Con el fin de probar la eficacia de uno de estos programas de formación se diseñó un estudio experimental con dos grupos de profesionales: Uno de intervención y otro de control. Ambos grupos recibieron formación sobre cuestiones generales de enfermería. Sin embargo, sólo el primero recibió información específica sobre medidas preventivas dirigidas a evitar pinchazos accidentales. En el estudio participaron 25 profesionales de enfermería de un Centro de Salud. Cada uno de ellos fue asignado de forma aleatoria al grupo de intervención o al grupo control. Tras el periodo de formación se realizó un seguimiento de todos los profesionales durante 6 meses, observando si durante ese periodo se produjo algún accidente.

Hipótesis

La principal hipótesis de investigación era que el programa de formación es eficaz para disminuir los accidentes por pinchazo, de manera que la proporción de accidentes sería menor en el grupo de intervención que en el grupo control.

Objetivos

1. Describir las características de los profesionales que participaron en el estudio.
2. Evaluar la eficacia del programa de formación en la disminución de accidentes por pinchazo.
3. Estudiar la relación del sexo y la edad con los accidentes por pinchazo.

Variables

La información correspondiente a cada uno de los profesionales se recogió en una ficha individual con un código personal de identificación. En ella se registraron las siguientes características:

Código: Número de identificación del profesional

Grupo: Grupo al que fue asignado dentro del programa de formación específica
- 1 Formación
- 2 No formación

Estado: Estado al final del seguimiento
- 1 Accidentado
- 2 No accidentado

Edad: Edad del profesional en años

Sexo:
- 1 Hombre
- 2 Mujer

Base de datos

El archivo *Accidentes por pinchazo* contiene los datos de las personas que participaron en el estudio con la siguiente estructura:

Código	Grupo	Estado	Edad	Sexo
00004	Formación	No accidentado	45	Hombre
00006	No Formación	No accidentado	50	Hombre
00014	No Formación	No accidentado	55	Hombre
00015	Formación	No accidentado	26	Mujer
00018	Formación	No accidentado	58	Mujer
00019	Formación	No accidentado	21	Mujer
00022	Formación	No accidentado	52	Mujer
00024	Formación	No accidentado	51	Mujer
00001	Formación	Accidentado	22	Hombre
00002	No Formación	Accidentado	22	Hombre
00003	No Formación	Accidentado	22	Hombre
00005	Formación	Accidentado	30	Hombre
00007	Formación	Accidentado	34	Hombre
00008	Formación	Accidentado	23	Hombre
00009	No Formación	Accidentado	28	
00010	No Formación	Accidentado	21	Hombre
00011	No Formación	Accidentado	40	Hombre
00012	Formación	Accidentado	30	Hombre
00013	No Formación	Accidentado	35	Hombre
00016	No Formación	Accidentado		Mujer
00017	No Formación	Accidentado	50	Mujer
00020	No Formación	Accidentado	25	Mujer
00021	Formación	Accidentado	47	Mujer
00023	No Formación	Accidentado	23	Mujer
00025	No Formación	Accidentado	23	Mujer

Base de datos con información numérica y caracteres de texto

VOLUMEN ESPIRATORIO EN PROFESIONALES DE LA MINERÍA

El presente estudio fue diseñado para estudiar la función pulmonar de 83 sujetos expuestos a altos niveles de polvo en una mina.

Hipótesis

La hipótesis principal del estudio era que el tiempo de exposición al polvo, la edad y el tabaco son factores importantes que intervienen en la alteración del volumen espiratorio.

Objetivos

1. Describir las características de los sujetos de estudio.
2. Estudiar el efecto del tiempo de exposición al polvo de la mina sobre el volumen espiratorio.
3. Estudiar el efecto del tabaco, la edad y la altura sobre el volumen espiratorio.

