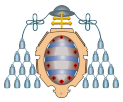


Universidad Oviedo
Dpto. Estadística e I.O. y D.M.

ESTADÍSTICA
E.P.I. Gijón
Prácticas de laboratorio

Febrero 2021

Índice general

Práctica 1

Introducción a R-Commander y Estadística descriptiva

1.1. R-Commander

Antes de nada, presentaremos el programa informático con que se realizarán las tareas de estas prácticas de laboratorio. Se trata de R-Commander, una interfaz gráfica (es decir, de ventanas, menús, etc.) al famoso entorno de computación estadística llamado R.

1.1.1. Instalación

Teniendo conexión a la red, la instalación de R-Commander en los sistemas operativos habituales resulta sencilla.

Sistema operativo Windows

Desde la página `http://knuth.uca.es/R/doku.php`, abra el enlace *Versión X.Y.Z Paquete R-UCA para windows*, de manera similar a como se muestra en la figura 1.1 (puede cambiar el número de versión).

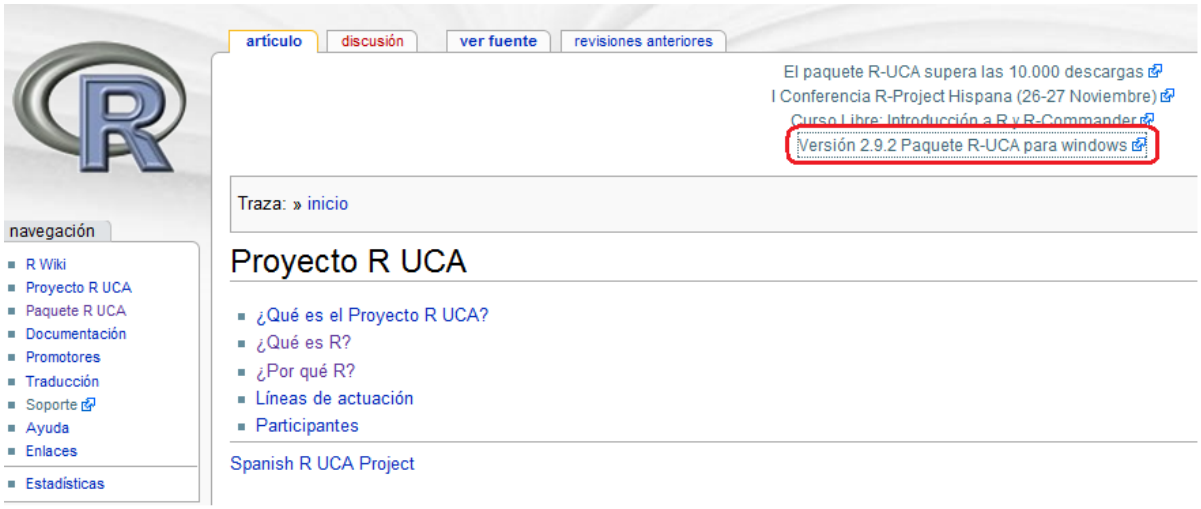


Figura 1.1: Página web del proyecto R-UCA.

Una vez que se haya descargado el paquete, ejecútelo para proceder con la instalación, que abrirá una ventana parecida a la figura 1.2.

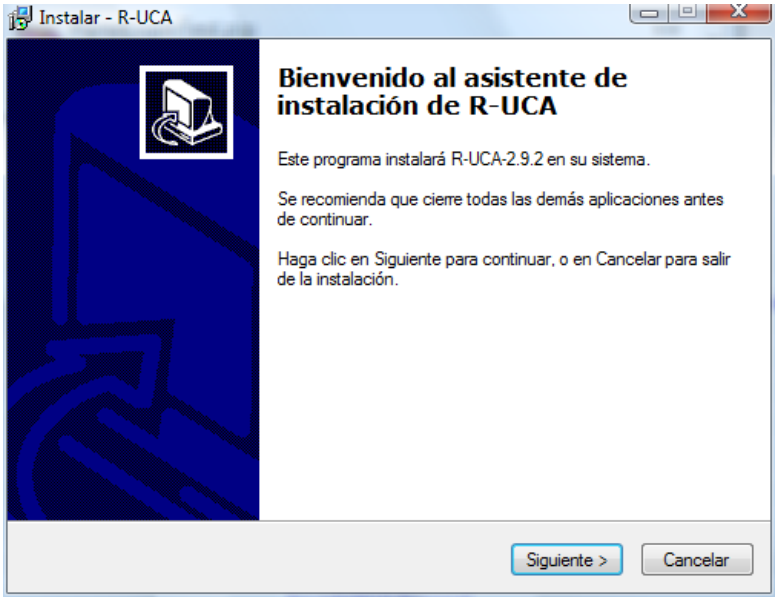


Figura 1.2: Ventana inicial de instalación de R-UCA.

Una vez completada la instalación, busque entre la lista de programas del menú *Inicio* la entrada *Rterm*. Al ejecutarlo, aparecererá una ventana de DOS que a su vez arrancará R-Commander (figura 1.3).

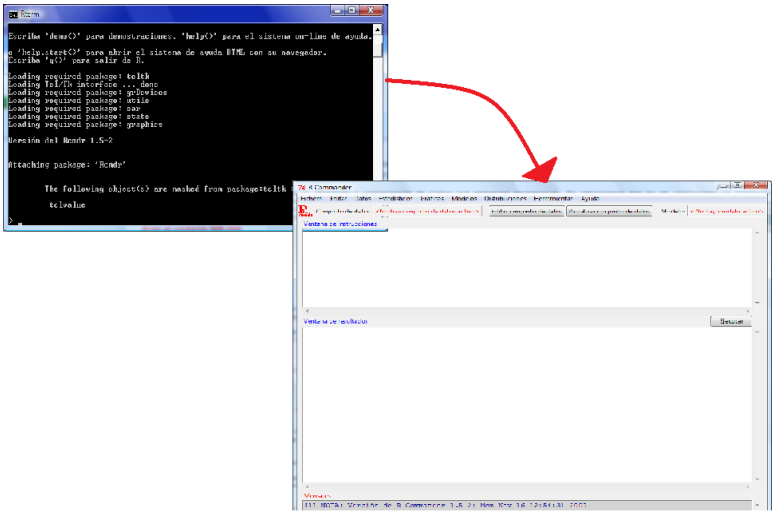


Figura 1.3: Ventanas de Rterm y R-Commander.

Con los pasos descritos anteriormente se instala la última versión disponible en R-UCA. Esta versión puede diferir de aquella que está instalada en las salas de ordenadores de la Escuela, habría que ir a http://cran.es.r-project.org/ y descargarla de ahí directamente. Es posible que en alguna versión antigua, el R-Commander, no se abra solo y haya que proceder bien sea tecleando `library(Rcmdr)`, bien sea siguiendo los siguientes pasos en el menú:

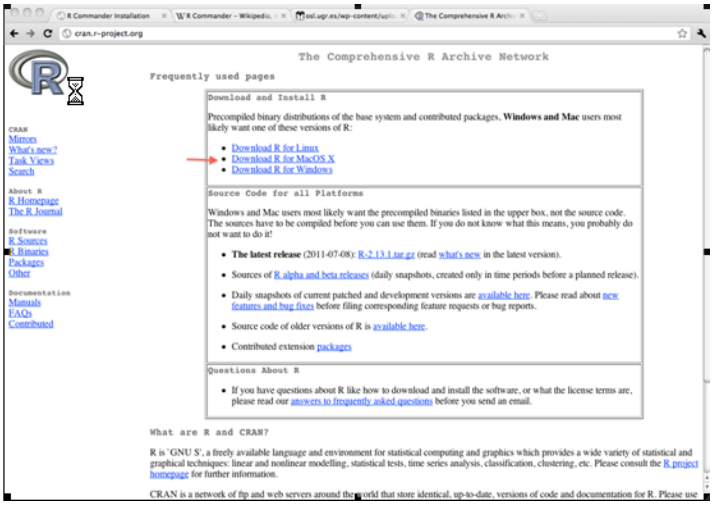
Packages
➥ Load package
➥ Rcmdr
➥ OK

Sistema operativo Ubuntu

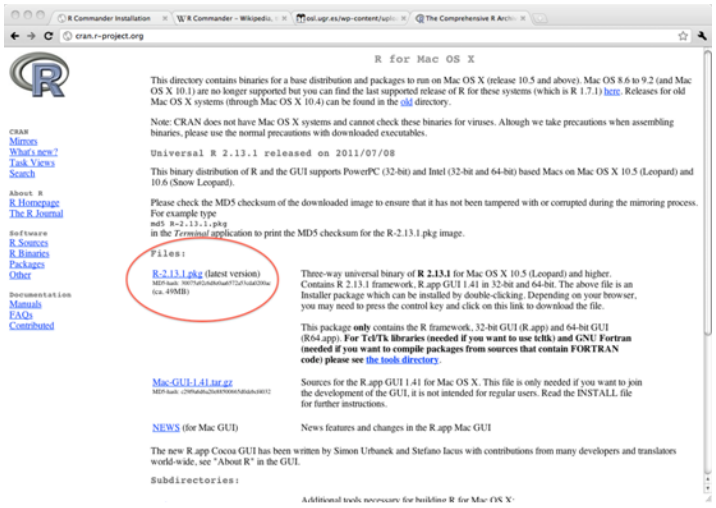
En el menú *Aplicaciones* arranque el *Centro de Software de Ubuntu*, seleccione *R-Commander* y pulse *Instalar*. El programa aparecerá en el menú *Aplicaciones*, submenú *Ciencia*.

Sistema operativo Macintosh

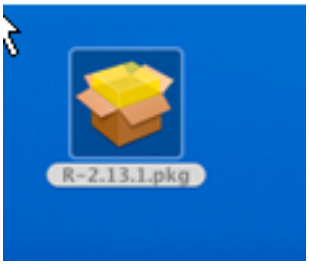
Visitamos la web http://cran.r-project.org/ y clicamos en *Download R for MacOS X*.



En la siguiente ventana descargaremos el archivo cuya extensión acaba en .pkg.
Una vez descargado procederemos a instalar el paquete: pinchamos en el archivo e instalamos el paquete.



Una vez instalado el paquete nos creará en nuestra carpeta de aplicaciones dos programas, el R, y el R-64 (siempre que nuestro procesador sea de 64 bits). Hay que instalar el Tcl/Tk que

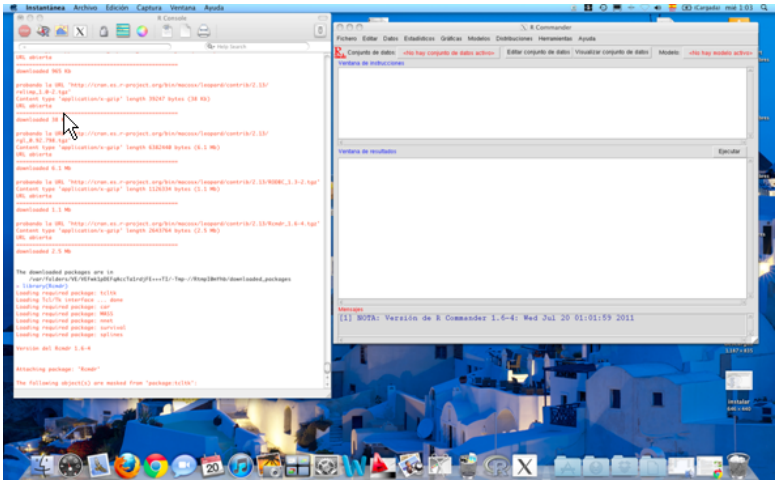


está en http://cran.r-project.org/bin/macosx/tools.
http://cran.r-project.org/bin/macosx/tools/tcltk-8.5.5-x11.dmg

Ahora tenemos que bajarnos la librería Rcmdr, que es la interfaz con la que trabajaremos en las clases.Para ello abrimos el R y escribimos lo siguiente:

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

La instalación de los paquetes es un proceso largo, no salga del programa hasta que finalice, una vez finalizada podremos trabajar con nuestro Mac sin problema. El programa es prácticamente igual que en Windows con la diferencia que para iniciar el paquete Rcmdr lo tenemos que hacer es escribir un comando en la consola: *library(Rcmdr)*.



Nota: Es necesario tener instalado un modulo denominado X11. Sino nos dará un error y no nos dejará iniciar la librería Rcmdr.

http://support.apple.com/kb/DL641

Sobre el X11 (Ayuda de Apple) : X11 está disponible en el disco de instalación de Mac

OS X, de modo que puede instalarse al mismo tiempo que Mac OS X. Para instalar X11 en un sistema que ya tiene Mac OS X instalado, inserte el disco de instalación de Mac OS X y haga doble clic en el paquete "Instalaciones opcionales". (Es posible que deba desplazarse hacia abajo para verlo.) Siga las instrucciones que se muestran en pantalla. Una vez terminada la instalación, escribimos library(Rcmdr) y vemos el Rcmdr corriendo en Mac OS.

1.1.2. Estructura

La ventana del R-Commander consta de las siguientes partes: barra de menús, barra de elementos activos (conjuntos de datos y modelos), área de instrucciones, área de resultados y área de mensajes (Fig. 1.4).

Figura 1.4: R-Commander.

1.2. Datos

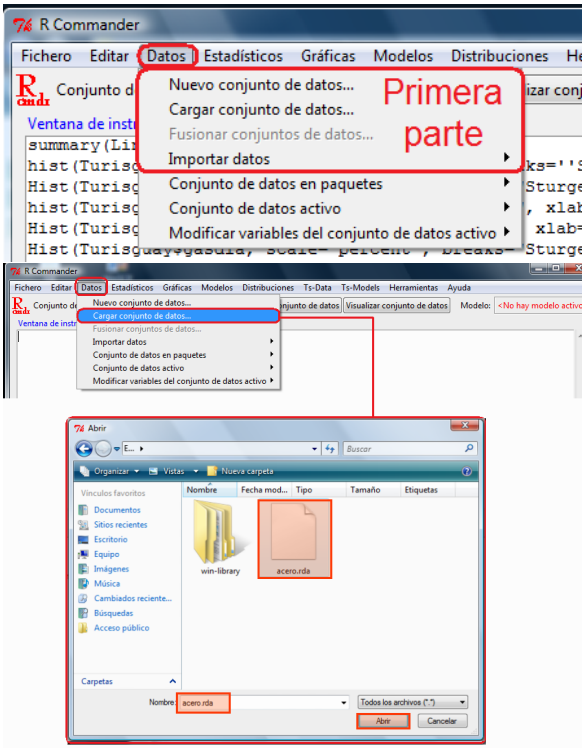
1.2.1. Banco de datos `acero.rda`

Con el fin de analizar el consumo energético de una empresa productora de acero se inspeccionó la producción de dicha empresa. La inspección duró cinco días; en los cuatro primeros, se inspeccionó cada una de las ocho horas del turno; el último día, sólo se inspeccionaron siete horas (todas menos la última). La inspección suponía registrar los valores de las variables más relevantes.

