Introducción a

para el Análisis de Datos en Ciencias Sociales

Eduardo Bologna

Colección Cartografías: Materiales para la investigación y el aprendizaje

Serie Cuadernos de Investigación Cuantitativa Nº 1

Bologna, Eduardo León
Introducción a R para el análisis de datos en Ciencias Sociales / Eduardo León Bologna. - 1a ed . - Ciudad Autónoma de Buenos Aires : CONICET - Consejo Nacional de Investigaciones Científicas y Técnicas , 2016.
Libro digital, PDF
Archivo Digital: descarga y online
ISBN 978-950-692-135-4
1. Metodología. 2. Análisis de Datos. I. Título.
CDD 300.1

Comité de referato para este número:

Dra. Silvina Brussino CIPSI, Grupo vinculado al CIECS (CONICET - UNC)

Dra. Patricia Caro Profesora Adjunta de Estadística Facultad de Ciencias Económicas UNC

Lic. Jorge Lorenzo Cátedra de Estadística Educativa UNC

Dra. Alicia Maccagno Programa de Estadísticas Universitarias UNC

Dr. Martín Saino Facultad de Ciencias Económicas UNC

Imagen de tapa: logo R, reproducido bajo licencia GNU General Public License version 2 (GPL-2).

Índice

Introducción

Actualmente el análisis de datos viene ganando importancia en diferentes áreas de conocimiento. Además de la investigación científica, está presente en estudios sociales, en los negocios, en el deporte. Los estudios de mercado, encuestas preelectorales o sondeos de opinión recurren al análisis de datos para obtener resultados que les sirvan para tomar decisiones. La disponibilidad cada vez mayor de bases de datos de gran tamaño exige herramientas estadísticas adecuadas para resumir información y poder extraer de ellas significado.

Los datos constituyen conjuntos de información expresada en lenguaje estandarizado, es decir que acerca de un conjunto de individuos (personas, países, instituciones, etc.) se conocen determinadas características, el análisis de los datos implica organizarlos de modo que se puedan dar respuestas a los problemas planteados o bien explorarlos para detectar tendencias y patrones.

Los procedimientos para el tratamiento de información sistematizada se perfeccionan continuamente, las técnicas se vuelven más sofisticadas para abordar problemas de mayor complejidad. Cada vez se otorga más importancia a la comunicación eficaz de los resultados a fin de incidir en la toma de decisiones en salud, educación, políticas públicas, en los negocios. Para ello, los análisis se enriquecen con expresiones gráficas novedosas que suman potencialidad heurística a las conclusiones que se alcanzan.

Para realizar estos análisis existen numerosos programas informáticos, que se ocupan de los procesos computacionales, de manera que el usuario solo deba decidir qué procedimiento aplicar y realizar una lectura correcta y completa del resultado que se obtiene, sin involucrarse con las operaciones de cálculo. Estos programas o "paquetes estadísticos" reúnen en un entorno único las operaciones más frecuentemente usadas por investigadores y analistas de datos y las ponen al alcance del usuario no especializado.

De entre las diversas opciones disponibles, en este curso se usará un software que se llama R, a través de una interfaz gráfica de usuario, amigable: R-commander. Por este medio, el investigador que se inicia en el análisis de datos o busca solo una aplicación concreta puede ingresar de manera gradual a la programación, para usarla en operaciones más avanzadas. El usuario experimentado hallará en R una herramienta muy versátil para diversos tipos de análisis y de gran potencialidad gráfica.

A fin de ejemplificar la aplicación de los procedimientos, se usan datos de la Encuesta Nacional de Factores de Riesgo 2013, realizada conjuntamente entre el Instituto Nacional de Estadística y Censos (INDEC) y el Ministerio de Salud de la Nación de la República Argentina.

¿Qué es R?

Algunas característica de R son mencionadas en el sitio de la comunidad inside-R (Analytics, 2015), allí cuentan que:

R (R Team Core, 2015) es un software para análisis de datos: lo usan estadísticos y analistas de datos para extraer significado de información cuantitativa, descripciones e inferencias, visualización de datos y modelización predictiva.

Es un lenguaje de programación orientado a objetos, diseñado por estadísticos y para el uso de estadísticos: el análisis se hace escribiendo sentencias en este lenguaje, que provee objetos, operadores y funciones que hacen muy intuitivo el proceso de explorar, modelar y visualizar datos.

Es un ambiente para el análisis estadístico: en R hay funciones para prácticamente todo tipo de manejo de datos, modelización y representaciones gráficas que pueden hacer falta. No solo cuenta con los métodos estándar sino que, debido a que los principales avances en procedimientos estadísticos se realizan en R, las técnicas más actualizadas están usualmente primero disponibles en R. R integra programas llamados paquetes, que sirven para realizar análisis específicos. Los paquetes son rutinas que realizan conjuntos de operaciones especializadas y una de las potencialidades de R es que diferentes investigadores pueden desarrollar paquetes para determinados tipos de análisis y ponerlos a disposición de los demás usuarios. En la actualidad hay más de 7100 paquetes y el conjunto crece porque la comunidad R es muy activa y continuamente se hacen aportes.

Es un proyecto de código abierto: esto significa no solo que se lo puede descargar y usar gratis, sino que el código es abierto y cualquiera puede inspeccionar o modificar las rutinas. Como sucede con otros proyectos de código abierto, como Linux, R ha mejorado sus códigos tras varios años de "muchos ojos mirando" y aportando soluciones. También como otros proyectos de código abierto, R tiene interfaces abiertas, por lo que se integra fácilmente a otras aplicaciones y sistemas.

Es una comunidad: R fue inicialmente desarrollado por Robert Gentleman y Ross Ihaka[1], del Departamento de Estadística de la Universidad de Auckland, en 1993 y desde entonces el grupo que dirige el proyecto ha crecido hasta llegar a tener actualmente más de 20 estadísticos y analistas de computación de todo el mundo. Además, miles de otras personas han contribuido con funcionalidades adicionales por medio del aporte de "paquetes" que utilizan los 2 millones de usuarios de todo el mundo. Como resultado existe una intensa comunidad de usuarios de R on-line, con muchos sitios que ofrecen recursos para principiantes y para expertos.

¿Por qué elegir R?

Considerando que en el mercado existen muchos programas para hacer análisis de datos, conviene explicar lo que hace que R sea diferente.

• Es gratis y abierto, no se pagan licencias y si se cambia de trabajo no hace falta aprender a usar un nuevo software. Se distribuye con licencia GNU GPL (General Public License: `http://www.gnu.org/licenses/gpl.html`), puede ser copiado sin ningún

[1] R&R, por los nombres de sus autores dio origen a R como denominación del lenguaje.

inconveniente. Los códigos pueden editarse.

• Lleva más de 20 años de desarrollo y hay una gran comunidad de usuarios que aporta continuamente desarrollos y soluciones.

• Hay redes de usuarios y foros a donde se puede recurrir para consultas, salvar dificultades, son muy activas y dispuestas a ayudar.

• Puede usarse en sistemas operativos Windows, Mac y Linux

• Existen interfaces de usuario gráficas que facilitan su uso.

• Tiene alta capacidad gráfica. Es un principio en el diseño de R que la visualización de los datos constituye una parte importante del análisis, por ello dispone de conjuntos de herramientas que permiten la creación de una variedad de gráficos. El diseño de los sistemas de representación gráfica de R ha sido influido por los trabajos de importantes referentes de visualización de datos, como Cleveland & McGill (1984), Cleveland (1993), Tufte (2003). Actualmente, los principales medios de comunicación usan R para expresar datos de manera gráfica.

• Da acceso a las técnicas de análisis más recientemente desarrolladas, porque los principales académicos e investigadores usan R para construir nuevos procedimientos estadísticos.

Organización del material

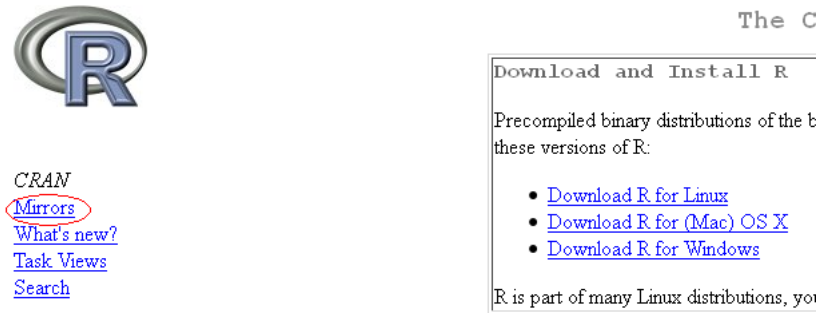
Este texto está previsto para ser usado como soporte de un curso sobre el tema o bien de manera autónoma. Cualquiera sea el caso, su lectura supone que simultáneamente se tenga abierto R y R Commander, a fin de poner a prueba los procedimientos que se muestran. Además del contenido conceptual y práctico, se ofrece un anexo llamado "Visitando la sintaxis" dirigido a quienes estén interesados en conocer muy inicialmente el modo de operación de R cuando no se tienen ventanas y botones.

Se incluye al final una selección arbitraria de paquetes, de los más de siete mil disponibles en el repositorio CRAN. La lista refiere paquetes para: inferencia bayesiana, análisis demográfico y epidemiológico, análisis de ítems bajo el modelo TRI, comparación de grupos —útil para investigación sobre evaluación de impacto—, representaciones gráfica y cartográficas. Es solo una muestra para estimular la búsqueda de opciones adecuadas al propio campo de investigación.

Los ejercicios están pensados para invitar al lector explorar otras posibilidades, en particular, a experimentar con sus propios datos.

Descarga e instalación

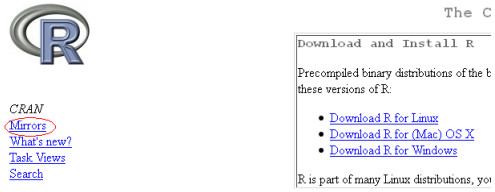
R se descarga gratuitamente de la "Red Integral de Archivos R", o CRAN en `https://cran.r-project.org/`.



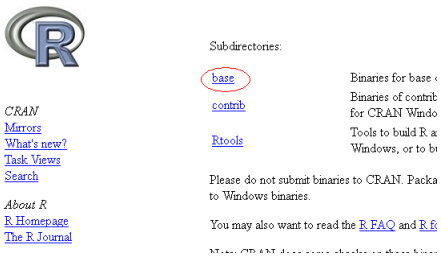
Una vez en esa página, se elige el sitio espejo desde donde se realizar la descarga, el de Argentina (Universidad Nacional de La Plata) es aconsejable, o también 0-Cloud.

.

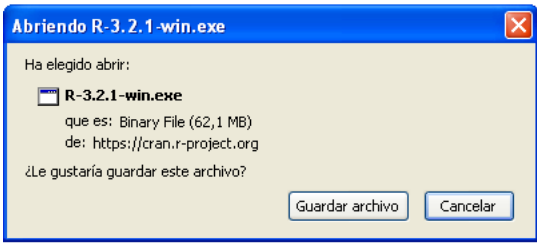
Luego se selecciona el sistema operativo:



Sigue la elección de "base" porque es la primera instalación de R:



Y se inicia la descarga



El proceso siguiente es automático y solicita alguna confirmaciones. Luego de esto, R está instalado y en condiciones de ser utilizado.

Apertura

Con un doble clic sobre el icono que se ubica en el escritorio luego de la instalación, se abre R. Aparece una ventana denominada R Gui (Graphical user interface, Interfaz gráfica de usuario), que tiene dentro otra ventana: la consola de R. Allí están los créditos y algunas indicaciones sobre la versión, además hay un "$>$" rojo, se llama prompt e indica que R está listo para recibir un comando. Si se fuera a usar así, se necesitaría escribir una línea de comando para que realice alguna operación, todo lo que aquí se escriba distingue mayúsculas y minúsculas.

El ingreso a la ayuda se logra con:

- help()

La instalación de origen tiene el paquete *base*. Para conocer lo que contiene se solicita

- help(base)

Se obtiene poca información, pero indica que para conocer las funciones disponibles debe solicitarse

- library(help = "base")

Cuyo resultado es el listado de operaciones que contiene el paquete.

Al momento de usar R para publicaciones debe citarse del modo en que los autores lo sugieren, para saberlo, se digita *citation()* en la consola y se obtiene la expresión de la cita y también la sintaxis en formato bibtex para quienes escriben en Latex.

El comando installed.packages() muestra un listado de paquetes ya instalados en R.

Los objetos en R

Las entidades que R crea y manipula se denominan *objetos*. Pueden ser variables, arreglos numéricos, textuales o lógicos, funciones, y también estructuras más generales que R crea a partir de ellos. Todos los objetos creados durante una sesión pueden ser guardados de manera permanente en un archivo para uso futuro. Los siguientes son algunos de los más frecuentemente usados:

- Vectores (vector): es la entidad más simple y consiste en un conjunto ordenado de elementos de un mismo tipo: lógico, entero, real complejo, de caracteres
- Matrices (matrix): es un arreglo bidimensional, en filas y columnas. Todos los elementos deben ser del mismo tipo.
- Listas (list): es una forma generalizada de vector, en la que los elementos no tienen que ser del mismo tipo. Es el objeto que contiene a los resultados de un análisis
- Funciones (function): relaciones entre otros objetos, que se almacenan en el espacio de trabajo
- Conjunto de datos o matriz de datos (data frame): un arreglo bidimensional cuyas columnas son vectores, que pueden ser de diferente tipo.
- Para conocer de qué tipo es un objeto, se usa el comando class(nombre.del.objeto).

Si se interesa por conocer un poco sobre el modo de operación por línea de comando, el Anexo provee una breve introducción para usar R sin interfaz gráfica; es decir, escribiendo las instrucciones directamente en la consola.