Variables

La información correspondiente a cada uno de los profesionales se recogió en una ficha individual en la que se registraron las siguientes características:

Edad: Edad del trabajador en años

Altura: Altura del trabajador en centímetros

Tiempo: Años de exposición al polvo

Tabaco: Hábitos sobre el tabaco
- 1 Nunca fumó
- 2 Exfumador
- 3 Fuma actualmente

Volumen: Volumen espiratorio (ml/seg)

Base de datos

El fichero *Volumen espiratorio* contiene los datos de las personas que participaron en el estudio con la siguiente estructura:

Identificador	Edad	Altura	Tiempo	Tabaco	Volumen
1	50	172	34	Exfumador	3480
2	51	168	28	Exfumador	2190
3	54	169	31	Fuma actualmente	3200
4	41	174	2	Nunca fumó	4220
5	31	191	3	Fuma actualmente	4900
6	50	178	25	Fuma actualmente	3920
7	48	175	6	Fuma actualmente	3700
8	29	182	3	Fuma actualmente	5190
9	28	170	1	Exfumador	4150
10	44	174	3	Fuma actualmente	4370
11	30	183	0	Exfumador	5900
12	48	168	26	Nunca fumó	3200

Identificador	Edad	Altura	Tiempo	Tabaco	Volumen
13	28	174	4	Fuma actualmente	3970
14	29	174	3	Nunca fumó	3120
15	37	166	2	Fuma actualmente	4200
16	58	169	23	Fuma actualmente	2180
17	31	166	3	Fuma actualmente	3900
18	27	183	3	Fuma actualmente	5320
19	28	175	4	Fuma actualmente	5220
20	29	169	3	Nunca fumó	3640
21	30	166	8	Exfumador	3890
22	52	174	23	Fuma actualmente	2850
23	46	158	15	Fuma actualmente	3170
24	41	169	20	Fuma actualmente	2980
25	35	179	3	Fuma actualmente	3640
26	52	166	12	Exfumador	4070
27	39	178	22	Nunca fumó	2850
28	55	180	22	Nunca fumó	3710
29	49	170	32	Fuma actualmente	3440
30	20	167	0	Fuma actualmente	4370
31	29	180	3	Exfumador	4930
32	62	174	20	Exfumador	2570
33	29	183	6	Fuma actualmente	5780
34	26	170	3	Fuma actualmente	3500
35	41	164	15	Nunca fumó	3520
36	50	174	22	Fuma actualmente	2420
37	39	173	3	Fuma actualmente	3640
38	32	190	14	Exfumador	5480
39	57	163	28	Fuma actualmente	3500
40	38	183	4	Fuma actualmente	4620
41	53	168	22	Fuma actualmente	3310
42	55	175	34	Fuma actualmente	2910
43	44	175	24	Fuma actualmente	3330
44	51	162	23	Exfumador	3550
45	51	170	12	Fuma actualmente	3530
46	47	179	24	Exfumador	4190
47	50	175	31	Exfumador	4200
48	27	178	0	Nunca fumó	5480
49	37	169	2	Exfumador	4420
50	22	166	2	Fuma actualmente	3350
51	27	171	3	Nunca fumó	4530
52	43	178	21	Exfumador	4350
53	30	175	1	Fuma actualmente	3450
54	63	166	37	Fuma actualmente	1990
55	31	174	0	Fuma actualmente	4210
56	46	177	3	Fuma actualmente	4130
57	28	180	4	Fuma actualmente	4350
58	57	165	25	Nunca fumó	2350
59	27	180	1	Nunca fumó	5060
60	38	182	2	Exfumador	4400
61	30	180	7	Fuma actualmente	4870
62	40	174	2	Fuma actualmente	4270
63	31	179	4	Nunca fumó	4150
64	50	168	29	Fuma actualmente	3050
65	60	170	34	Exfumador	1720
66	33	183	3	Exfumador	5230
67	61	165	26	Fuma actualmente	2210
68	56	173	26	Fuma actualmente	3470