Ejemplo 1.1. *Abra el banco de datos* `acero.rda`.

Solución: Para abrir un banco de datos, accedemos al menú de *Datos*; y si deseamos trabajar con un fichero con el formato nativo de R (`.rda`), escogemos la opción *Cargar conjunto de datos*.

Datos
➥Cargar conjunto de datos



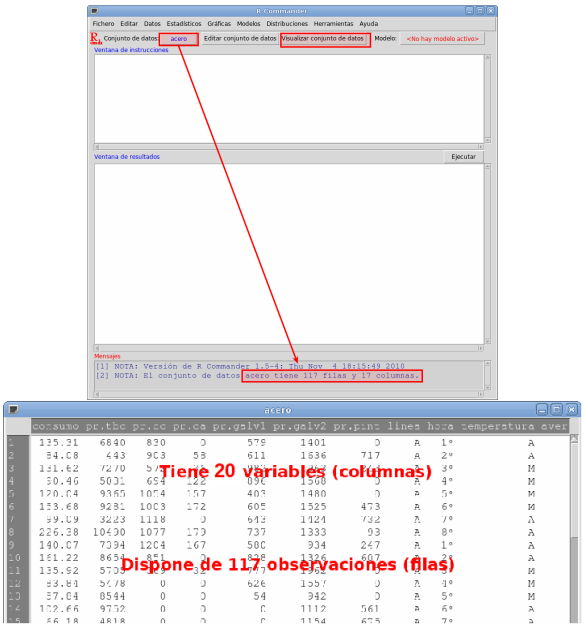
Seleccionamos el banco correspondiente: acero.rda

□

Ejemplo 1.2. *Identifique el número de variables y el número de registros del banco de datos* `acero`.

Solución: Hay varias formas de proceder. Tal vez la más sencilla consista en visualizar el banco de datos.

Conjunto de datos
➥Seleccionamos `acero` (si hubiera varios)
➥Visualizar conjunto de datos



Aparece una ventana con los datos disponibles. Moviendo el cursor hacia la izquierda o hacia abajo podemos recorrer todo el banco de datos.

En total se disponen de 117 mediciones, correspondientes a 117 horas de trabajo de la empresa, en las que se han recogido las siguientes variables:

`consumo` Consumo energético de la empresa por hora (megavatios).

`pr.tbc` Producción del tren de bandas calientes por hora (toneladas de acero).

`pr.cc` Producción de colada continua por hora (toneladas de acero).

`pr.ca` Producción del convertidor de acero por hora (toneladas de acero).

`pr.galv1` Producción de galvanizado de tipo I por hora (toneladas de acero).

`pr.galv2` Producción de galvanizado de tipo II por hora (toneladas de acero).

`pr.pint` Producción de chapa pintada por hora (toneladas de acero).

`linea` Línea de producción empleada: A o B.

`turno` Turno en la que se recogieron los datos: mañana (M), tarde (T) o noche (N).

`temperatura` Temperatura del sistema esa hora laborable: Alta, Media y Baja.

`pres.aver` Presencia de averías en esa hora: hubo averías (A) o no hubo averías (NoA).

`num.aver` Número de averías detectadas por hora.

`sistema` Activación de un sistema de detección de sobrecalentamiento en esa hora de trabajo: encendido (ON), apagado (OFF).

`ProdTotal` Producción total de la empresa por hora (toneladas de acero).

`NOx` Emisiones de mezcla de óxidos de nitrógeno por hora (toneladas).

`CO` Emisión de monóxido de carbono por hora (toneladas).

`COV` Emisión de compuestos orgánicos volátiles por hora (toneladas).

`SO2` Emisión de dióxido de azufre por hora (toneladas).

`CO2` Emisión de dióxido de carbono por hora (toneladas).

`N2O` Emisión de óxido nitroso por hora (toneladas).

□

1.2.2. Tipos de variables estadísticas

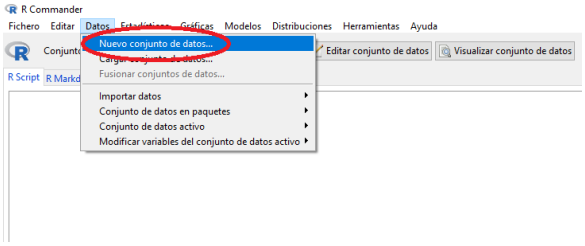
Los valores que toma una variable estadística se llaman modalidades. Si las modalidades son cantidades (números) la variable se dice cuantitativa o numérica (por ejemplo, las variables que representan alguna magnitud: velocidad, edad, tiempo, etc.); si las modalidades son nombres (etiquetas, atributos, niveles...) entonces la variable se dice cualitativa o categórica o factor.

A la hora de trabajar con una variable en R-Commander, es importante ser consciente de si se trata de una variable numérica o de un factor, pues hay procedimientos que sólo se pueden aplicar a uno de los tipos. Por ejemplo, sólo se puede hacer un gráfico de barras si la variable es factor; si es numérica, hay que convertirla antes a factor.

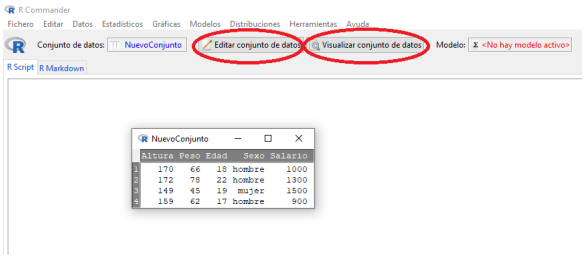
1.2.3. Otros tipos de conjuntos de datos

En RCommander también es posible crear nuestro propio conjunto de datos. Para ello, debemos seleccionar *Datos->Nuevo conjunto de datos*. Como en el ejemplo anterior con el conjunto de datos *acero*, cada columna representa una variable y cada fila un elemento de la muestra. Podemos introducir estos valores directamente rellenando la tabla.

Datos
➥Seleccionamos `Nuevo conjunto de datos`
➥Introducimos el nombre que queremos asignar al nuevo conjunto de datos

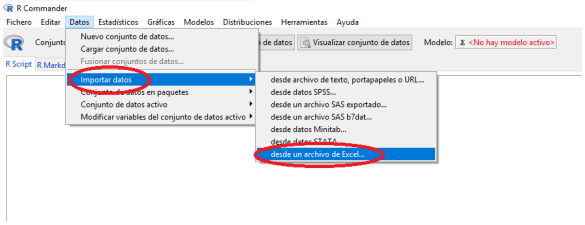


Una vez hecho esto, podemos editar y visualizar nuestro conjunto de datos seleccionando las opciones *Editar conjunto de datos* y *Visualizar conjunto de datos* en el menú.



También es posible importar conjuntos de datos de otros programas como, por ejemplo, Excel y SPSS:

Datos
➥Importar datos
➥Seleccionar el tipo adecuado



Por último, podemos guardar nuestros conjunto de datos de la siguiente manera:
Datos ->Conjunto de datos activo ->Guardar el conjunto de datos activo.

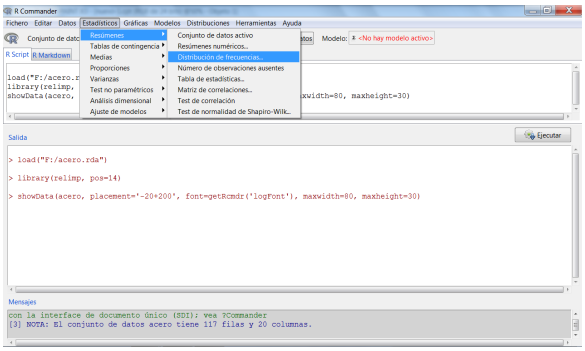
1.3. Frecuencias y porcentajes

Veamos cómo obtener las frecuencias de las modalidades de una variable.