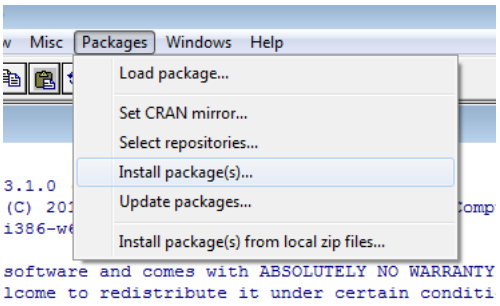
R Commander (Rcmdr)

R Commander (Fox, 2005) es un conjunto de paquetes disponibles en R, que transforma muchas de las instrucciones que se escribirían en la consola, en opciones de menúes, por lo que R se operará de modo muy similar a otros programas de análisis de datos: SPSS, SAS, INFOSTAT, STATA, etc. Técnicamente, R Commander provee una Interfaz Gráfica de Usuario ("GUI") para R. Entre las operaciones que pueden hacerse con R Commander se destacan:

- introducción manual de datos
- importación de datos en diferentes formatos
- gráficos de datos
- análisis descriptivo
- análisis inferencial

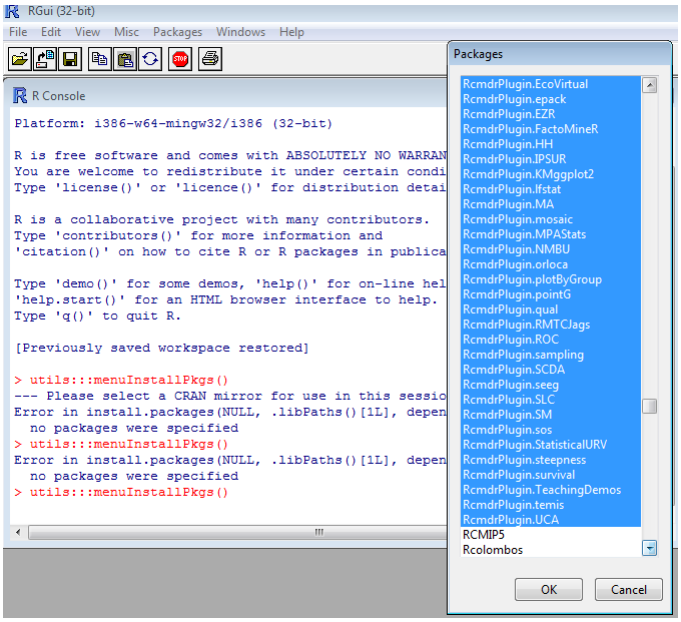
Instalación del paquete R Commander

En la ventana de R puede verse la opción *Paquetes* y dentro de ella *Instalar paquetes*.



Cuando esto se solicita, se abre una ventana para elegir desde que sitio espejo se hará la descarga, el de Argentina (La Plata) es adecuado, si no está disponible, la nube (cloud) también funciona. A continuación, una nueva ventana mostrará el conjunto de paquetes que están disponibles en la CRAN.

Los paquetes están ordenados alfabéticamente, se busca el que se llama *Rcmdr*, se lo selecciona y con él a todos los que siguen hasta *RcmdrPlugin.UCA* (usando Shift + ↓).



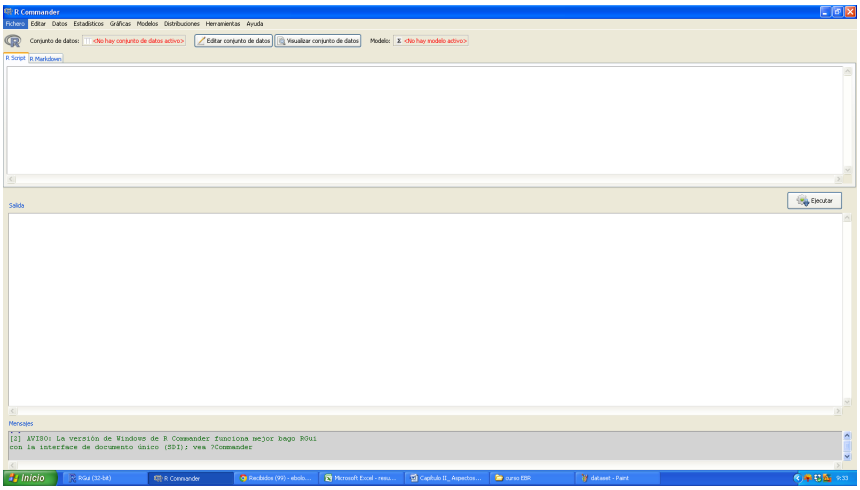
Con todos ellos seleccionados, se elige OK para instalarlos. Este proceso es largo, dependiendo de la velocidad de la conexión, cuando se completa, vuelve a aparecer el prompt rojo en la consola.

Carga de R Commander

La operación anterior se realiza una sola vez en la computadora, de ahora en más el conjunto de paquetes está instalado. Sin embargo, cada vez que se use será necesario *cargarlo*, es decir tenerlo activo en la sesión. Para cargar R Commander, simplemente se escribe en la consola

require(Rcmdr) <Enter>

Se recibe una advertencia que indica que otros plugins necesitan ser también cargados, se debe aceptar en las dos ventanas siguientes y comienza el proceso de carga. La próxima vez que se requiera no sucederá esto y se abrirá directamente. Cuando se completa, se abre una ventana diferente:



Las tres ventanas de R Commander son: *R Script*, *Salida* y *Mensajes*. Para poder empezar a ver el uso de cada una de ellas, debe contarse con un conjunto de datos activo.

Ejercicio 1

Instale R y R Commander en su equipo.

Los datos

El objetivo de esta introducción es usar R para analizar datos, por lo que se necesita primero disponer de un conjunto de ellos. La forma que tienen los datos que provienen de estudios cuantitativos es tal que se releva la misma información para diferentes individuos; es decir que se observan los mismos aspectos en todos los casos. Por ejemplo, si se trata de analizar los resultados de un examen, éstos no existen en el vacío, los resultados del examen pertenecen a alumnos que los rindieron; de cada uno de esos alumnos, interesa observar sólo la nota que obtuvieron. Pero si el interés es además comparar los resultados que obtienen quienes cursan por la mañana y quienes lo hacen por la tarde, entonces de cada alumno se necesita conocer también en qué horario cursa. En ese sentido se entiende "relevar la misma información", a todos los alumnos se les pregunta lo mismo: su nota y el turno en que cursan. Sea la siguiente tabla un conjunto ficticio de datos, correspondiente a 10 alumnos que rindieron un examen:

Datos ficticios sobre la nota en un examen y el turno en que cursan 10 alumnos.

caso	alumno	nota	turno
1	Daniel	8	M
2	Marcos	7	M
3	Susana	7	T
4	Ximena	8	T
5	Laura	8	T
6	Matías	5	T
7	Marta	7	M
8	Susana	4	T
9	Carlos	9	T
10	Ulises	6	T

Esta es la forma habitual de organizar la información cuantitativa, aquí hay filas (las horizontales) y columnas (las verticales). La primera fila se distingue del resto, porque son rótulos, son los nombres que indican lo que hay en cada columna: caso, alumno, nota, turno. La primera columna se llama caso y es un número correlativo asignado a cada alumno, ese número es único y así se distingue entre dos alumnos que tengan el mismo nombre. Además, ese número permite no poner el nombre y respetar el anonimato de la información. La segunda columna es el nombre de pila del alumno, en este caso no tiene ninguna utilidad, solo está allí para transmitir la idea que cada fila informa acerca de un individuo. La tercera columna indica la nota que cada uno obtuvo en el examen y la cuarta el turno al que asiste. Esta información puede estar impresa en un papel o en una hoja de cálculo o también en un documento de texto, sin embargo para el uso que aquí interesa, se la requiere cargada en el programa con que serán analizados los datos.

Convenciones de vocabulario

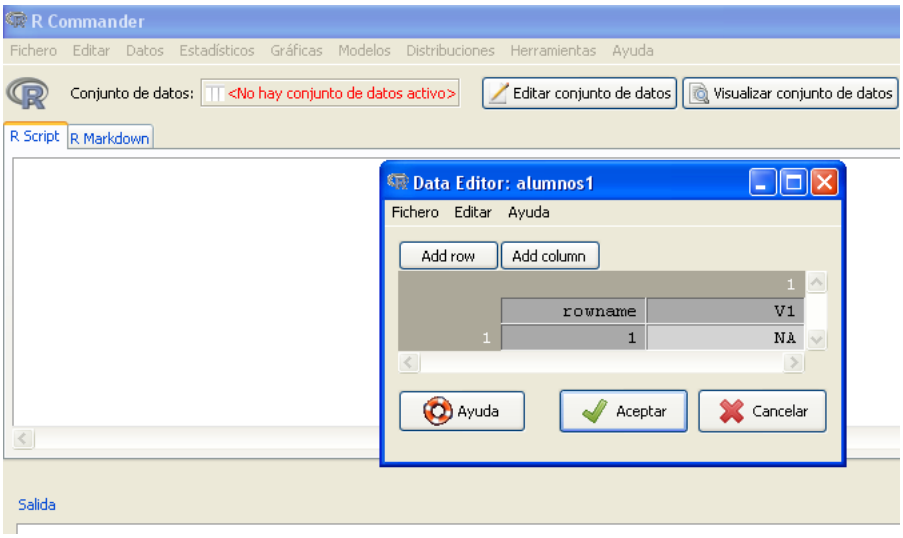
Una *matriz de datos* o *conjunto de datos* o *tabla de datos* es una organización en filas y columnas en la que cada fila corresponde a un individuo y cada columna a un aspecto.
Los individuos cuya información se registra se llaman *unidades de análisis.*
Los aspectos que se registran de cada individuo se llaman *variables.*
La característica de cada individuo en cada aspecto se llama *valor* o *categoría de la variable.*

Hay tres formas para disponer de una matriz de datos en R:

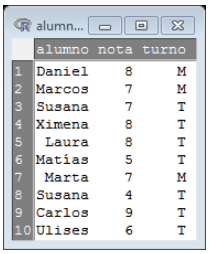
- digitandola de manera manual
- cargándola si está en formato legible
- importándola desde otros formatos

Digitación manual

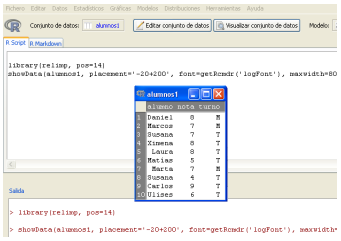
Cuando los datos no están cargados, sino que se cuenta con planillas impresas o bien con cuestionarios que han sido rellenados en papel, entonces se realiza la carga de manera manual. En la barra de herramientas, *datos* → *nuevo conjunto de datos...* y se debe definir el nombre de esta matriz de datos (en este campo no se aceptan espacios). En este ejemplo, la matriz se llamará *alumnos1*. Se abre una pequeña ventana que constituye el editor de datos.



Las opciones *add row* (agregar fila) y *add column* (agregar columna) permiten incluir tantas filas como casos haya y tantas columnas como variables. Para el ejemplo anterior, se agregan 10 filas y 3 columnas, no es necesaria la columna caso, porque ya esta prefijada bajo el rótulo *rowname*. A continuación se colocan los nombres de las variables y se cargan los datos, al completar la operación se elije *OK*. Luego se solicita ver la matriz con el comando *visualizar conjunto de datos*, para obtener:



Por su parte, la ventana principal de R Commander se divide en tres:



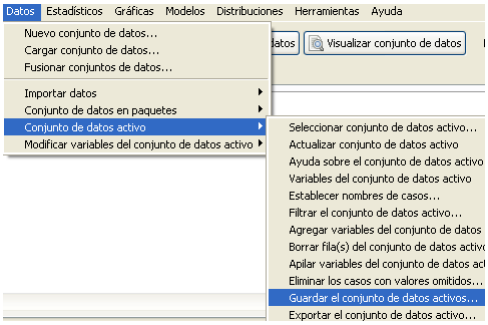
R Script contiene las líneas de comando que corresponden a las instrucciones que han sido introducidas por medio de los botones. De esto se trata: R Commander transforma las operaciones introducidas desde los menúes, en líneas de comando de R. Cuando se escriben instrucciones allí, se las ejecuta con la combinación Ctrl+R o bien con el botón *ejecutar*; si se las pide desde los botones, se ejecutan inmediatamente[2].

Salida reproduce la línea de comando cuando ésta es ejecutada y muestra los resultados de algunas operaciones, por ejemplo cuando se soliciten resúmenes de datos.

Mensajes información que puede servir al usuario, en este caso menciona cuántas filas y columnas tiene el archivo de datos generado. También es el lugar donde aparecen los mensajes de error.

Guardar los datos

Los datos que han sido cargados deben guardarse para tenerlos disponibles cuando vuelva a abrirse el programa. Para ello:



El formato en que se guarda es *.RData.

[2]Existen operaciones que no están disponibles desde los botones. En esos casos será necesario escribir aquí la sintaxis, en el Anexo se ofrece una breve introducción a ella.

Carga de datos en formato RData

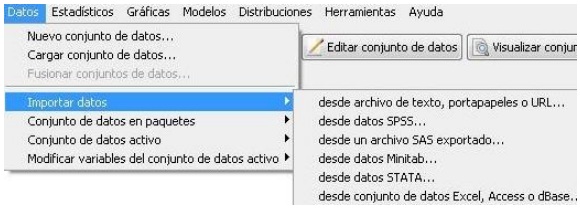
Un archivo guardado con el formato propio de R se abre simplemente desde *datos→cargar conjunto de datos*. Aparecerá así en el botón *conjunto de datos*, el nombre de la matriz cargada. Al hacer doble clic en un archivo RData no se abre R Commander, sino la consola de R, desde donde debe solicitarse R Commander (require(Rcmdr)) y luego elegir el conjunto de datos para tenerlo activo en la sesión.

Ejercicio 2

Cargue los datos de la tabla mostrada y guárdelos en formato RData. Cierre R Commander y R y abra nuevamente el archivo *alumnos 1.RData*.
Digite *class(alumnos1)* para ver el tipo de objeto de que se trata.
Solicite *names(alumnos1)*.

Importación desde otros formatos

Cuando se dispone de un conjunto de datos en alguno de los formatos más usados, como xls, xlsx, sav, txt, csv, dbf, es necesario importarlo para que R los pueda manejar. En la barra de herramientas de R Commander, se usa la combinación *Datos→Importar datos* y se elige el formato en que se encuentran los datos originales



Luego debe indicarse el nombre que tendrá este conjunto de datos, que puede ser el mismo que tiene en el otro formato o bien uno diferente, y, según el formato de origen que se haya elegido habrá que especificar opciones.