Identificador	Edad	Altura	Tiempo	Tabaco	Volumen
69	49	178	28	Exfumador	4180
70	42	174	24	Fuma actualmente	3370
71	31	180	4	Nunca fumó	4820
72	30	172	3	Fuma actualmente	4980
73	44	176	11	Exfumador	4930
74	27	182	1	Fuma actualmente	4760
75	38	176	3	Nunca fumó	5040
76	34	190	3	Exfumador	5150
77	39	165	5	Fuma actualmente	3210
78	34	163	2	Fuma actualmente	4260
79	44	173	7	Fuma actualmente	3360
80	57	171	28	Exfumador	4140
81	59	161	28	Fuma actualmente	1770
82	30	174	7	Fuma actualmente	4580
83	61	174	38	Fuma actualmente	2660

Bibliografía

Culebro M, Gómez WG, Torres S. Software libre vs software propietario: Ventajas y desventajas. México, 2006.

De Irala J, et al. ¿Qué es una variable de confusión? Medicina Clínica (Barcelona) 2001; 117: 337-385.

Fox J. The *R* Commander: A Basic-Statistics Graphical User Interface to R. Journal of Statistical Software 2005; 11(9): 1-42.

GNU Operating System. La definición de software libre. Disponible en: www.gnu.org/philosophy/free-sw.es.html

Gómez-Gómez M, Danglot-Banck C, Huerta-Alvarado SG, García de la Torre G. El estudio de casos y controles: su diseño, análisis e interpretación en investigación clínica. Revista Mexicana de Pediatría 2003; 70(5): 257-263.

Hornik K. The *R* FAQ. 2011. Disponible en: http://cran.r-project.org/doc/FAQ/R-FAQ.pdf

Lewis JA. Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline. Statistics in Medicine 1999; 18: 1903-1942.

Ripley BD, Murdoch. *R* for Windows FAQ (Version for R-2.13.0) [www.r-project.org]

The *R* Foundation for Statistical Computing. R: Regulatory Compliance and Validation Issues. A Guidance Document for the Use of *R* in Regulated Clinical Trial Environments. Viena: The *R* Foundation, 2008. Disponible en: http://www.r-project.org/doc/R-FDA.pdf

Descubriendo *R-Commander*

MONOGRAFÍAS EASP

R es un lenguaje de programación muy flexible orientado a la estadística computacional, el análisis de datos y el desarrollo de gráficos, características que lo ha convertido en un lenguaje muy popular entre estadísticos y matemáticos especializados en estadística computacional.

A pesar de sus cualidades técnicas, el uso de *R* puede resultar complejo para personas que no están familiarizadas con los lenguajes de programación. Por este motivo, John Fox, profesor de Sociología de la Universidad McMaster (Canadá), desarrolló en 2005 el paquete *Rcmdr*, una Interfaz Gráfica de Usuario denominada *R-Commander* que permite trabajar en un entorno de ventanas similar al de otros programas estadísticos como SPSS.

Durante los últimos años ha habido un interés creciente entre profesionales de Ciencias de la Salud por el uso del lenguaje de programación *R* y de la interfaz *R-Commander* en sus investigaciones, más debido al carácter gratuito de los mismos que a la necesidad de programar complejos algoritmos para el análisis estadístico de la información. En la actualidad existe una amplia bibliografía sobre el lenguaje de programación *R* y sus procedimientos para el análisis de datos. Sin embargo, la documentación sobre *R-Commander* es escasa, especialmente en lengua castellana.

El propósito de esta monografía es proporcionar una guía de ayuda sencilla para el análisis estadístico de datos mediante la interfaz *R-Commander*, dirigida a profesionales no especializados en Estadística que utilizan esta aplicación durante el desarrollo de actividades formativas básicas o de forma puntual en sus investigaciones. No se tratan, por tanto, cuestiones relacionadas con la programación en *R* o el uso de secuencias de comandos, cuyo abordaje requeriría conocimientos computacionales más avanzados y estaría orientado a especialistas que utilizan métodos estadísticos de forma intensiva en su labor profesional diaria.