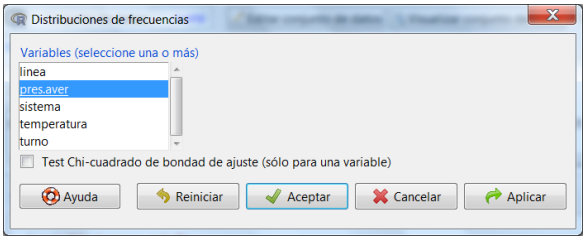
Ejemplo 1.3. *Halle la distribución de frecuencias de la variable estadística* `pres.aver`.

Solución: Procédase de la siguiente forma

Estadísticos
➥Resúmenes
➥Distribución de frecuencias



Seleccionar la variable `pres.aver`
➥Aceptar



Los pasos anteriores proporcionan el siguiente resultado:

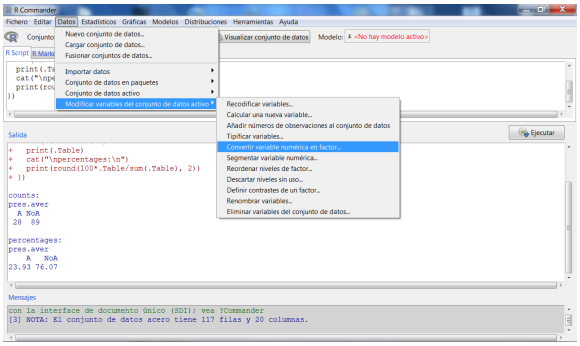
```
counts:
pres.aver
  A NoA
 28  89

percentages:
pres.aver
    A   NoA
23.93 76.07
```

Así, se han obtenido el número de casos y el porcentaje de cada modalidad dentro de la muestra. □

Ejemplo 1.4. *Halle la distribución de frecuencias de la variable estadística* `num.aver`.

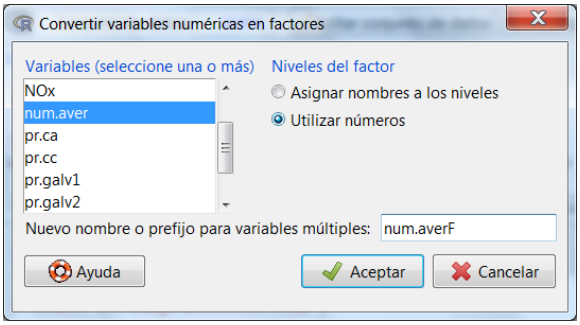
Solución: En este caso, al tratarse una variable numérica, R-Commander la considera continua y, por tanto, no nos permite realizar el cálculo de frecuencias[1]. Debemos pues, crear en primer lugar una nueva variable de tipo factor con estos datos.

Datos
➥Modificar variables del conjunto...
➥Convertir variable numérica en factor



Al convertir a factor, tenemos dos opciones. Para variables cuantitativas suele ser conveniente usar los mismos valores numéricos como etiquetas de las modalidades:

Seleccionar la variable `num.aver`
➥Utilizar números
➥Aceptar



[1]Teóricamente, la frecuencia de cada modalidad en una variable cuantitativa continua debería ser 1.

Por otro lado, téngase en cuenta que, si para *Nuevo nombre* del factor se mantiene la opción por omisión *<igual que las variables>*, se pierde el carácter cuantitativo de la variable. Esto significa que, si posteriormente se desean obtener descriptivos cuantitativos (p.ej. una media), hay que dar explícitamente un *Nuevo nombre* distinto del original. En este caso le hemos dado el nombre `num.averF`.

Una vez hecho esto, puede procederse como en el ejemplo anterior y obtener las siguientes salidas:

```
counts:
num.averF
 0  1  2  3  4
89  2  9  9  8

percentages:
num.averF
    0     1     2     3     4
76.07  1.71  7.69  7.69  6.84
```

□

1.4. Gráficas

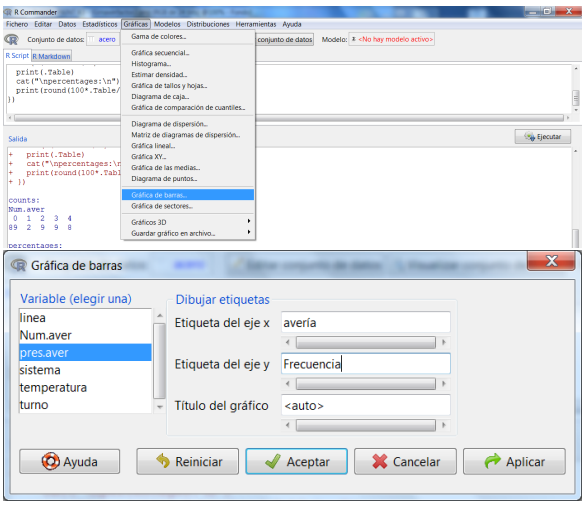
1.4.1. Diagrama de barras

Ejemplo 1.5. *Represente gráficamente la distribución de la variable* `pres.aver` *mediante una gráfica de barras.*

Solución: Se trata de una variable cualitativa, por lo que una forma adecuada de representarla gráficamente sería utilizar un diagrama de barras. Los gráficos de barras se obtienen con la opción del menú *Gráficas*; en concreto,

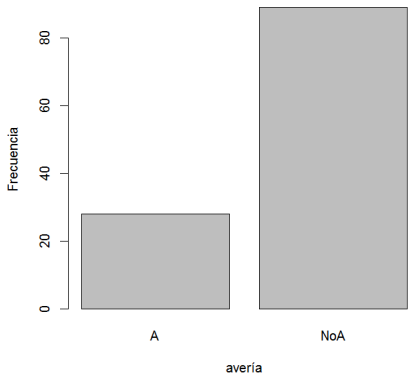
Gráficas
➥Gráfica de barras

Seleccionar la variable `pres.aver`
➥Aceptar



Como se puede observar en la ventana anterior, hemos etiquetado los ejes de coordenadas, añadido las etiquetas `avería` en `Etiqueta del eje x` y `Frecuencia` en `Etiqueta del eje y`

Con este procedimiento se obtiene el siguiente diagrama de barras:

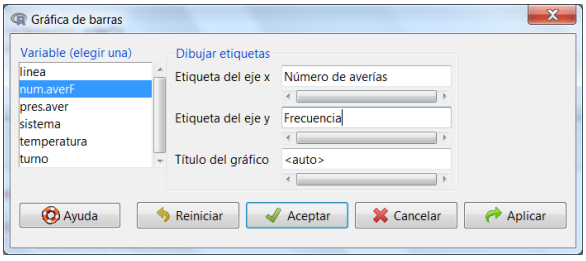


□

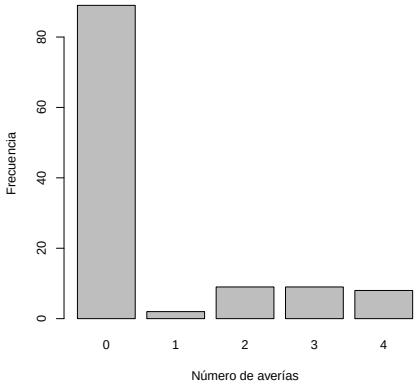
Ejemplo 1.6. *Obtenga el gráfico de barras de la variable "número de averías detectadas".*

Solución: Consideramos la variable `num.averF` que tiene los datos relativos al número de averías detectadas. Para ella realizamos el gráfico de barras como en el ejemplo 1.5:

Gráficas
➥Gráfica de barras



con lo que obtenemos un gráfico similar al siguiente:



□

1.4.2. Diagrama de sectores

Ejemplo 1.7. *Represente gráficamente la distribución de la variable* `pres.aver` *mediante un gráfico de sectores.*

Solución: Los gráficos de sectores se obtienen con la opción del menú *Gráficas*; en concreto,

Gráficas
➥Gráfica de sectores
➥Seleccionar la variable `pres.aver`
➥Aceptar

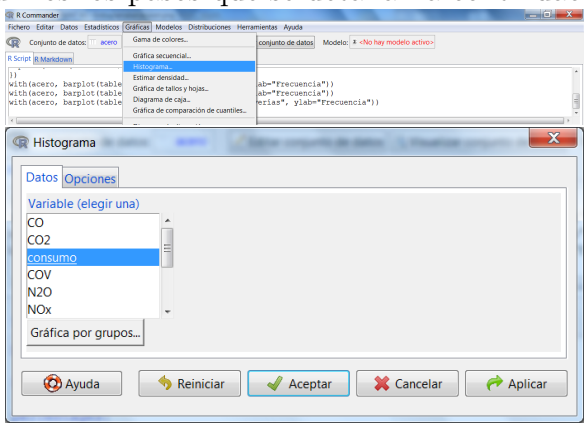
□

1.4.3. Histograma

Ejemplo 1.8. *Obtenga el histograma de la variable* `consumo`.

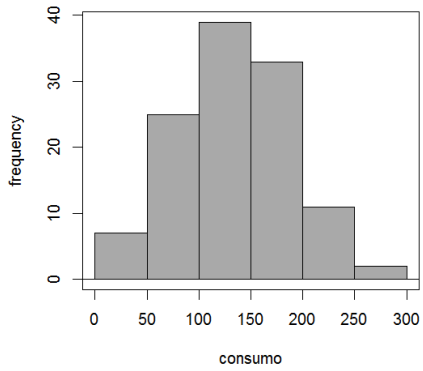
Solución: Para representar el histograma, seguimos los pasos que se detallan a continuación:

Gráficas
➥Histograma...



Seleccionar la variable `consumo`
➥Aceptar

Así se obtiene el siguiente histograma para la variable `consumo`:



□

Téngase en cuenta un par de detalles:

- Por omisión, el número de barras del histograma se calcula automáticamente. Puede indicarse un número concreto en la opción *Número de clases*; sin embargo, nótese que `R` considera ese número como una sugerencia (el cálculo del número de barras lo hace de forma de que los hitos de los ejes queden entre dos barras y sean números redondos).
- Por omisión, la altura de las barras está expresada en frecuencias. Se pueden escoger también porcentajes o densidades, en la opción *Escala de los ejes* de la pestaña *Opciones* en la ventana anterior.

1.4.4. Diagrama de caja

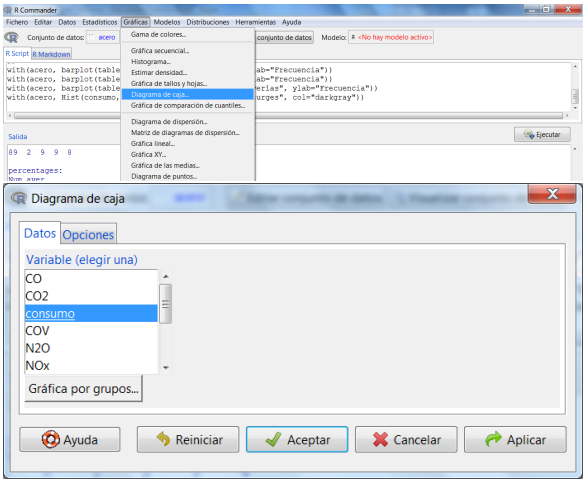
Una vez conocidas las medidas del apartado anterior, podemos abordar un último tipo de gráfica, útil para representar variables estadísticas cuantitativas, en concreto para

- detectar valores atípicos;
- comparar la distribución de una variable estadística en distintas muestras.

Ejemplo 1.9. *Obtenga el diagrama de caja de la variable* `consumo`.

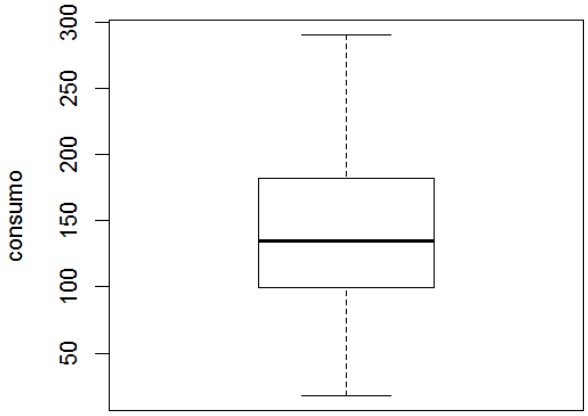
Solución: Los pasos por seguir son:

Gráficas
➥Diagrama de caja...



Seleccionar la variable `consumo`
➥Aceptar

El resultado es:



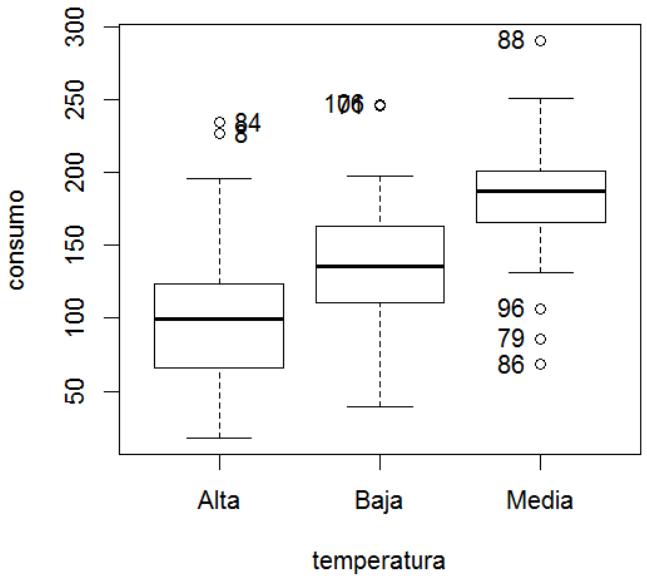
A partir de dicho diagrama se observa, por ejemplo, que no existen datos atípicos para la variable `consumo` en esta muestra.

□

Ejemplo 1.10. *Obtenga los diagramas de caja de la variable* `consumo` *para cada nivel de* `temperatura`.

Solución: Los pasos por seguir son:
Gráficas
➥Diagrama de caja...
➥Seleccionar la variable `consumo`
➥Pulsamos *Gráfica por grupos...*
➥Elegimos el factor *temperatura*
➥Aceptar
➥ El botón cambia su nombre a *Gráfica según: temperatura*
➥Aceptar

El resultado es:



A partir de dicho diagrama se ve claramente que el consumo baja con temperaturas extremas. También se ven algunos valores que podrían ser considerados atípicos, que vienen identificados por la línea que ocupan en la base de datos.