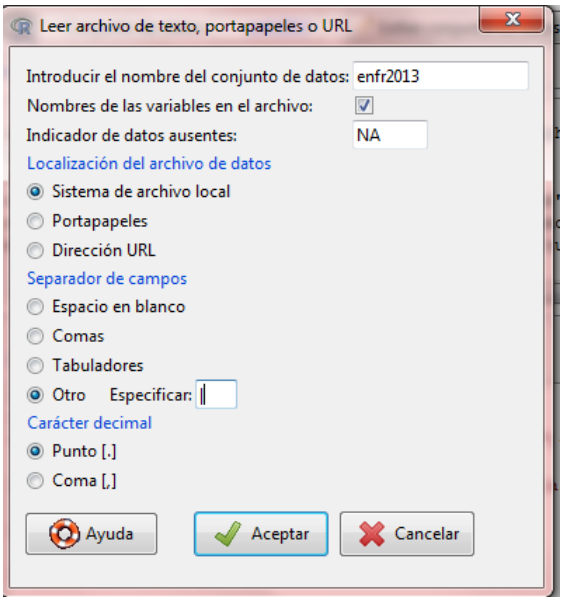
- desde texto, portapapeles o URL: la ubicación del archivo, la codificación de los casos perdidos, el carácter que se usa para separar campos y el que indica decimales (punto o coma).
- desde SPSS o STATA: la conversión de las etiquetas de las variables en factores (categorías de una variable nominal), el formato de fechas (solo desde STATA) y si se escriben los nombres de las variables en minúsculas (solo desde SPSS).
- desde SAS: la importación es directa
- desde Excel o Minitab: solo el nuevo nombre del archivo

Aplicación a datos reales

Para los ejemplos y ejercicios, se trabajará con base de la Encuesta Nacional de Factores de Riesgo realizada por el INDEC y el Ministerio de Salud en 2013 (INDEC, 2015). La base de microdatos está disponible en `http://www.indec.gov.ar/ftp/cuadros/menusuperior/enfr/ENFR2013_baseusuario.rar`. Además, las especificaciones necesarias para la importación y su uso están en un documento metodológico disponible en `http://www.indec.gov.ar/ftp/cuadros/menusuperior/enfr/doc_base_usuario_enfr2013.pdf`; allí se indica los siguiente:

Tipo de archivo: texto plano.
Delimitador: "|" (pipe, barra vertical, ASCII 124).
Calificador de texto: comilla doble (ASCII 34).
Encabezado en la primera línea: sí.
Codificación: UTF-8.
Salto de línea: Windows (CRLF).

Con esto es suficiente para importar el archivo a R Commander, se usará como nombre para el archivo importado *enfr2013*.



Cuando los datos estén cargados, aparecerá su dimensión en la ventana de mensajes.

Exploración de la matriz de datos

El documento *doc_base_usuario_enfr2013* explica la metodología de la encuesta, y en particular ofrece un listado de las variables que tiene la base, así como sus categorías. Ahora que está cargada, puede hacerse una visualización de la matriz de datos, por medio del comando *ver datos*. La visualización sucede con el formato de una hoja de cálculo. También se admite la edición de los datos, pero ésto tiene riesgo para la integridad de la información, por lo que no es aconsejable.

Ejercicio 3

Baje el archivo comprimido ENFR2013_baseusuario.rar, guárdelo y luego descomprima su contenido.
Abra el archivo de texto.
Obtenga el documento llamado *doc_base_usuario_enfr2013* de `http://www.indec.gov.ar/ftp/cuadros/menusuperior/enfr/doc_base_usuario_enfr2013.pdf` y guárdelo.
Importe la base desde R Commander.
Verifique en la ventana de mensajes cuántas filas y columnas tiene el conjunto de datos.
Guarde la base con el nombre enfr2013.

Modificación de variables

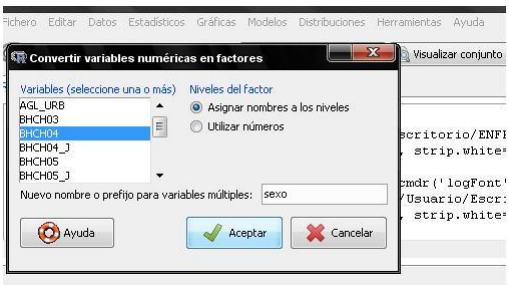
Antes de iniciar un proyecto de análisis de datos es necesario familiarizarse con las variables presentes en el conjunto que se cargó y eventualmente hacer algunos cambios. En esta sección se muestran algunos de ellos: la transformación de una variable cuantitativa en cualitativa (o nominal), la definición de nuevas variables, y la segmentación de variables cuantitativas de manera manual y automática.

Transformar en factor

La lectura automática que hace el programa consiste en considerar como numéricas a todas las variables que tienen números como categorías. Para lograr, por ejemplo, distribuciones de frecuencia es necesario indicar explícitamente cuáles variables son nominales (factores). Para ello, *datos* → *modificar variables del conjunto de datos activo* → *convertir variables numéricas en factores* permite que se de un nuevo nombre a la variable categorizada y nuevos nombres a sus categorías.

Cuando de una variable cuantitativa u ordinal que no tenga muchas categorías se quiere tener distribuciones de frecuencia y también medidas descriptivas, entonces será necesario generar una segunda variable. Si la que está cargada se llama *variable*, se denomina *variable.f* a la que tiene a sus categorías con nombres. El comando de conversión en factores ofrece la posibilidad de mantener los números como nombres de las categorías o bien de asignar nuevos nombres. Si se elije que sigan siendo caracteres numéricos, para las operaciones serán considerados como texto, es decir admitirá distribuciones de frecuencia, gráficos de barras, etc.

En la base enfr2013, la variable BHCH04 corresponde al sexo del entrevistado, con categorías 1 y 2 asignadas a varones y mujeres respectivamente, Dado que está cargada con los números 1 y 2, ha sido tomada por el sistema como cuantitativa y no será posible solicitar una distribución de frecuencias. Para corregir esto, se usa la transformación a factor de BHCH04:



Y, en la ventana siguiente indicamos las etiquetas para cada valor:

Ahora la ventana de mensajes informa que el conjunto de datos tiene una columna más, que corresponde a la variable que acaba de crearse.

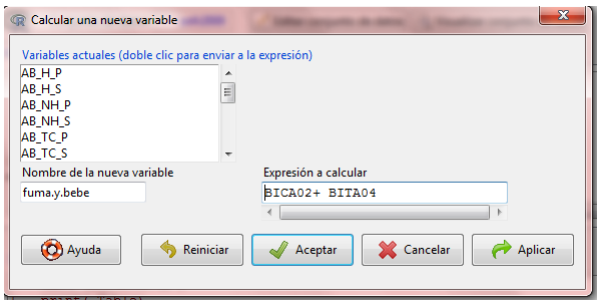
Definir una variable nueva

Una variable nueva resulta de modificaciones introducidas a las existentes o combinaciones de ellas. Por ejemplo, pasar la talla de centímetros a metros, requiere definir una variable nueva y también la generación de un puntaje de síntesis en una prueba, que resulta de la suma de puntos de cada una de una serie de preguntas. El ingreso total de una persona se compone también de la suma de lo ingresos provenientes de distintas fuentes. Como ilustración del procedimiento, se construirá a continuación una nueva variable en la base enfr2013, que permitirá distinguir a quienes fuman actualmente y además han consumido bebidas alcohólicas en los últimos 30 días, de quienes no tiene estas dos conducta simultáneamente (aquellos que no fuman o no han bebido en el último mes, o bien no manifiestan ninguna de las dos conductas). Las variables que se usan son:

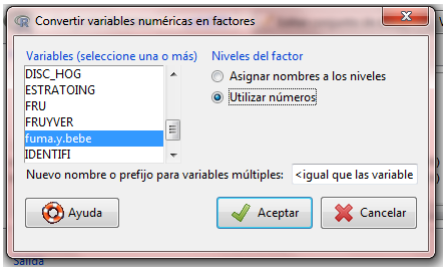
BICA02 ¿Cuándo fue la última vez que tomó alguna de estas bebidas alcohólicas?
1 Durante los últimos 30 días
2 Hace más de un mes, pero menos de un año
3 Hace más de un año
9 Ns/nc

BITA04 Actualmente ¿fuma usted cigarrillos...
1 ...todos los días?
2 ...algunos días?
3 ...no fuma?

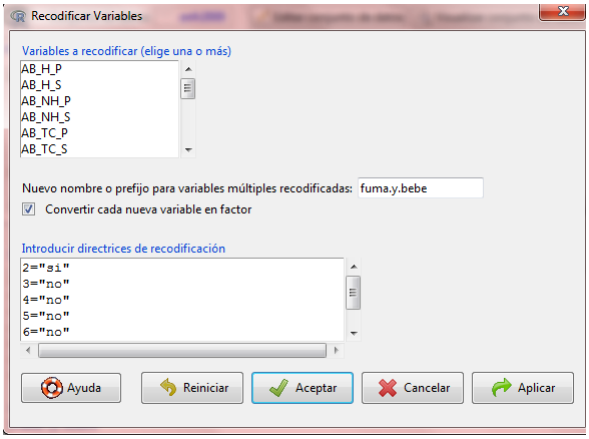
En primer lugar se define la variable fuma.y.bebe como la suma de los valores de estas dos, *datos → modificar variables del conjunto de datos activo→ Calcular una nueva variable*:



Luego se la definirá como factor, manteniendo los números como categorías, *datos → modificar variables del conjunto de datos activo→ Convertir variables numéricas en factores*:



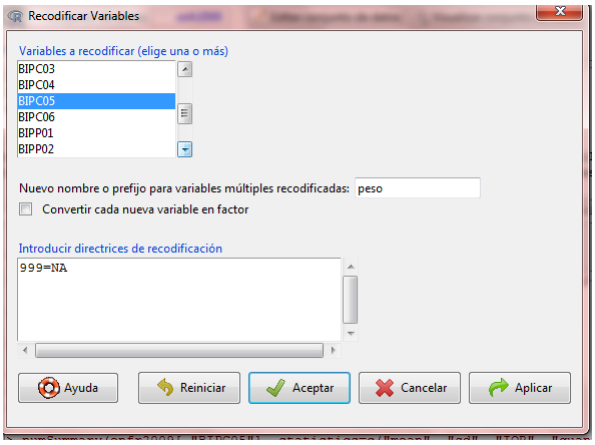
De los números que resultan de las sumas de todas las combinaciones posibles de los valores de BICA02 y BITA04, interesa distinguir el 2, porque resulta de haber respondido que se consumió alcohol en los últimos 30 días y que además se fuma actualmente. Se recodifica esta variable asignando si a la categoría 2 y no al resto, *datos* → *modificar variables del conjunto de datos activo*→ *recodificar variables*:



Si la variable es cuantitativa pueden ser necesarios otros recaudos. Para ilustrarlo, se construye la variable Índice de masa corporal, según la definición:

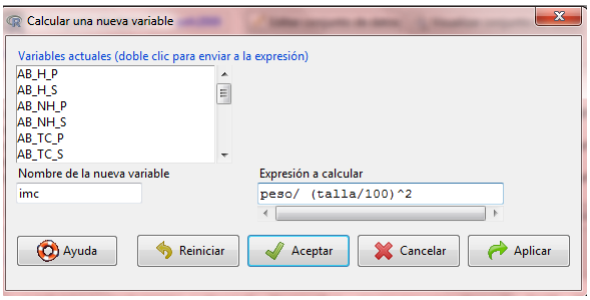
$$IMC = \frac{peso}{talla^2}$$

Donde el peso está medido en kilogramos y la talla en metros. Las variables numéricas BIPC05 y BIPC06 registran estas dos medidas, pero en el diseño de la ENFR se ha usado el código 999 para la no respuesta. Dado que este valor numérico no corresponde a un peso o tallas observables, será necesario indicar que son "valores perdidos". Se recodifican esta variables, con la sola instrucción de dar al 999 por perdido:



Y del mismo modo para BIPC06 a la que se renombra como *talla*. Con el cuidado de destildar la opción por defecto "Convertir cada nueva variable en factor", dado que se quiere conservar el carácter numérico del peso y la talla, para luego operar con ellas.

Solo ahora es posible definir el IMC:



Donde la talla ha sido dividida por 100 antes de elevar al cuadrado porque la variable original la expresa en centímetros.

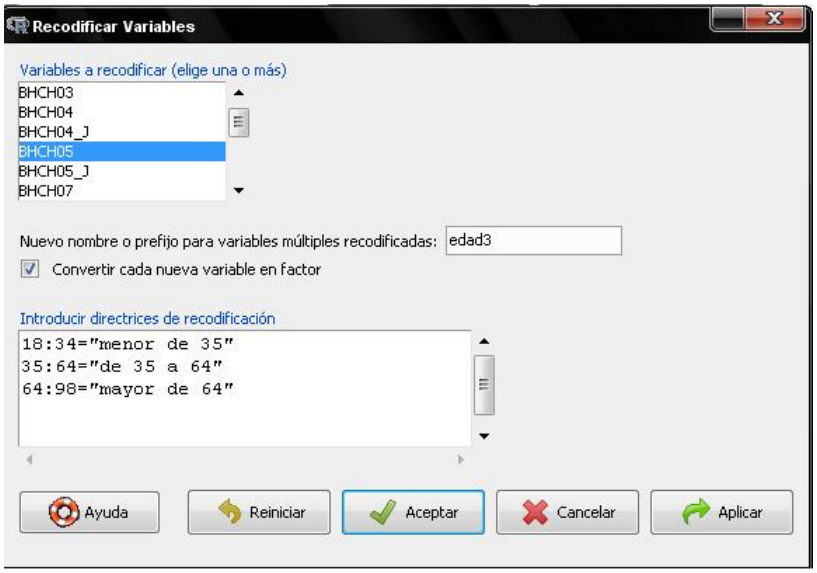
Categorización manual

La categorización de una variable cuantitativa se utiliza para construir grupos de valores y trabajar con esos grupos como si fueran categorías de un factor, por ejemplo, para construir una tabla de contingencia.