□

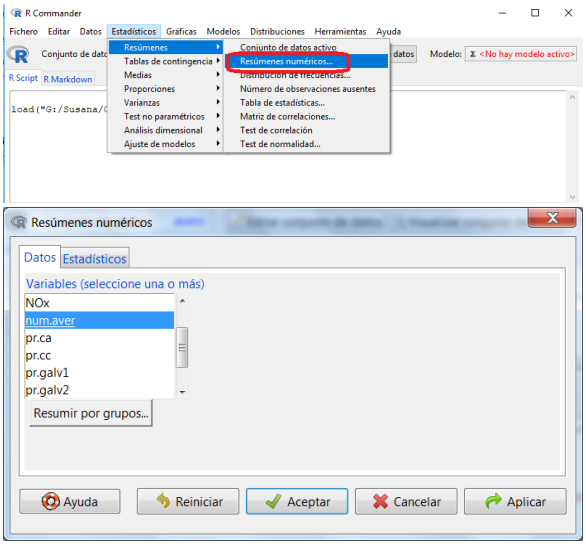
1.5. Medidas de centralización y dispersión

Como ejemplos de variables numéricas disponemos en el banco de datos de las variables `num.aver` y `consumo`. Para describir estas variables interesa obtener sus medias, sus desviaciones típicas y algunos de sus percentiles (habitualmente, los cuartiles).

Ejemplo 1.11. *Calcule la media, desviación típica y percentiles de la variable* `num.aver`.

Solución: Estos valores se obtienen de la siguiente forma:

Estadísticos
➥Resúmenes
➥Resúmenes numéricos



Seleccionar la variable `num.aver`
➥Aceptar

Las salidas del procedimiento anterior son:

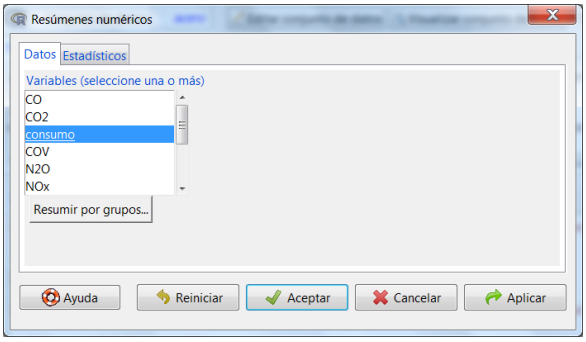
```
     mean       sd IQR 0% 25% 50% 75% 100%   n
0.6752137 1.292078   0  0   0   0   0    4 117
```

Los resultados nos indican que la media es de 0′6752137 averías por hora, con una desviación típica de 1′29. El número de averías varía desde 0 hasta 4, y al menos el 75 % de la observaciones no presentaron averías. En total disponemos de 117 observaciones. □

Ejemplo 1.12. *Calcule los principales estadísticos descriptivos de la variable* `consumo`.

Solución: Estos valores se consiguen mediante el siguiente procedimiento:

Estadísticos
➥Resúmenes
➥Resúmenes numéricos
➥Seleccionar `consumo`
➥Aceptar



con el que se obtiene:

```
     mean       sd    IQR    0%   25%   50%    75%    100%   n
 135.6771 56.90756 83.39 17.5 99.09 135.1 182.48 290.72 117
```

Con esta información podemos concluir que el consumo medio se sitúa en torno a 135′68 megavatios-hora, con una desviación típica de 56′91 MWh. El consumo mínimo desciende hasta 17′5 y el máximo asciende hasta 290′72. El 25 % de los casos analizados consumen 99′09 MWh o menos; el 50 %, menos de 135′1; y un 25 % consume más de 182′48.

□

1.6. Generar nuevas variables

A partir de una variable numérica se pueden obtener nuevas variables, tanto numéricas como de texto. Vamos a ver como se haría en ambos casos:

1.6.1. Calcular una nueva variable

Las variables del fichero de datos se pueden manipular numéricamente a través de la opción del menú

> Datos
> ➥ Modificar variables del conjunto de datos activo
> ➥ Calcular una nueva variable...

que nos permite calcular una nueva variable, a partir de una ya existente.

Las notaciones para las operaciones elementales en *Expresión a calcular* son las habituales: + − * / y ^.

Así, por ejemplo, si queremos generar la variable `coste` a partir del `consumo`, si la relación entre ambas fuese $coste = 2'34 \cdot consumo$, deberíamos utilizar la expresión `2.34*consumo`. El nombre de la variable que se está transformando se puede teclear o simplemente pasarlo haciendo doble click sobre ella en la lista de *Variables actuales*.

Es importante comentar, porque puede dar lugar a error, que si introducimos números decimales en *Expresión a calcular*, tal como hemos hecho en el ejemplo anterior, el separador decimal es el punto.

1.6.2. Recodificar variables

La opción del menú

> Datos
> ➥ Modificar variables del conjunto de datos activo
> ➥ Recodificar variables...

nos permite crear una nueva variable, generalmente es útil para crear una variable discreta a partir de una continua.

Las expresiones que irán en *Introducir directrices de recodificación* son:

- un valor simple: `"Alta"=1`
- varios valores separados por comas: `7,8,9="alto"`
- un rango de valores indicados por dos puntos: `7:9="alto"`. El caso especial de los valores no acotados `lo` (lowest) y `hi` (highest) son admitidos.
- el comando especial `else`, con el que hay que tener cuidado en su utilización, puesto que es aplicable en cualquier caso que no sean los anteriores, incluso si la celda está en blanco.

Así, por ejemplo, la recodificación de la variable `consumo` en la variable `Grupoconsumo` siguiente:

$$\text{Grupoconsumo} = \begin{cases} Bajo & \text{si consumo } \leq 100 \\ Medio & \text{si } 100 < \text{ consumo } \leq 200 \\ Alto & \text{si consumo } > 200 \end{cases}$$

se haría con los comandos:

```
lo:100="Bajo"
100:200="Medio"
200:hi="Alto"
```

Se pueden recodificar, con el mismo criterio, varias variables a la vez, sin más que seleccionarlas todas juntas.

Un caso particularmente interesante de recodificación es el de la dicotomización de una variable numérica. Vamos a verlo en el siguiente ejemplo.

Ejemplo 1.13. *Definir una nueva variable que llamaremos* `Produccion` *que tome el valor* `Fracaso` *si la producción total fue de como mucho 10000 toneladas y el valor* `Exito` *si fue mayor.*

Solución: Para ello simplemente deberíamos repetir los pasos del ejemplo anterior, pero tecleando

```
lo:10000="Fracaso"
10000:hi="Exito"
```

en `Introducir directrices de recodificación` y `Produccion` en `Nuevo nombre o prefijo para variables múltiples recodificadas:` dentro de la ventana Recodificar variables. □

1.7. Filtros

Al igual que acabamos de ver para el diagrama de caja, se podrían obtener resúmenes numéricos de una variable numérica para cada nivel de otra variable factor. No obstante, en ocasiones nos conviene analizar solamente una parte de los datos, que cumplen determinada característica. Vamos a ver en este epígrafe como se realizaría dicho filtrado, poniendo como ejemplo el análisis de una variable estadística condicionada.

Ejemplo 1.14. *Halle la distribución de frecuencias de la variable estadística* `pres.aver` *para aquellos casos en que la temperatura es alta.*

Solución: Para filtrar los datos según la condición enunciada, usamos la ruta
La opción del menú

Datos
➥ Conjunto de datos activo
➥ Filtrar el conjunto de datos activo...

Aparece una ventana titulada *Filtrar el conjunto de datos*. La parte superior de la misma permite escoger columnas (es decir, variables) concretas; lo habitual será que nos interese dejar seleccionada la opción por omisión: *Incluir todas las variables*.

Expresión de selección. En nuestro caso, el criterio ha de escribirse como sigue:

```
temperatura=="Alta"
```

Nombre del nuevo conjunto de datos. Por omisión aparece <*igual que en el conjunto de datos activos*>. Conviene cambiarlo a otro nombre, porque si lo dejamos así el resultado del filtro sustituye al conjunto original, y los datos originales se pierden. En nuestro caso escogeremos, por ejemplo, `acero.temp_alta`. Nótese que el nombre puede usar letras, cifras, puntos y subrayados (guiones bajos), pero no espacios, ni guiones, ni muchos otros.

Tras cerciorarse que en la barra de elementos activos aparece

Conjunto de datos: `acero.temp_alta`

procédase como en el ejemplo 1.3 (pág. 10) para obtener

```
counts:
pres.aver
  A NoA
  8  38

percentages:
pres.aver
    A   NoA
17.39 82.61
```

□

Puede que le haya llamado la atención la expresión por usar dos símbolos igual (==) en lugar de uno. El siguiente cuadro muestra las correspondencias entre expresiones habituales a la hora de establecer una condición y los símbolos que usa R-Commander:

igual (=)	`==`
distinto (≠)	`!=`
menor o igual (≤)	`<=`
mayor o igual (≥)	`>=`
conjunción (y)	`&`
disyunción (o)	`|`

También se pueden usar paréntesis para agrupar.

Por otro lado, es importante destacar que al usar

- un valor numérico, emplee punto decimal (no coma);
- un valor textual, entrecomíllelo con el símbolo " o el apóstrofo '.

Advertencia: Es habitual olvidarse de que se ha establecido un filtro, y realizar subsiguientes cálculos con el conjunto filtrado sin pretenderlo. Recuerde volver al conjunto de datos original pinchando el botón *Conjunto de datos* en la barra de elementos activos de R-Commander.

1.8. Ejercicios propuestos

Ejercicio 1.1. *Represente la variable* `num.aver` *mediante un diagrama de sectores. ¿Es adecuado el gráfico?*

Ejercicio 1.2. *¿Qué gráfico es apropiado para representar la producción de tren de bandas en caliente, un diagrama de barras o un histograma? Represéntelo usando porcentajes en el eje de ordenadas.*

Ejercicio 1.3. *Observe la distribución de la producción de colada continua. ¿Cuánto vale la producción media? ¿Cuánto vale la mediana?*

Ejercicio 1.4. *¿En qué tipo de producción se alcanza la máxima producción por hora?*

Ejercicio 1.5. *¿Qué gráfico es apropiado para detectar valores atípicos en la emisión de dióxido de azufre? Represéntelo. ¿Qué observaciones son atípicas? ¿Cuánto vale la emisión para estas observaciones atípicas?*

Ejercicio 1.6. *a) ¿Cuántos de los 117 datos fueron tomados a temperatura media?, ¿y a temperatura alta?*

b) ¿Qué porcentaje de los datos se han tomado a temperatura baja?

c) ¿Qué porcentaje no se tomaron a temperatura alta?

d) ¿Cuál ha sido la temperatura más frecuente durante la toma de estos datos? ¿y la menos frecuente?

e) ¿Se pueden representar los datos de la temperatura mediante un diagrama de barras? ¿y mediante un diagrama de sectores? ¿y mediante un histograma?

f) Realice un diagrama de sectores para la variable temperatura, tal que el título del gráfico sea "Temperatura del sistema".

Ejercicio 1.7. *Represente gráficamente los datos tomados respecto al consumo energético de la empresa, de forma que en el eje de ordenadas (Y) aparezcan representados los porcentajes y los ejes lleven las etiquetas "Consumo" y "%", según corresponda. En función de dicho gráfico responda a las siguientes preguntas:*

a) Si este estudio se ha llevado a cabo en la empresa para determinar si se están cumpliendo los objetivos de que el consumo no sobrepase los 400 megavatios-hora, ¿apoyan estos datos dicha hipótesis? Si el objetivo fuese que el consumo se mantuviese por debajo de los 200 megavatios-hora, ¿se estarían alcanzando los objetivos de la empresa?

b) Según estos datos, ¿es creíble suponer que aproximadamente el 40 % del tiempo el consumo es mayor de 250 megavatios-hora?

c) Considerando que estos datos se han tomado con un proceso estable y, por tanto, representan el comportamiento habitual de esta empresa, ¿alrededor de qué valores se sitúa más frecuentemente el consumo energético de esta empresa?

Ejercicio 1.8. *Realice un diagrama de barras para la variable* `consumo` *y comente dicho gráfico.*

Ejercicio 1.9. *El diseño del experimento sugería que se tomasen aproximadamente la misma cantidad de datos con el sistema de detección de sobrecalentamiento encendido que con él apagado. ¿Se ha trabajado de acuerdo con dicho diseño? ¿Cuántos datos se tomaron con el sistema apagado?, ¿qué porcentaje supone del número total de datos analizados?*

Ejercicio 1.10. *Realice un gráfico para el consumo energético de esta empresa en el que aparezcan dos diagramas de caja, uno para el consumo cuando el sistema de detección de sobrecalentamiento está encendido y otro cuando está apagado. Comente dicho gráfico. ¿Cuánto vale el consumo medio en cada uno de estos dos casos?*

Ejercicio 1.11. *Si la producción de un séptimo producto X se puede obtener como la diferencia entre la producción total y la suma de las seis producciones dadas (tren de bandas calientes, colada continua, convertidor de acero, galvanizado tipo I, galvanizado tipo II y chapa pintada). ¿Cuánto vale la producción media del producto X?*

Ejercicio 1.12. *Si definimos la variable* `Grupoprod` *como sigue:*

$$Grupoprod = \begin{cases} Moderada & si\ ProdTotal \leq 8000 \\ Adecuada & en\ otro\ caso \end{cases}$$

¿qué porcentaje de veces la producción ha sido moderada?

Ejercicio 1.13. *Para analizar el comportamiento del sistema de detección de sobrecalentamiento, se consideran sólo los datos cuándo éste ha estado encendido. En tal caso, se pide:*

a) ¿Cuál es el número medio de averías que se han producido?

b) ¿Qué número de averías deja la mitad de los datos por encima de él y la mitad por debajo?

c) ¿Cuál ha sido el número de averías más frecuente?

d) ¿Qué porcentaje de veces se tomaron datos de la línea A?

e) Realice un gráfico que representa la línea de producción empleada y comente dicho gráfico.

f) Realice un gráfico que representa el número de averías detectadas y comente dicho gráfico.

g) Realice un gráfico que representa la producción de chapa pintada y comente dicho gráfico.

Ejercicio 1.14. *Del total de la muestra, ¿puede asegurarse que en menos de un 25 % de los datos el consumo ha sido mayor de 150?*

Ejercicio 1.15. *Calcula la media, mediana, rango, desviación típica y varianza de las siguientes variables en los casos, en los que sea posible, y comenta dichos resultados:*

a) Presencia de averías.

b) Número de averías detectadas.

c) Producción del convertidor de acero.