Como puede verse en el documento metodológico de la ENFR, la base ya trae una categorización de edad, que es la variable que se llama RANGEDAD, que ha sido construida con categorías:

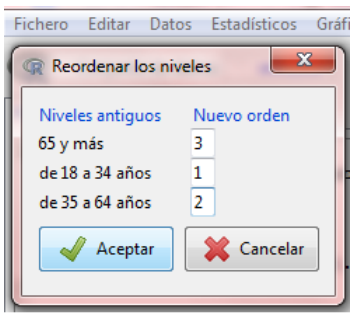
edad (años)	categorías de RANGEDAD
18 a 24	1
25 a 34	2
35 a 49	3
50 a 64	4
65 años y más	5

Pero para un análisis determinado, puede suceder que no resulten adecuadas y que sea necesario contar con grupos más amplios, por ejemplo, solo tres grupos así conformados: hasta 34 años, entre 35 y 49, 50 y más. A estos grupos de edades los construimos por medio de *datos* → *modificar variables del conjunto de datos activo*→ *recodificar variables*:



La notación 18:34 indica la secuencia de valores desde 18 hasta 34. Los nombres de las categorías van entrecomillados.

La ventana de mensajes indica ahora que la base tiene una variable (columna) más que antes. La variable que resulta de la recodificación es un factor y, según lo interpreta R, el orden de sus categorías (o niveles) no importa, por lo tanto, usa orden alfabético, como podría haber usado cualquier otro arbitrariamente. Pero las categorías de edad3 están originalmente ordenadas, y es necesario respetar ese orden para que aparezcan en las tablas primero la más baja y luego las siguientes en orden creciente. Para ello es necesario que se reordenen los niveles del factor: *datos* → *modificar variables del conjunto de datos activo*→ *reordenar niveles de factor*. Se elige la variable, conservando su nombre, por lo que se sobreescribirá sobre la existente y en la ventana siguiente:



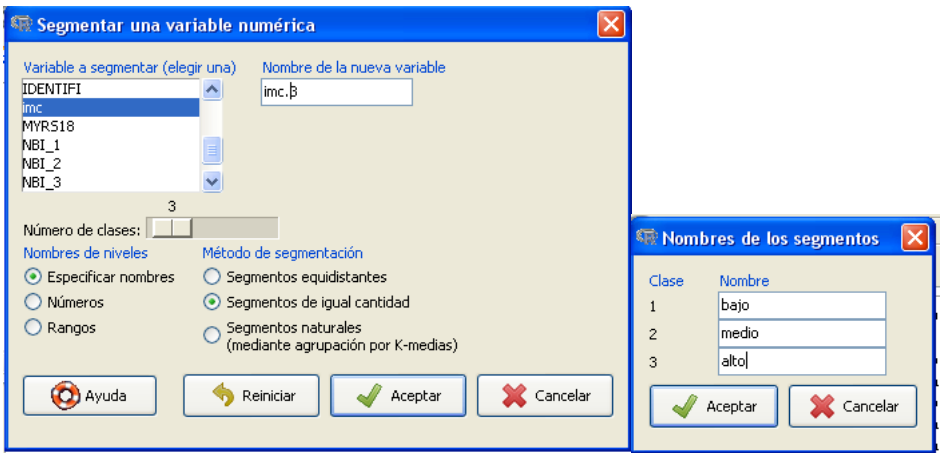
Se colocan los niveles en el orden correcto.

Categorización automática

Una variable cuantitativa puede necesitar ser categorizada en niveles, de manera similar a un factor ordenado. Para ello se usa el comando *datos* → *modificar variables del conjunto de datos activo*→ *segmentar variable numérica*. Esta instrucción exige decidir varios aspectos de la categorización:

- La cantidad de categorías
- Si cada categoría tendrá:
  - un nombre que el usuario asigna
  - códigos numéricos
  - la especificación del conjunto de valores que contiene cada intervalo
- El criterio para realizar los cortes:
  - intervalos de igual amplitud
  - intervalos de igual cantidad de casos
  - cortes naturales basado en agrupamientos de k medias[3]

Para el caso del imc, a modo de ejemplo, se la categoriza en tres grupos de igual cantidad de casos cada uno, a las que se denominará: bajo, medio, alto.



Por el contrario si se requiere crear una categorización en cuatro grupos de acuerdo a la clasificación BMI de la Organización Mundial de la Salud (`http://apps.who.int/bmi/index.jsp?introPage=intro_3.html`):

[3]Que particiona el conjunto de observaciones en k grupos, de modo que cada observación pertenece al grupo más próximo a la media.

Clasificación	intervalo de IMC
Bajo peso	menos de 18.50
Peso normal	18.50 - 24.99
Sobrepeso	25.00-29.99
Obesidad	30 o más

Deberá usarse el procedimiento de categorización manual, que se describe antes.

Ejercicio 4

Defina como factor a la variable BICA02 ¿Cuándo fue la última vez que tomó alguna de estas bebidas alcohólicas?
Con categorías:

- 1 Durante los últimos 30 días
- 2 Hace más de un mes, pero menos de un año
- 3 Hace más de un año
- 9 Ns/Nc

Etiquete sus categorías incluyendo al 9 como NA. Ordene los niveles de ese factor

Categorice BIPC06 (altura en centímetros) en cuatro grupos con la misma cantidad de casos, llame talla.4 a esta variable.
Solicite que muestre las categorías como intervalos.
No olvide indicar al 999 como NA antes de la categorización.

Análisis de datos

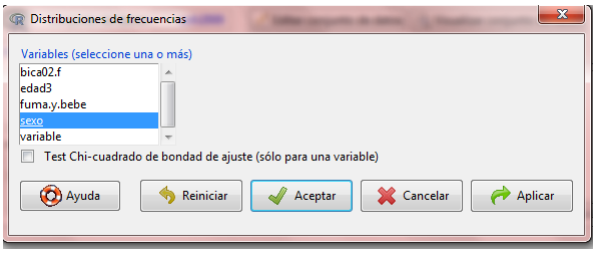
Una vez que la matriz de datos está cargada y hay cierta familiaridad con la información que contiene, se la debe resumir, a fin de hacerla inteligible. No es posible sacar conclusiones solo mirando el conjunto de datos que está cargado. Estas operaciones de resumen se realizan con las variables individualmente y luego estableciendo relaciones entre dos o más de ellas.

Descripciones univariadas

El procedimiento de descripción univariada depende del tipo de variable con que se trabaje y de su nivel de medición. Se realizan tablas de contingencia para los factores (o variables nominales) y se calculan medidas descriptivas para las cuantitativas.

Variables categóricas

Tablas de frecuencia Sea como ejemplo, la distribución de frecuencias de la variable sexo, a la que ya se ha transformado en factor y dado nombre a sus categorías. La instrucción *Estadísticos→Resúmenes→Distribución de frecuencias* abre una ventana para elegir la variable que se va a describir:



La salida tiene la forma:

```
counts:
sexo
varón mujer
14317 18048

percentages:
sexo
varón mujer
44.24 55.76
```

Que puede formatearse como[4]:

Allí se leen los recuentos absolutos de las dos categorías y los porcentajes en cada una más abajo.

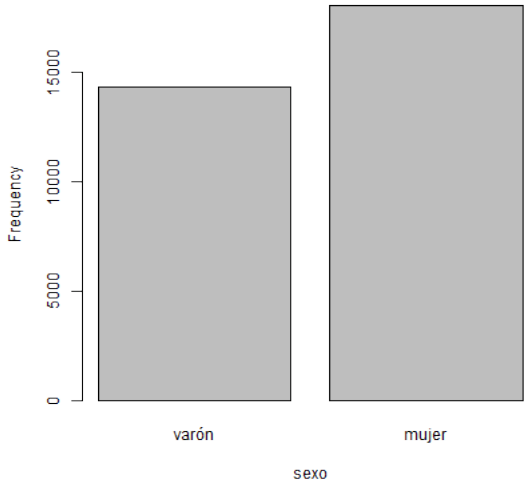
[4]Ver el paquete xtable para formatear objetos R (`https://cran.r-project.org/web/packages/xtable/xtable.pdf`.

	sexo
varon	14317
mujer	18048

	sexo
varon	44.24
mujer	55.76

En R Commander no es posible obtener una distribución de frecuencias para una variable cuantitativa, si se lo necesita, se la debe transformar en factor primero. En ese caso es conveniente usar un nombre diferente a fin de preservar la variable cuantitativa para cálculos posteriores[5].

Gráficos Los gráficos más usados para representar descripciones univariadas de variables cualitativas, son los gráficos de barras y, si las categorías son pocas, el de sectores circulares. Todas ellos se encuentran en el comando *Gráficas*. Por ejemplo, un gráfico de barras para la variable sexo se solicita con *Gráficas→Gráfica de barras* y se obtiene como resultado:



Para mejorar la estética del gráfico se debe ingresar a la sintaxis que lo generó, la que originalmente es:

with(enfr2013, Barplot(sexo, xlab="sexo", ylab="Frequency"))

Allí, es posible modificar aspectos:

[5]Esta es una condición para usar los comandos bajo el formato de ventanas, pero siempre es posible escribir las instrucciones en la consola. Por ejemplo, para la pregunta si fuma actualmente (BITA04), las categorías son códigos numéricos y, aunque así lo entiende R, se puede obtener una distribución de frecuencias solicitando:

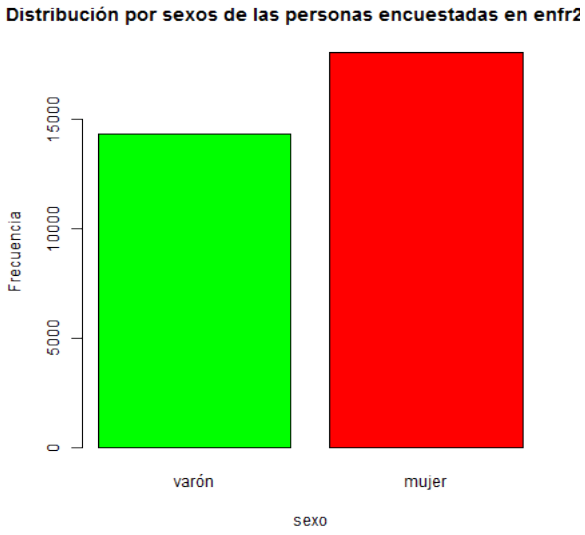
table(enfr2013$BITA04)

lo cual produce la tabla:

1	2	3
6039	2408	8028

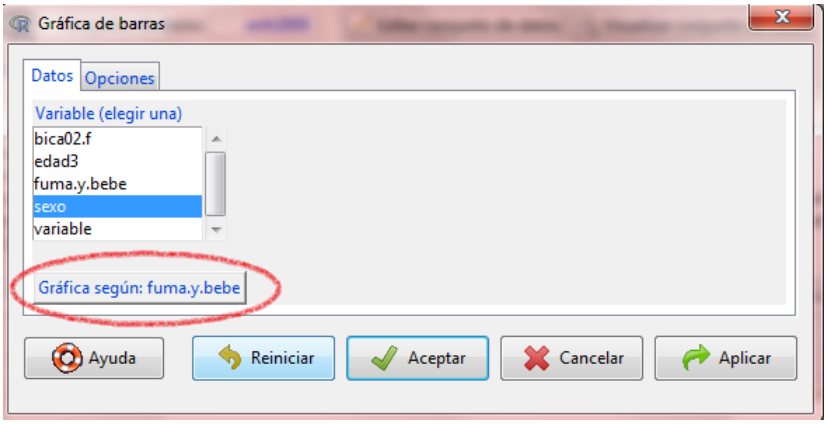
with(enfr2013, barplot(table(sexo), xlab="sexo", ylab="Frecuencia",main="Distribución por sexos de las personas encuestadas en enfr2013",col=c("green","red")))

donde se ha cambiado la etiqueta del eje y (ylab), se agregó un título (main) y se indicaron los colores para cada barra, de lo que resulta:

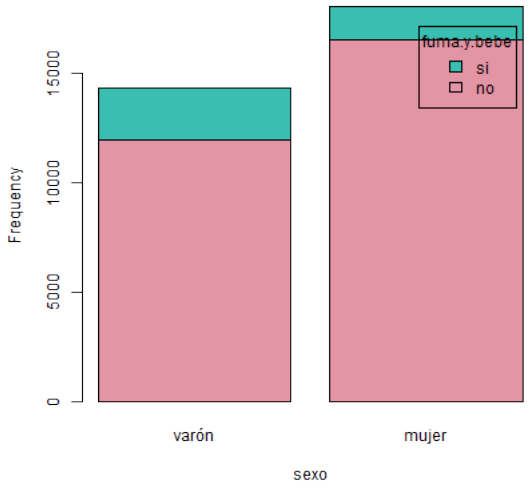


Para ver un listado de todos los nombres de colores posibles, solicite *colors()* en la consola.

Los gráficos pueden particionarse para comparar grupos, por ejemplo:

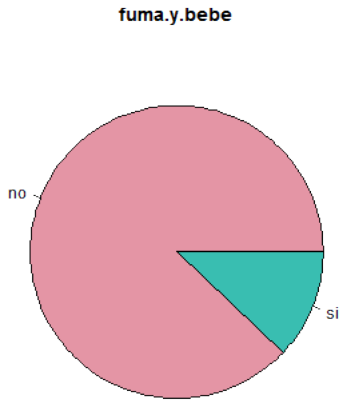


Produce:



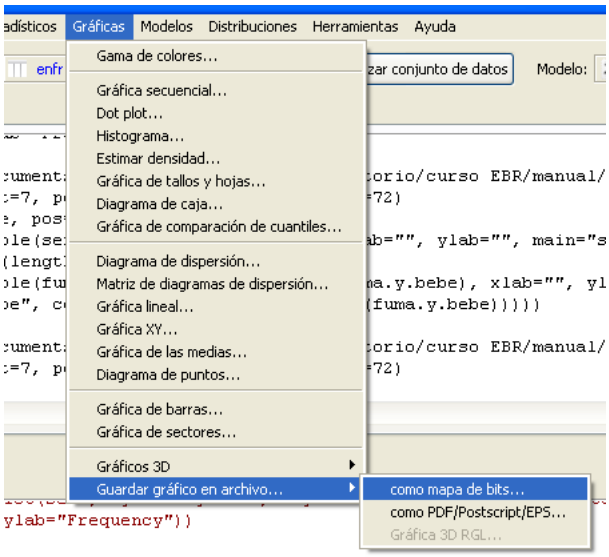
Cuyos rótulos y colores pueden también editarse.

El gráfico de sectores también es adecuado, a condición de no tener más de cuatro categorías y que no haya dos que sean muy similares, ya que el ojo humano no detecta la diferencia entre ángulos de sectores circulares tan bien como la diferencia de altura entre barras. Para el caso de las variables anteriores resulta:



Que también admite ajustes en los colores y las leyendas.

Guardar un gráfico El guardado de los gráficos se realiza en *Gráficas→Guardar gráfico en archivo...→Como mapa de bits...*



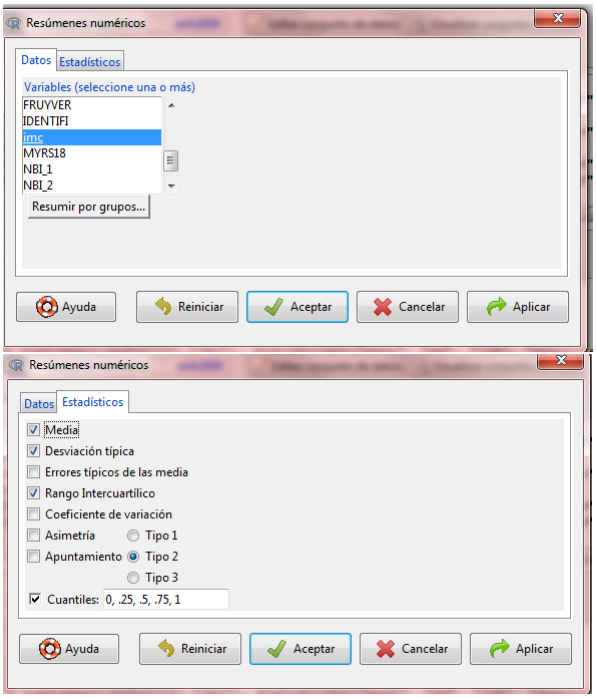
Luego puede elegirse si el formato será png o jpeg y el tamaño de la imagen.

Ejercicio 5

Construya una distribución de frecuencia para talla.4. Aunque proviene de una variable continua, está categorizada en cuatro grupos y admite un tratamiento como categórica. Grafique esa distribución.
Compare los gráficos de varones y mujeres.

Variables cuantitativas

Medidas descriptivas Para describir variable numéricas se usan medidas resumen: de posición, dispersión y forma, que están disponibles en: *Estadísticas→Resúmenes→ Resúmenes numéricos*. La descripción del Índice de Masa Corporal se solicita:

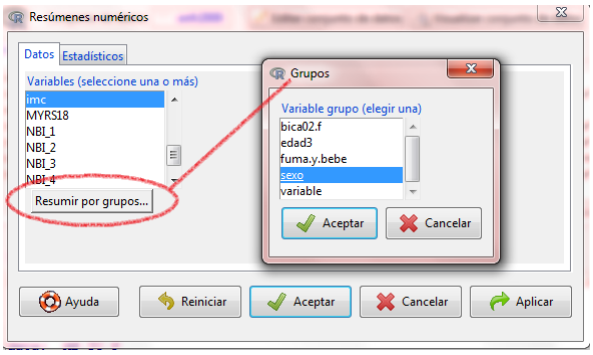


Que ofrece por defecto los percentiles 0 (valor mínimo de la variable), 25, 50, 75 y 100 (valor máximo), pero que pueden editarse para solicitar los que se elija. La salida es:

mean	sd	IQR	0 %	25 %	50 %	75 %	100 %	n	NA
26.69198	5.348918	6.182372	11.08033	23.18339	25.96953	29.36576	259.1068	30290	2075

La salida muestra: media, desviación estándar, recorrido intercuartílico, valor mínimo, primer cuartil, mediana, tercer cuartil, valor máximo, número de casos y número de casos perdidos.

Las medidas descriptivas pueden particionarse a fin de comparar grupos:

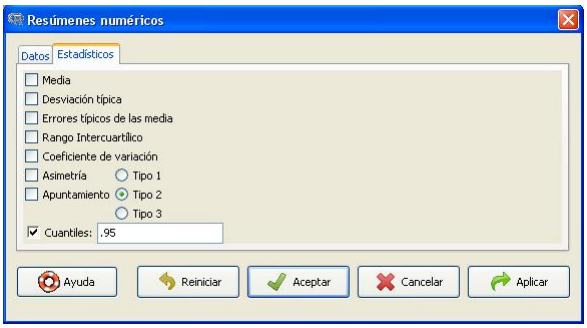


Para obtener como salida:

	mean	sd	IQR	0 %	25 %	50 %	75 %	100 %	data:n	data:NA
varon	27.38984	5.239027	5.485967	13.84083	24.22145	26.77551	29.70742	259.10684	13626	691
mujer	26.12135	5.370300	6.530570	11.08033	22.37568	25.25952	28.90625	70.17236	16664	1384

Estos resultados muestran una alta variabilidad, que confunde la identificación de alguna tendencia general. Además, el grupo de varones muestra al menos un caso excepcionalmente elevado: el máximo es 259, que es un índice de masa corporal 8 veces superior a lo que OMS considera obesidad. Una estrategia posible consiste en separar los análisis; por un lado observar en detalle los valores extremos (outliers) y por otro analizar la tendencia del grupo menos extremo. Luego el análisis estadístico continúa con los valores menos extremos de la distribución, dejando los excepcionales para una mirada particularizada.

Un criterio elemental[6] para considerar valores extremos superiores es tratar como tales a los que se ubican por encima del percentil 95. Éste se solicita en el mismo comando que el anterior indicando ahora que solo se pide el cuantil 0.95:



El percentil 95 resulta ser 35.75[7]. Ahora se recorta la base para seleccionar solo los casos con imc inferior a 35.75: *Datos→Conjunto de datos activo→Filtrar el conjunto de datos activo*:

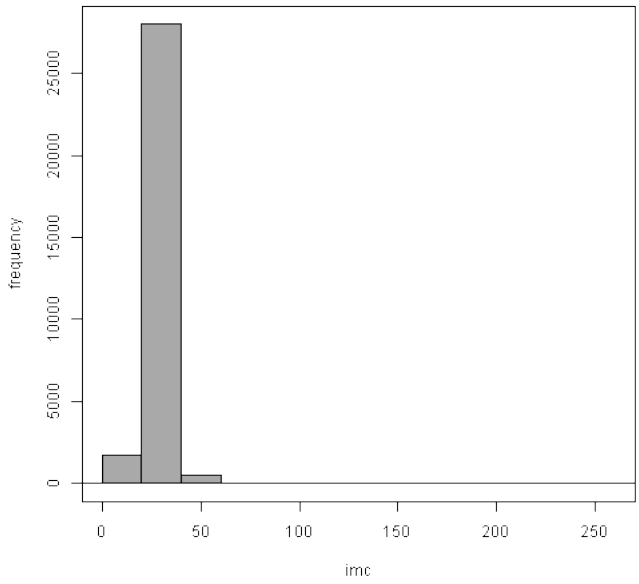
El nuevo conjunto de datos ha sido llamado enfr.recorte y tiene menos casos que el original, porque se han quitado los valores extremos superiores. Eso queda indicado en la ventana de mensajes, este conjunto es el que queda activo. La descripción del imc es ahora:

mean	sd	IQR	0 %	25 %	50 %	75 %	100 %	n	NA
26.16596	4.242347	5.888196	11.08033	23.11111	25.80645	28.99931	37.74133	29327	2075

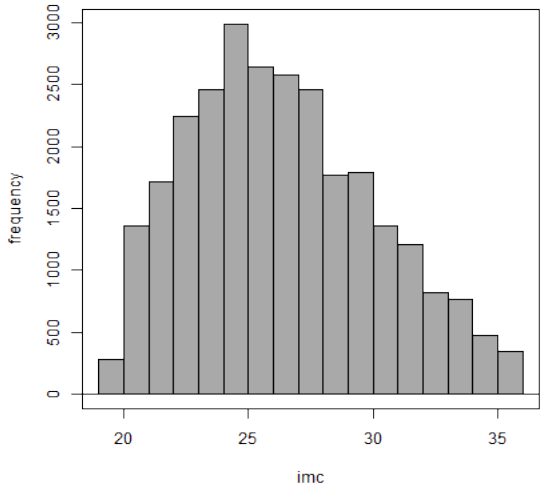
[6]Los criterios (de Chauvenet, de Peirce) para decidir si un valor es un caso atípico toman en consideración la probabilidad de hallarlo y la cantidad de casos bajo análisis.

[7]El criterio de Chauvenet establece como punto de corte el valor 48.9 del imc.

Gráficos El histograma muestra la distribución de frecuencias de una variable cuantitativa y se obtiene en *Gráficas→Histograma*. Sobre la base original, que incluye los valores extremos de imc, el histograma tiene la forma:



Que no es adecuada para describir el comportamiento de la variable, debido a que la escala que se usa debe incluir a los valores extremos superiores. Por el contrario, cuando se usa la base recortada, se obtiene:

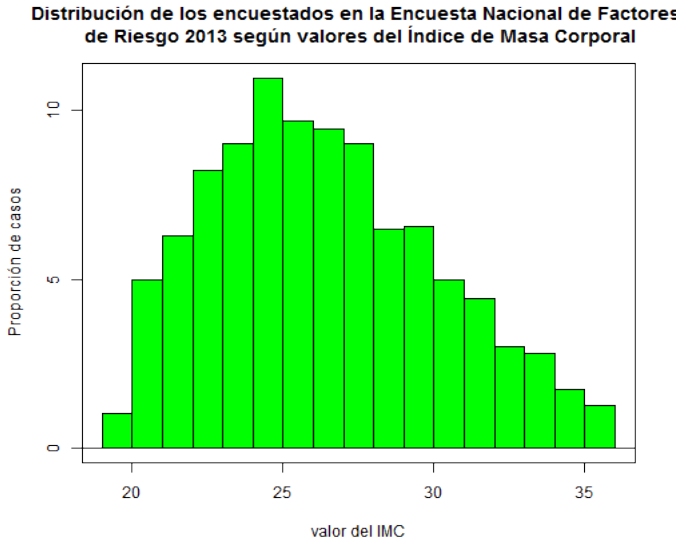


Cuyo aspecto puede mejorarse si se interviene en la sintaxis:

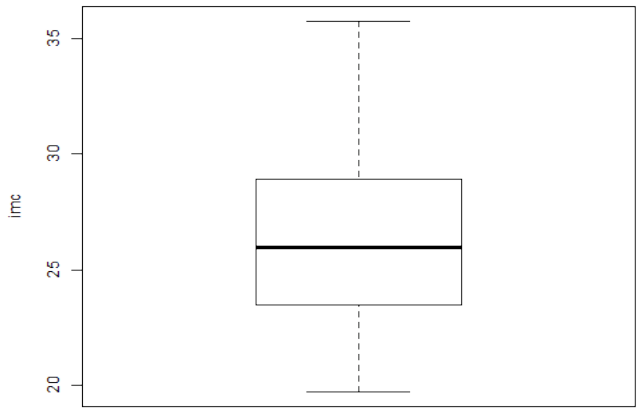
with(enfr, Hist(imc, scale="percent", breaks="Sturges", col="green", xlab="valor del IMC", ylab="Proporción de casos",

main="Distribución de los encuestados en la Encuesta Nacional de Factores \n de Riesgo 2013 según valores del Índice de Masa Corporal"))

Que da por resultado

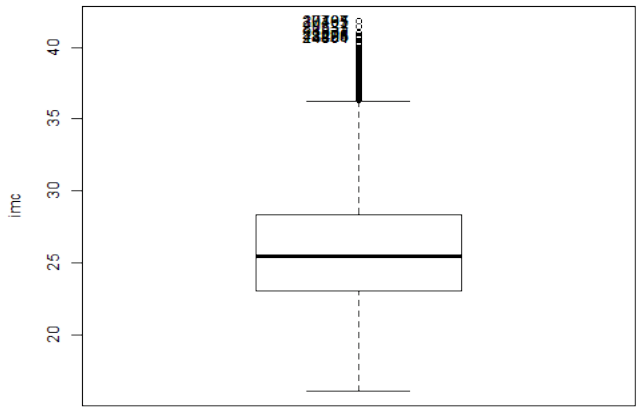


Las medidas descriptivas de una variable numérica son bien representadas por el diagrama de caja, o box plot. Para el imc tiene la forma:



Que admite el mismo tipo de edición que los demás gráficos.

Si se usa la base anterior, enfr, que contiene la variable imc original, el diagrama de caja provee una manera gráfica de identificar a los valores extremos, a los que define como aquellos que se ubican más de tres semirecorridos intercuartílicos por encima del tercer cuartil. De este modo se detectan los casos que contienen esos valores, ubicarlos en la base y analizarlos por separado. El gráfico que resulta es:



Ejercicio 6

Describa la variable BIPC05 (peso corporal), pero antes indique que 999 es NA. Verifique si tiene casos muy extremos.

Defina una nueva variable, peso.rec para la cual los valores superiores al percentil 95 e inferiores al percentil 5 sean casos perdidos (NA).

Aclaración: no filtre el conjunto de datos, por el contrario, construya una nueva variable peso.rec que en primer lugar es igual a BIPC05 y luego recorta sus valores extremos.

Grafique peso.rec con histograma y diagrama de caja

Descripciones bivariadas

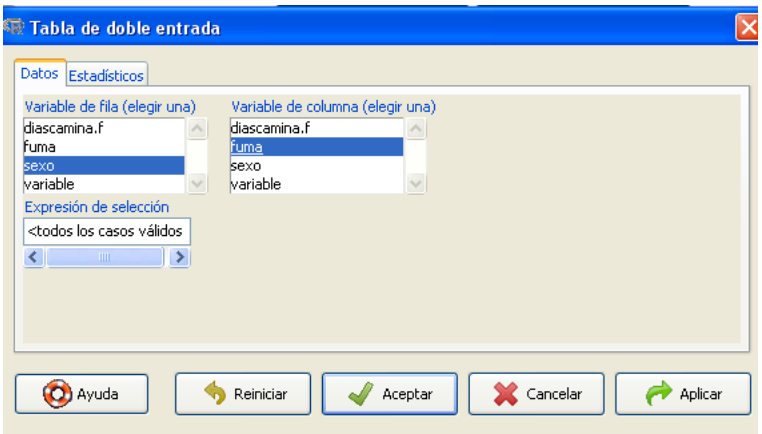
La descripción de dos variables de manera simultanea permite hallar casos de variación conjunta. Esto puede expresarse a través de la visualización en una tabla, de la comparación de proporciones o de medidas sintéticas: los coeficientes de asociación. Los análisis difieren según se trate de variables categóricas o numéricas.

Variables categóricas

Tablas de contingencia Un tabla de contingencia es un arreglo en filas y columnas determinadas por las categorías de cada una de dos variables que se "cruzan". Las celdas de la tabla contienen recuentos de casos, es decir frecuencias que suceden conjuntamente entre una categoría de la variable de las filas y una de la variable de las columnas. La construcción de una tabla de contingencia exige que ambas variables sean factores. Para cruzar sexo y habito de fumar, se debe llevar a factor BITA04:

Actualmente ¿fuma usted cigarrillos...
1 ...todos los días?
2 ...algunos días?
3 ...no fuma?

Una vez hecho esto y construida la variable fuma, se solicita, en *Estadísticos*→*Tablas de contingencia*→ *Tablas de doble entrada*. En la ventana del selector de variables solo aparecen las que son factores, el orden de selección es primero filas, luego columnas.



La salida tiene la forma de la tabla siguiente:

Distribución conjunta de la condición de fumador y sexo

	todos los días	algunos días	no fuma
varón	2849	1121	3589
mujer	2329	888	3232

Pearson's Chi-squared test

data: .Table

X-squared = 10.035, df = 2, p-value = 0.006622

La lectura de las frecuencias absolutas es tal que "se cuentan 2849 mujeres que fuman todos los días" y del mismo modo el resto.

La prueba Ji Cuadrado (χ^2) indica que, con el criterio usual, se rechaza la hipótesis de independencia ($p<0.05$), con lo que puede sostenerse cierta relación entre las dos variables. Para apreciar el modo en que ésta sucede, se solicitan frecuencias relativas por filas, ya que interesa comparar la condición de fumador entre varones y mujeres. Se obtiene la tabla que se muestra a continuación:

Distribución de la condición de fumador según sexo

	todos los días	algunos días	no fuma	Total	Count
varon	37.70	14.80	47.50	100.00	7559.00
mujer	36.10	13.80	50.10	100.00	6449.00

La tabla muestra que el 37.7 % de los varones fuma todos los días, mientras que así lo hace el 36.1 % de las mujeres. Los no fumadores son el 47.5 % de los varones y el 50.1 % de las mujeres. Así, aunque es significativa, la diferencia entre los grupos es pequeña[8].

Medida de la asociación Una cuantificación de la distancia a la que se encuentra la distribución bivariada de la que resultaría si las variables fueran independientes, es ofrecida por el coeficiente V de Cramer, que está basado en el puntaje χ^2. Este coeficiente no es una función directa en R Commander, por eso servirá como ejemplo del modo de resolver limitaciones de esta interfaz y requerirá un pequeño uso de sintaxis. Como se mencionó al comienzo, hay comunidades de usuarios de R que regularmente construyen paquetes para resolver tareas específicas. En este caso, luego se una búsqueda en el sitio *inside R* (`http://www.inside-r.org/`)se encuentra que ese coeficiente está (entre otros) en un paquete que se llama lsr. La instalación debe hacerse desde R, no desde R Commander, ya que este último solo admite cargar paquetes que ya han sido instalados. Como antes, se selecciona el sitio espejo desde donde hacer la instalación y luego se busca lsr en la lista. Una vez instalado, desde R Commander, en *Herramientas→ Cargar paquetes*, se lo elige de la lista (solo aparece aquí si ha sido instalado desde la consola de R). Luego se define la tabla que cruza las dos variables, para disponer de ella como objeto, bajo un nombre arbitrario (para ejecutar cada comando que se escribe se usa Ctrl+R o bien el botón ejecutar):

```
tabla=xtabs(~sexo+fuma,data=enfr.recorte)
```

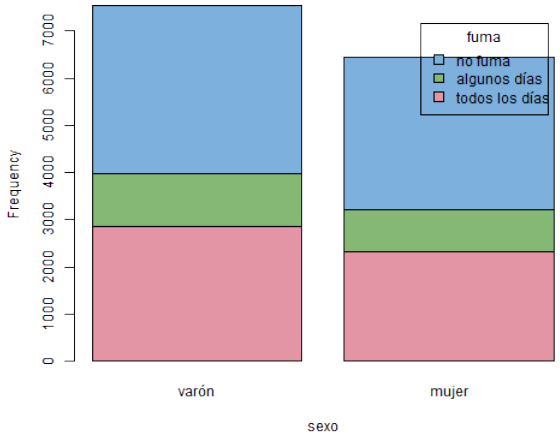
Y se solicita el coeficiente V de Cramer para este objeto:

[8]Como sucede con la prueba χ^2, las muestras de gran tamaño dan resultados significativos aunque se trate de relaciones muy débiles.

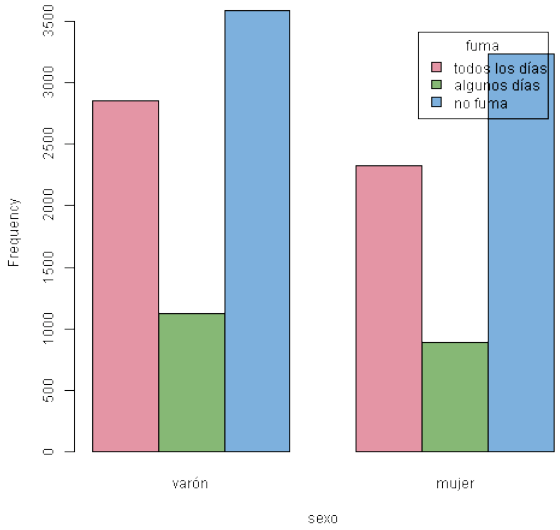
cramersV(tabla)

Que devuelve su valor, 0.0268, cuya pequeña magnitud confirma que es poca la diferencia entre los grupos que se comparan.

Gráficos La representación gráfica de las relaciones entre variables debe hacerse cuidadosamente, porque corre el riesgo de mostrar resultados engañosos. Para dos variables categóricas, el gráfico más usado es el de barras, que pueden ser superpuestas o adyacentes. En *Gráficas→Gráfica de barras* se encuentra la opción de elegir una variable para graficar y otra para definir los grupos de comparación. En este ejemplo, se elige graficar sexo y agrupar por fuma para obtener:



Alternativamente, las barras pueden ubicarse al lado:



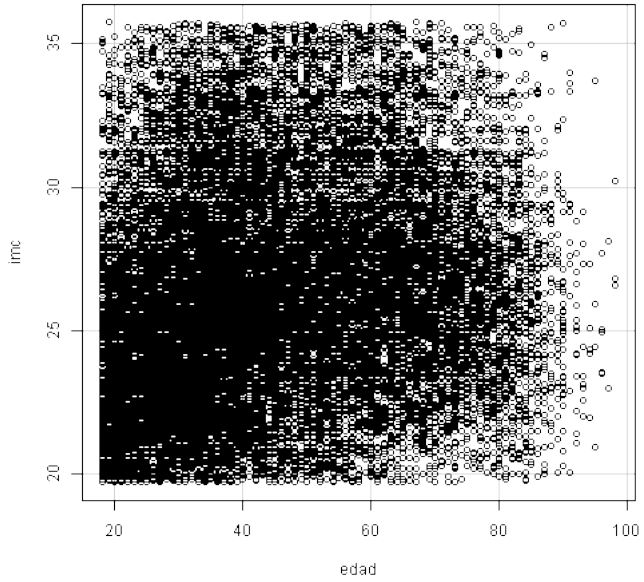
Ejercicio 7

Realice una tabla bivariada que cruce NIVEL_ACTIVIDAD_FISICA con el índice de masa corporal en tres categorías de igual número de casos. Antes de hacerlo deberá definir como factor a la primera variable.
Calcule frecuencias relativas.
Evalúe la intensidad de la asociación.
Represente gráficamente.

Variables cuantitativas

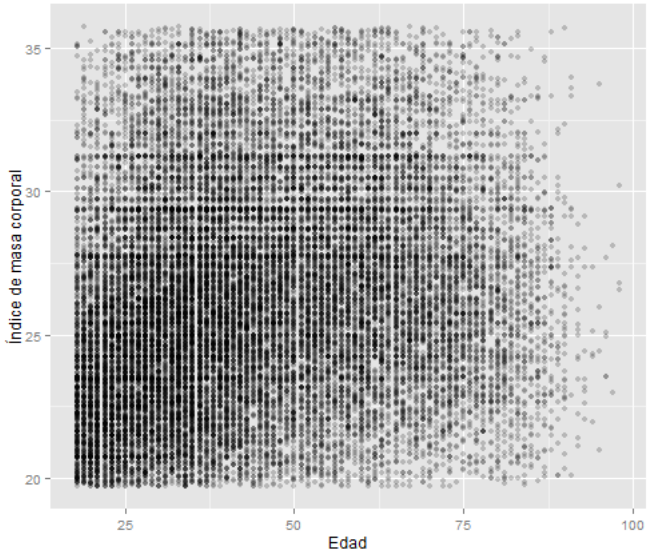
La diferencia más importante es que ahora no es posible construir tablas que contengan toda la información que ofrece una variable cuantitativa. Para hacerlo habría que categorizarla, con lo que se pierde el carácter cuantitativo.

Dos métricas: diagrama de dispersión Como contrapartida de las tablas de doble entrada, las variables cuantitativas se "cruzan" por medio de una diagrama de dispersión. Allí, las filas y las columnas de la tabla se convierten en ejes cartesianos y las celdas son puntos del plano. Por ejemplo, para visualizar la relación entre el índice de masa corporal y la edad, se solicita *Gráficas→Diagrama de dispersión*

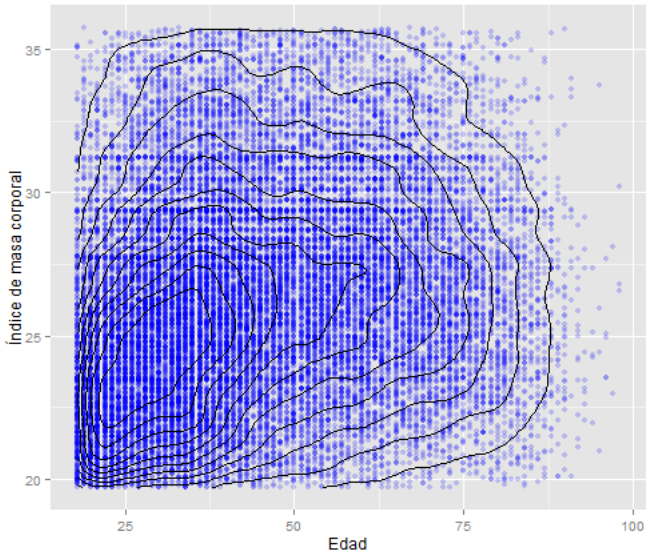


La información que provee este gráfico es escasa; para apreciar alguna tendencia es necesario que se vea donde hay más densidad de puntos. Esto puede hacerse con un paquete especializado en gráficos, llamado ggplot2 (Wickham, 2009), que no está (aun) como aplicación en R Commander, por lo que es necesario cargar el paquete[9] y luego escribir la sintaxis en la consola. Para el caso que se está usando como ejemplo, el gráfico puede mejorarse agregando transparencia a los puntos:

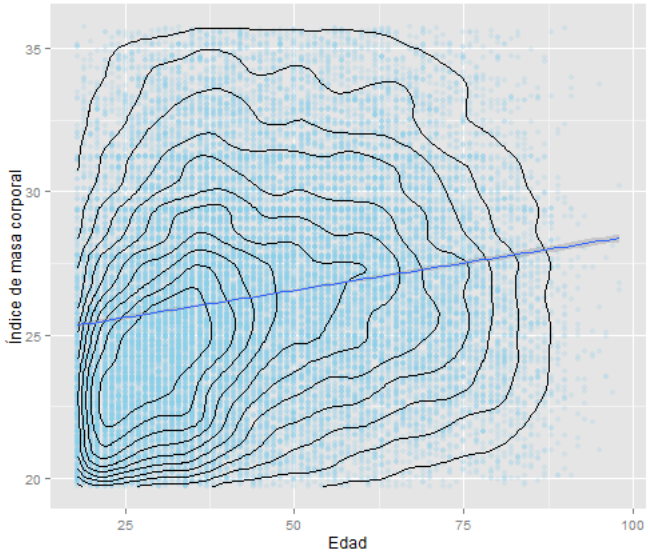
[9]No requiere la instalación desde la consola de R, porque viene instalado por defecto en R Commander, solo su carga para la sesión



También conviene agregar líneas que indiquen zonas de mayor concentración:



Y la recta de regresión, aclarando un poco el color de los puntos:



Las instrucciones para los tres últimos gráficos son las siguientes:

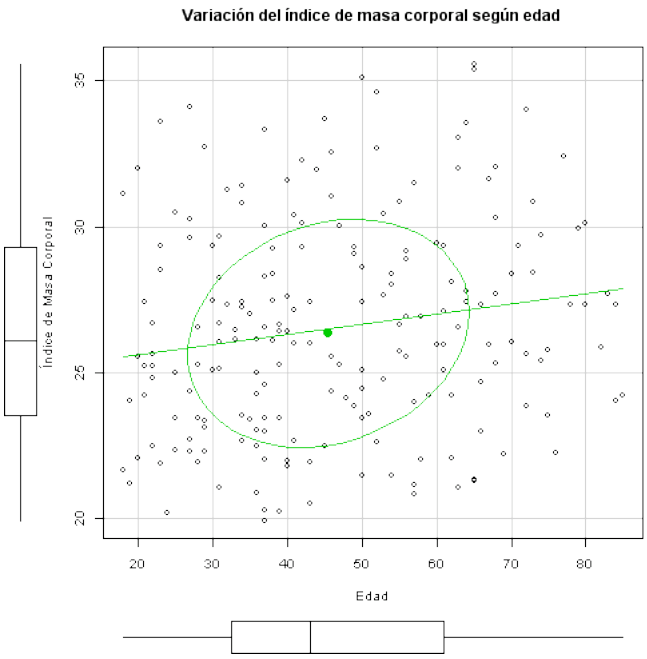
- ggplot(enfr.recorte,aes(x=BHCH05,y=imc)) + geom_point(colour="black", alpha=0.2) +xlab("Edad")+ylab("Índice de masa corporal")
- ggplot(enfr.recorte,aes(x=BHCH05,y=imc)) + geom_point(colour="blue", alpha=0.2) + geom_density2d(colour="black")+xlab("Edad")+ylab("Índice de masa corporal")
- ggplot(enfr.recorte,aes(x=BHCH05,y=imc)) + geom_point(colour="skyblue", alpha=0.2) + geom_density2d(colour="black")+xlab("Edad")+ylab("Índice de masa corporal")+stat_smooth(meth = "lm")

A fin de apreciar lo que ofrece R Commander para los diagramas de dispersión, se trabajará con una selección de 200 casos aleatoriamente elegidos de la matriz enfr.recorte. Para ello se usa la instrucción:

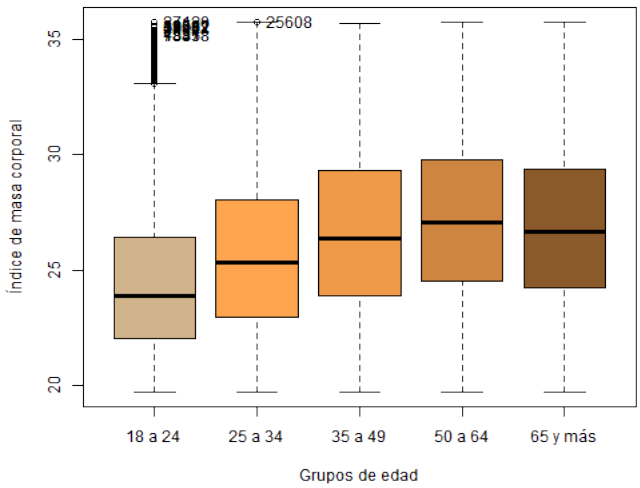
enfr200=enfr.recorte[sample(nrow(enfr.recorte), 200),]

Que define otro conjunto de datos consistente en una muestra aleatoria de enfr.recorte. Esta nueva matriz no queda inmediatamente activa luego de creada, sino que debe ser explícitamente cargada con el botón al lado de *conjunto de datos*. En la ventana de mensajes se lee que tiene 200 filas (casos) y conserva todas las columnas (variables).

Sobre este nuevo conjunto de datos, se ha agregado la recta de regresión, un box plot para cada variable y una elipse que incluye al 50 % de las observaciones. Todo ello está disponible en la solapa "opciones" de la ventana *Gráficas→Diagrama de dispersión* en R Commander. El gráfico toma la forma:



Una métrica y una ordinal: sucesión de box-plots Cuando una variable es ordinal y la otra cuantitativa, conviene graficar una sucesión de diagramas de caja, que ilustren una posible tendencia. Por ejemplo, para mostrar diferencias en el índice de masa corporal entre diferentes grupos de edades, se solicita el diagrama de caja según grupos de edad, para obtener:



Los valores extremos son identificados por el número de caso. Aunque para el primer grupo de edad ésto resulta ilegible en el gráfico, el listado de casos con valores atípicos se muestra en la ventana de resultados.

Con dos variables cuantitativas resulta ilustrativo el gráfico llamado bagplot (Rousseeuw et al., 1999) es un método gráfico que permite visualizar datos bidimensionales, y

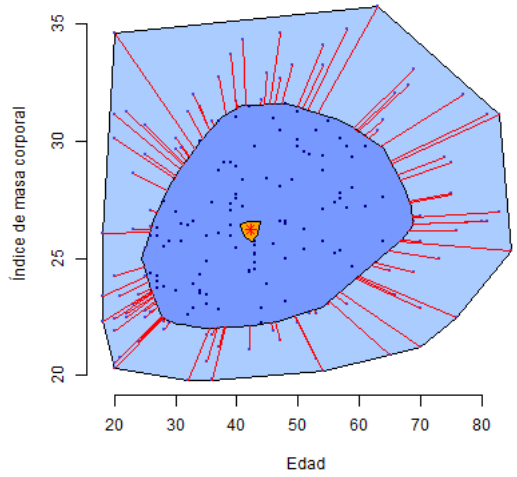
consiste en una versión bivariada del diagrama de caja. El bag plot muestra la posición, dispersión, asimetría y valores atípicos en un conjunto de datos. Este gráfico no está disponible directamente en R Commander, por lo que debe instalarse un paquete específico, que se llama aplpack (Wolf & Bielefeld, 2015), en R y luego cargarlo desde R Commander. Hecho esto, se solicita en la consola la definición del objeto "bagplotimcxedad":

bagplotimcxedad=compute.bagplot(edad,imc)

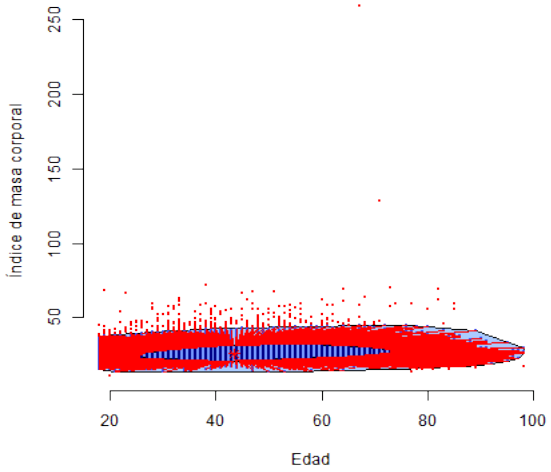
Y luego su representación gráfica:

plot(bagplotimcxedad, xlab="Edad", ylab="Índice de masa corporal")

Que especifica los nombres de los ejes. El gráfico resulta:



De manera alternativa, si se construye sobre la base original, que conserva los valores extremos del índice de masa corporal, el gráfico resulta menos ilustrativo:



Medida de la asociación La intensidad de la asociación entre variables cuantitativas se evalúa por medio del coeficiente r de Pearson. Si una o más de las variables son de nivel ordinal, se utiliza el coeficiente r de Spearman. La operación está disponible en *Estadísticas→Resúmenes→Matriz de correlacion*es. Una vez elegidas las dos o más variables, se opta por el coeficiente adecuado y se obtiene una salida en forma matricial, de la que cada elemento es el coeficiente entre la variable de la fila y la de la columna. Para este ejemplo se elije correlacionar edad, peso, talla e imc. Este último tiene alta correlación con el peso, porque guarda con él una relación lineal, pero no con la talla.

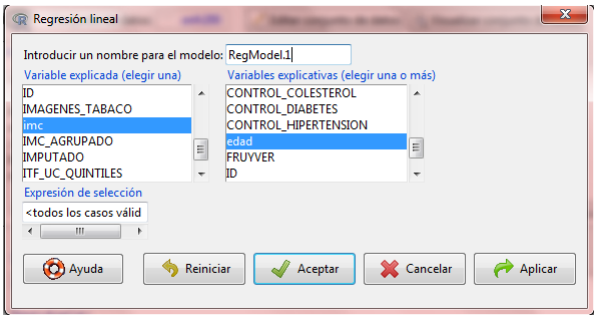
	edad	imc	peso	talla
edad	1.00	0.23	0.07	-0.17
imc	0.23	1.00	0.77	0.01
peso	0.07	0.77	1.00	0.64
talla	-0.17	0.01	0.64	1.00

Con el mismo formato se obtienen los valores de significación para cada coeficiente, si se marca la opción *p-valores pareados:*

	edad	imc	peso	talla
edad	-	0.0034	0.7177	0.0458
imc	0.0034	-	<0.0001	0.9439
peso	0.7177	<0.0001	-	<0.0001
talla	0.0458	0.9439	<0.0001	-

Modelo lineal

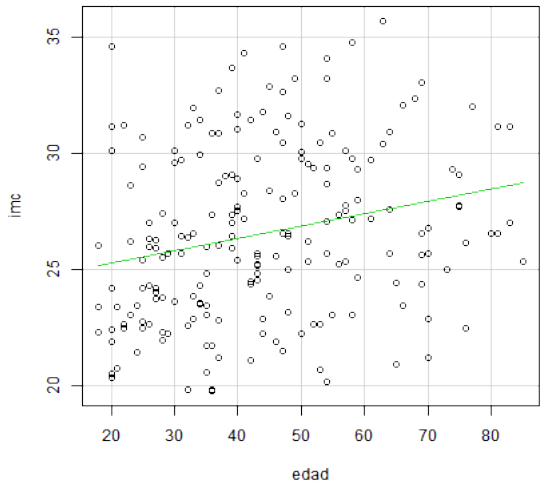
Una variable independiente Para ilustrar la construcción de un modelo de regresión lineal simple, se usará la variable edad como explicativa del índice de masa corporal, bajo la hipótesis de un efecto lineal de la primera variable sobre la segunda. Con la instrucción *Estadísticos→Ajuste de modelos→Regresión lineal* aparece la ventana que permite elegir una variable dependiente (explicada o de salida) y una o más regresoras (explicativas o independientes).



Y se obtiene como salida:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.6619	0.0602	409.61	0.0000
edad	0.0380	0.0013	30.40	0.0000

La leve (pero significativa) pendiente positiva de la recta puede verse si se pide el diagrama de dispersión para este modelo:

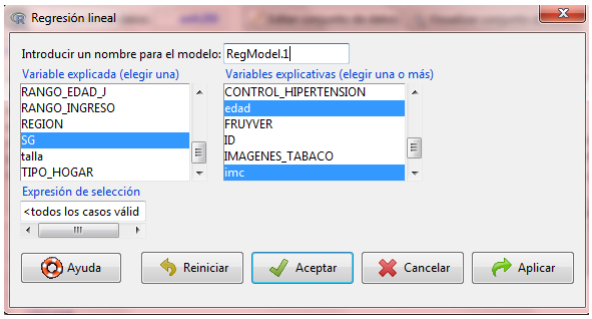


Ejercicio 8

Analice la relación entre el peso y los días de caminata semanales (no olvide declarar al 8 como perdido). Calcule una medida adecuada de asociación.

Dos variables independientes La ENFR incluye una variable llamada BISG01 que recoge la evaluación subjetiva por parte del encuestado, de su estado general de salud, con categorías ordinales de 1 = Excelente hasta 5 = Mala. Para el ejemplo que sigue, se ha construido una nueva variable, llamada "SG" (por Salud General) como promedio de las puntuaciones estandarizadas de las respuestas dadas a las seis preguntas del bloque "Salud General", que, además de BISG01, indagan por: movilidad, cuidado personal, actividades cotidianas, dolor/malestar y ansiedad/depresión.

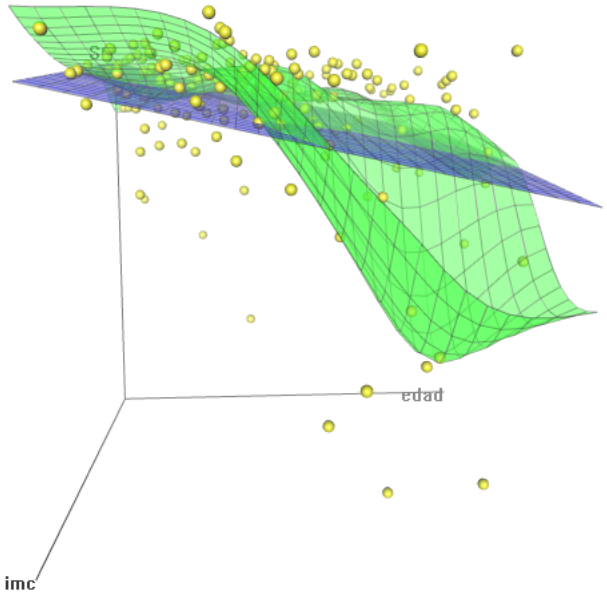
A modo ilustrativo, se prueba el ajuste del modelo lineal que relaciona esta puntuación con la edad y el índice de masa corporal de las personas encuestadas. Para ello se solicita: *Estadísticos*→*Ajuste de modelos*→*Regresión lineal* y en la ventana siguiente se eligen las variables, usando Ctrl para elegir varias regresoras.



La salida tiene la forma:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	100.8620	8.3666	12.06	0.0000
BHCH05	-0.3223	0.0580	-5.56	0.0000
imc	-0.3298	0.3195	-1.03	0.3033

Que muestra que habría un efecto negativo de ambas variables sobre la percepción del estado de salud, pero solo el de la edad sería un efecto significativo. Cuando las variables explicativas son dos, el paquete rgl (Adler & Murdoch, 2014)—que R Commander carga por defecto—, permite que este modelo sea representado con un gráfico tridimensional interactivo, del que aquí solo se muestra una imagen estática:


edad
imc

Comentario final

La gran variedad de paquetes y su permanente actualización hacen de R una herramienta muy flexible y recomendable para casi cualquier campo de investigación. Si bien al comienzo de su desarrollo tuvo un carácter un tanto elitista, reservado a quienes pudieran escribir las instrucciones para hacer cada proceso, esto ha sido superado con la disponibilidad de interfaces gráficas como R Commander.

El control que permite tener sobre los gráficos abre un amplio espacio para mejorar la comunicabilidad de los resultados de investigación y llegar a públicos refractarios a la información cuantitativa.

Creemos que es una muy buena opción ingresar en la aventura de explorar este lenguaje. Es un camino inacabable, por el que transitan muchas personas con gran voluntad para ayudar a los que se suman.

Anexo: Visitando la sintaxis

Este anexo muestra la operación de R sin usar una interfaz gráfica. Los comandos pueden escribirse directamente en la consola de R y se ejecutan con "Enter" o bien abrir un "script", que es un archivo de texto en el que se reúnen las instrucciones que se escriben durante la sesión y que se guarda para usar en otras ocasiones. Si se opta por escribir la sintaxis en el script, la ejecución se realiza con la combinación "Ctrl+R". Cuando se usa R Commander, la ventana script funciona del mismo modo, solo que agrega un botón "Ejecutar", equivalente a "Ctrl + R". Cuando se escriben instrucciones, el signo numeral # sirve para separar un comando de un comentario, todo lo escrito a la derecha del numeral no se ejecuta, porque es tomado como texto accesorio. En la consola pueden hacerse operaciones:

- > 2+3 # hace la cuenta.
    - [1] 5
- > x=7 # asigna a la letra x el valor 7. Crea el objeto "x".
- > x # se invoca a x para ver el valor asignado.
    - [1] 7
- > class(x) # indica el tipo de objeto que es x.
    - [1] "numeric"
- > 1:4 # produce y muestra la secuencia desde 1 hasta 4 que no queda almacenada en ningún lugar.
- > z=1:4 # ahora crea el objeto z y no lo muestra hasta que no se lo invoque.
- >t=c(1,3,5,7) # define t como la concatenación de los valores entre paréntesis.
- > y=x*4 # se define y como x por 4.
- > y # se invoca a y para obtener su valor.
    - [1] 28
- > operacion=function(a,b){5*a+3*b+5} # se define una función de los argumentos a y b que consiste en multiplicar a a por 5, luego sumarle b multiplicada por 3 y finalmente sumar 5.
- > operacion(2,5) # se aplica esa operación a los valores *a=2* y *b=5.*
    - [1] 30

- > class(operacion) # se solicita el tipo de objeto que es “operacion”.
- [1] ”function”

Entre las operaciones listadas en el paquete *base*, aparece *mean*, la media aritmética que, cuando es aplicada sobre el vector t, definido más arriba

- > mean(t)
  - [1] 4

Da la media de los cuatro números. ¿Qué sucede si entre los datos hay casos perdidos? R identifica los casos perdidos como NA (not available). Sea el vector:

- s=c(2,3,5,7,6,NA)

Que tiene cinco casos válidos y uno perdido. El número de observaciones se obtiene pidiendo la *longitud* del vector:

- length(s)

Al solicitar la media de s, se obtiene NA, debido a que el cálculo de la media se hace sobre todos los valores, incluyendo los perdidos. Para solucionar esto, se consulta sobre esta función:

- help(mean) (que conduce al mismo resultado que ?mean)

Allí se dice que la sintaxis de mean incluye una instrucción na.rm, que corresponde a quitar los NA y que, por defecto, es falsa (F); es decir, que por defecto no quita los valores perdidos. Si se quiere obtener la media de los casos válidos ignorando a los perdidos, será necesario indicarlo así:

- mean(s,na.rm=T)

Y ahora el promedio es el de los cinco casos válidos

Los [1] que acompañan a los resultados indican en qué línea se encuentran, cuando haya resultados más largos, se indicarán del mismo modo las líneas sucesivas. Las letras x, y, z y t definidas, así como la operación, son objetos que quedan almacenados en R y pueden ser llamados cuando se los necesite.

Digitación manual

La construcción de un conjunto de datos, como el “alumnos1” requiere que se definan en primer lugar las columnas que corresponden a cada variable de la base, como vectores (y se accione Ctrl+R para que cada línea se ejecute, o bien se seleccionan todas las líneas y se ejecute de una sola vez):

- alumno=c(”Daniel”, ”Marcos”, ”Susana”, ”Ximena”, ”Laura”, ”Matías”, ”Marta”, ”Susana”, ”Carlos”, ”Ulises”)
- nota=as.numeric(c(8,7,7,8,8,5,7,4,9,6))
- turno=c(”M”,”M”,”T”,”T”,”T”,”T”,”M”,”T”,”T”,”T”)

Luego, estos vectores se ligan como columnas para definir el objeto alumnos1:

- alumnos1=cbind(alumno, nota, turno)

Finalmente, se redefine a este objeto como un conjunto de datos:

- alumnos1=as.data.frame(alumnos1) # si se solicita alumnos1, puede verse el conjunto de datos

Datos externos

Para obtener la base de la Encuesta Nacional de Factores de Riesgo 2013 y darle el nombre enfr2013, se solicita, en la consola de R:

- enfr2013 <- read.table(”C:/Documents and Settings/usuario/Escritorio/ENFR2013_baseusuario.txt”, header=TRUE, sep=”|”, na.strings=”NA”, dec=”.”, strip.white=TRUE) # se especifica la ruta completa hasta el lugar del archivo, además se indica que la primera fila son los nombres de las columnas, el carácter que se usa para separar campos, la identificación para los valores perdidos, el separador de decimales.

Ahora pueden explorarse algunas características de este objeto

- class(enfr2013) # indica el tipo de objeto que es enfr2013, la respuesta es *data frame*, equivalente a *matriz de datos*
- str(enfr2013) # indica la estructura de la matriz de datos: lista las variables, indica el tipo de cada una y muestra algunos valores. La lista se trunca porque esta matriz tiene muchas variables.
- head(enfr2013) # las primeras filas de la matriz de datos
- tail(enfr2013) # las últimas filas de la matriz de datos
- enfr2013[30,10] # el elemento que está en la fila 30 y columna 10. El valor de la variable que se ubica en la columna 10, para el caso que está en la fila 30.
- enfr2013[30,] # todas las columnas para la fila 30
- enfr2013[,10] # todas las filas de la columna 10
- enfr2013$BHCH04 # los valores de la variable BHCH04 (sexo del encuestado) para todos los casos

Descripción de variables

Una tabla univariada de frecuencias absolutas se obtiene por medio de:

- table(enfr2013$sexo)#construye y muestra la tabla sin almacenarla como objeto (esto es opcional)

Las frecuencias relativas (o proporciones) resultan de:

- prop.table(table(enfr2013$sexo))

El gráfico de barras:

- barplot(prop.table(table(enfr2013$sexo)))

Cuyo aspecto puede ajustarse:

- barplot(prop.table(table(enfr2013$sexo)), col=c("blue", "red"), xlab="sexo del entrevistado", ylab="proporción de casos", main="Composición por sexos de las personas encuestadas \n en la ENFR") ## el símbolo $\backslash n$ indica un cambio de línea en el título del gráfico

Es similar la construcción de una tabla de doble entrada:

- fuma.x.sexo=table(enfr2013$fuma,enfr2013$sexo) # define el objeto "fuma.x.sexo", que se construye ubicando a la variable *fuma* en las filas y *sexo* en las columnas
- fuma.x.sexo # muestra el objeto
- fuma.x.sexo.relativa=prop.table(fuma.x.sexo, 2) # define el objeto "fuma.x.sexo.relativa" que es una tabla de frecuencias relativas por columnas. El número posterior a la coma indica el denominador de las frecuencias: 1 calcula frecuencias relativas por filas, 2 lo hace por columnas y, si se deja en blanco el espacio posterior a la coma, calcula frecuencias relativas al total)
- fuma.x.sexo.relativa # lo muestra
- barplot(100*fuma.x.sexo.relativa, xlab="sexo", ylab="porcentaje", col=c("green","red"), beside=T, legend=T,ylim=c(0,100)) # construye el gráfico de barras, rotula los ejes, define colores, ubica las barras al lado (no apiladas), coloca leyenda y lleva el límite del eje y hasta el 100 %
- Pruebe de jugar con los parámetros para ver los efectos en el gráfico.

Las medidas resumen de una variable se solicitan por medio de:

- summary(enfr2013$edad)

Que ofrece por defecto: valores mínimo y máximo, primero y tercer cuartiles, mediana y media. Si se requiere otra medida, como la desviación estándar:

- sd(enfr2013$edad)

Además puede definirse una función como se indicó antes, por ejemplo, para el coeficiente de variación:

- cv=function(x){sd(x)/mean(x)}

Material en español para profundizar en R y R Commander

- Apuntes de R http://lbe.uab.es/vm/desc/r/docs/apuntesr.pdf
- Aprenda a usar R http://www.tutorialr.es/es/index.html
- R para Principiantes https://cran.r-project.org/doc/contrib/rdebuts_es.pdf
- Métodos Estadísticos con R y R Commander https://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf
- Introducción al uso de R-commander http://www.dma.ulpgc.es/profesores/personal/asp/Documentacion/Manual%20R%20commander.pdf

Algunos paquetes de interés

BayesPop Sevcikova & Raftery (2015): produce proyecciones de población para todos los países usando varias componentes probabilísticas, como la tasa global de fecundidad y la esperanza de vida. https://cran.r-project.org/web/packages/bayesPop/bayesPop.pdf

BayesTFR Sevcikova et al. (2015): realiza proyecciones probabilísticas de la tasa global de fecundidad para todos los países del mundo, por medio de un modelo bayesiano jerárquico. https://cran.r-project.org/web/packages/bayesTFR/bayesTFR.pdf

demography Hyndman et al. (2015): ofrece funciones para análisis demográfico incluyendo cálculo de tablas de mortalidad, modelización de Lee-Carter, análisis de tasas de mortalidad y fecundidad, volúmenes migratorios y proyecciones estocásticas de población. https://cran.r-project.org/web/packages/demography/demography.pdf

epiR Stevenson (2015): un paquete para analizar datos epidemiológicos. Contiene funciones para estimar y ajustar directa e indirectamente medidas sobre la prevalencia de patologías, para cuantificar la asociación en tablas de contingencia, y cálculo de intervalos de confianza para estimar riesgos y tasas de incidencia. Algunas funciones adicionales sirven para meta-análisis, interpretación de pruebas diagnósticas y cálculo de tamaños muestrales. https://cran.r-project.org/web/packages/epiR/epiR.pdf

ggmap Kahle & Wickham (2013): un conjunto de funciones para visualizar datos espaciales y modelos usando datos provenientes de diversas fuentes online, como Google Maps y Stamen Maps. Incluye herramientas para geolocalización e itinerarios. https://cran.r-project.org/web/packages/maps/maps.pdf

ggplot2 Wickham (2009): es una implementación de la gramática de gráficos (Wilkinson, 2005) en R. Permite construir los gráficos paso a paso con datos de diferentes fuentes, también implementa un sofisticado sistema de condicionamiento y una interfaz que facilita la transformación de datos en atributos gráficos. https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf y http://ggplot2.org/

grapheR Hervé (2015): Es una interfaz gráfica de usuario (GUI) para graficar en R de manera personalizada. Es una ayuda valiosa para hacer gráficos rápidamente sin necesidad de conocer comandos en R. Admite seis tipos de gráfico: histograma, diagrama de caja, de barras, de sectores, curva de frecuencias y diagrama de dispersión. https://cran.r-project.org/web/packages/GrapheR/GrapheR.pdf y http://www.maximeherve.com/r-statistiques/grapher

matching Sekhon (2011): provee funciones para realizar emparejamiento según puntaje de propensión (propensity score matching, Sekhon & Grieve (2012)) y para hallar el balance óptimo, basado en un algoritmo genético de búsqueda. También ofrece varias medidas univariadas y multivariadas para determinar si se ha alcanzado el balance. `https://cran.r-project.org/web/packages/Matching/Matching.pdf`

mirt Chalmers (2012): para el análisis de datos de respuestas dicotómicas o politómicas por medio de modelos de rasgos latentes univariados o multivariados, bajo el paradigma de la Teoría de Respuesta al Ítem. Pueden analizarse modelos exploratorios y confirmatorios con métodos de cuadratura y estocásticos. `https://cran.r-project.org/web/packages/mirt/mirt.pdf`

oaxaca Hlavac (2015): realiza una descomposición de Oaxaca Blinder, un método estadístico que descompone la diferencia de medias entre dos grupos en una parte debida a diferencias en las características de los grupos y otra que no puede ser explicada por esas diferencias. Ha sido usado principalmente para analizar discriminación de género y de raza en el mercado de trabajo, pero puede aplicarse a cualquier variable continua cuya media se compare entre dos grupos. El paquete oaxaca implementa la mayoría de la variantes de la descomposición con modelos de regresión, calcula los errores estándar por bootstrap y ofrece claras visualizaciones del resultado. `https://cran.r-project.org/web/packages/oaxaca/oaxaca.pdf`

Referencias

Adler, D. & Murdoch, D. (2014). rgl: 3D visualization device system (OpenGL).

Analytics, R. (2015). inside-R | A Community Site for R.

Chalmers, R. P. (2012). {mirt}: A Multidimensional Item Response Theory Package for the {R} Environment. *Journal of Statistical Software*, 48(6), 1–29.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, 1 edition.

Cleveland, W. S. & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387), 531–554.

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

Hervé, M. (2015). A multi-platform GUI for drawing customizable graphs in R.

Hlavac, M. (2015). oaxaca: Blinder-Oaxaca Decomposition in R.

Hyndman, R. J., Booth, H., Tickle, L., & Maindonald, J. (2015). demography.

INDEC (2015). Encuesta Nacional de Factores de Riesgo 2013.

Kahle, D. & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1), 144–161.

R Team Core (2015). R: A Language and Environment for Statistical Computing.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The Bagplot: A Bivariate Boxplot. *The American Statistician*, 53(4), 382–387.

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, 42(7), 1–52.

Sekhon, J. S. & Grieve, R. D. (2012). A matching method for improving covariate balance in cost-effectiveness analyses. *Health economics*, 21(6), 695–714.

Sevcikova, H., Alkema, L., Raftery, A., Fosdick, B., & Gerland, P. (2015). Bayesian Fertility Projection.

Sevcikova, H. & Raftery, A. (2015). Probabilistic Population Projection.

Stevenson, M. (2015). Tools for the Analysis of Epidemiological Data.

Tufte, E. R. (2003). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 6 edition.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* New York: Springer New York.

Wilkinson, L. (2005). *The Grammar of Graphics.* Springer Science & Business Media.

Wolf, H. P. & Bielefeld, U. (2015). Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions.

Este cuaderno metodológico muestra el uso del entorno R para el análisis de datos a través de una interfaz gráfica de usuario amigable: R-Commander. Por este medio, el investigador que se inicia en el análisis de datos o busca solo una aplicación concreta hallará un software libre, semejante a otros paquetes comerciales. Quienes tengan interés por explorar la flexibilidad de este entorno, podrán ingresar de manera gradual a la programación, en vistas a operaciones más complejas que las provistas por defecto y para personalizar los análisis. El usuario experimentado hallara en R una herramienta muy versátil para aplicar diversas estrategias analíticas y de gran potencialidad gráfica. A fin de ejemplificar la aplicación de los procedimientos, se usan datos de la Encuesta Nacional de Factores de Riesgo 2013, realizada conjuntamente entre el Instituto Nacional de Estadística y Censos (INDEC) y el Ministerio de Salud de la Nación de la República Argentina.

El autor en docente e investigador en la Universidad Nacional de Córdoba. Su tema de investigación es el de las migraciones internacionales en perspectiva sociodemográfica. En docencia, se ocupa de cursos de metodología cuantitativa en carreras de grado y posgrado de Ciencias Sociales y de la Salud. Actualmente dirige la Especialización en Producción y Análisis de Información para Políticas Públicas en el Centro de Estudios Avanzados - Facultad de Ciencias Sociales - UNC.

ISBN 978-950-692-135-4