1.9. Soluciones de los ejercicios propuestos

Ejercicio 1.1. *Recuerde que ha de convertir la variable a factor, previamente. Es más adecuado un gráfico de barras en este caso, porque las modalidades de la variable son cantidades, y la relación de orden entre ellas queda diluida por la representación circular.*

Ejercicio 1.2. *Histograma.*

Ejercicio 1.3. *Realizamos el histograma para observar gráficamente la distribución de esta variable. Por otro lado, calculamos los principales estadísticos descriptivos de la variable* `pr.cc` *al seleccionar la opción del menú:* Resúmenes numéricos, *cuyos resultados son:*

```
    mean       sd IQR 0% 25% 50% 75% 100%   n
433.9316 276.8536 406 33 201 380 607 1204 117
```

Ejercicio 1.4. *Calculamos los descriptivos de las diferentes producciones:*

```
              mean        sd  IQR 0%  25%  50%  75%  100%   n
pr.ca     244.9231  167.5311  234 13   99  225  333   677 117
pr.cc     433.9316  276.8536  406 33  201  380  607  1204 117
pr.galv1  440.4530  235.8312  392 19  245  432  637   982 117
pr.galv2 1173.2222  511.7398  654 13  902 1333 1556  1963 117
pr.pint   349.6923  245.1241  423 20  135  270  558   908 117
pr.tbc   6916.6667 3017.5123 4451 22 4882 8062 9333 10955 117
```

*El tren de bandas en caliente alcanza la máxima producción por hora (*10955 *toneladas).*

Ejercicio 1.5. *Diagramas de caja. Hay varias observaciones atípicas. Claramente lo son las de las filas 68 y 85, pero el resto no se pueden ver con claridad. Si vamos a la ventana* Salida *observamos que las filas con observaciones atípicas son: 85, 87, 106 y 68. La emisión para ellas es de* 0′014, 0′002, 0′001 *y* 0′127, *respectivamente.*

Ejercicio 1.6. *Para responder a las cuatro primeras cuestiones obtenemos la distribución de frecuencias de la temperatura:*

```
> .Table  # counts for temperatura
 Alta  Baja Media
   46    38    33

> round(100*.Table/sum(.Table), 2)  # percentages for temperatura
 Alta  Baja Media
39.32 32.48 28.21
```

a) *33 a temperatura media y 46 a temperatura alta.*

b) 32′48 %

c) 60′68 %

d) *La más frecuente la temperatura alta y la menos la media.*

e) *Sí, se pueden representar mediante un diagrama de barras porque es una variable cualitativa. Por la misma razón, también se pueden representar mediante un diagrama de sectores. Mediante un histograma no, puesto que ni siquiera es numérica.*

f) *Simplemente sería necesario escribir en la etiqueta* Título del gráfico *de la ventana que se abre al seleccionar la opción del menú* `Gráficas` → `Gráfica de sectores` *el texto* Temperatura del sistema.

Ejercicio 1.7. *Para realizar dicho gráfico ir a la opción del menú* `Gráficas` → `Histograma`. *Una vez ahí, seleccionar la variable* `consumo` *en la pestaña* `Datos` *y seleccionar* `Porcentajes` *en* `Escala de los ejes` *dentro de la pestaña* `Opciones`. *Además debemos rellenar las opciones de etiquetas como corresponde a los requerimientos del enunciado.*

a) *Sí. No.*

b) *No.*

c) *Alrededor de 125 megavatios-hora.*

Ejercicio 1.8. *El programa no permite realizar dicho gráfico al tratarse el consumo de una variable numérica. La pasamos a factor tal como vimos en el guión (Datos → Modificar variables del conjunto → Convertir variable numérica en factor...) y ya podemos hacer el correspondiente diagrama de barras. En él no se ve nada, puesto que casi todas las frecuencias son 1 por ser continua. No es un gráfico adecuado para este tipo de variable.*

Ejercicio 1.9. *Sí, puesto que más o menos se tomaron datos en la misma proporción con el sistema encendido y apagado (50'43 % apagado y 49'57 % encendido) Con el sistema apagado se tomaron 59 datos de 117, lo que supone un 50'43 % de los datos.*

Ejercicio 1.10. *Gráfico de caja: parece que hay menos consumo con el sistema encendido. El consumo medio con el sistema encendido es* 124′24 *y con él apagado* 146′92.

Ejercicio 1.11. *En primer lugar calculamos la nueva variable* `prodX` *y luego calculamos su media, obteniendo que la producción media del producto X es de* 2287′556 *toneladas.*

Ejercicio 1.12. 17′09 %

Ejercicio 1.13. *a*) *Media*= 0′637931

b) *Mediana=0*

c) *Moda=0 (con un* 79′31 % *de los datos; 46 veces)*

d) 51′72 %

e) *Serviría el gráfico de barras o de sectores (variable cualitativa). Se observa que más o menos se han tomado datos en la misma proporción para las dos líneas.*

f) *Serviría el gráfico de barras o de sectores (variable cuantitativa discreta). Casi siempre 0 averías, nunca 1 y luego 2, 3 o 4 en una proporción similar.*

g) Serviría el histograma o el diagrama de caja (variable cuantitativa continua). En ellos se observa como la mayor parte de las veces la producción de chapa pintada está entre 0 y 300 toneladas, no habiendo sido nunca mayor de 900 toneladas.

Ejercicio 1.14. *No, puesto que el percentil 75 es* $182'48$*, con lo que mayores de 150 son al menos el* $25\,\%$*.*

Ejercicio 1.15. *a) Al ser cualitativa nominal, no tiene sentido calcular ninguna de las medidas propuestas.*

b) Al ser cuantitativa discreta se pueden calcular todos: $\overline{x} = 0'675$*,* $Me = 0$*,* $R = 4$*,* $s = 1'29$ *y* $s^2 = 1'29^2$*.*

c) Al ser cuantitativa continua, las medidas pedidas son: $\overline{x} = 244'92$*,* $Me = 225$*,* $R = 664$*,* $s = 167'53$ *y* $s^2 = 167'53^2$*.*

1.10. Notas

Opciones del diagrama de barras

Diagrama de barras de porcentajes en lugar de frecuencias

```
barplot(100*table(acero$temperatura)/sum(table(acero$temperatura)),
    xlab="temperatura", ylab="Porcentaje")
```

Cambiar color de las columnas

```
barplot(table(acero$temperatura), xlab="temperatura", ylab="Frequency",
    col="green")
```

Etiquetar barras con frecuencias

```
dibujo <- barplot(table(acero$temperatura), xlab="temperatura",
    ylab="Frequency") # guardamos el grafico que sale
numeros <- table(acero$temperatura) # numeros a dibujar
text(dibujo, numeros + 2, format(numeros, digits=2), xpd = TRUE)
```

Etiquetar barras con porcentajes

```
dibujo <- barplot(table(acero$temperatura), xlab="Temperatura",
    ylab="Porcentaje") # guardamos el grafico que sale
numeros <- 100*table(acero$temperatura)/sum(table(acero$temperatura))
    # porcentajes a dibujar
text(dibujo, numeros + 2, format(numeros, digits=2.2), xpd = TRUE)
```

Opciones del diagrama de sectores

Diagrama de sectores con etiquetas

```
pie(tabla, labels=paste(levels(acero$temperatura),table(acero$temperatura)),
    main="temperatura", col=rainbow_hcl(length(levels(acero$temperatura))))
```

Diagrama de sectores de porcentajes en lugar de frecuencias y con etiquetas

```
tabla<-round(table(acero$temperatura)*100/sum(table(acero$temperatura))
    *100+0.5)/100
pie(tabla, labels=paste(levels(acero$temperatura),tabla,"%"), main="temperatura",
    col=rainbow_hcl(length(levels(acero$temperatura))))
```

Ver todos los gráficos realizados

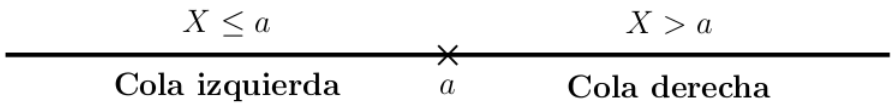
Para poder ir para adelante y atrás con las teclas `Re Pág` y `Av Pág` es necesario cuando se haga el primer gráfico de la sesión ir a la opción del menú *Histórico* → *Grabando*. Hay que repetir la operación cada vez que se abre `R`.

Práctica 2

Modelos de distribuciones

El menú *Distribuciones* de R-Commander contiene un amplio conjunto de distribuciones de probabilidad, agrupadas en discretas y continuas. Para cada una de ellas, hay cuatro posibilidades:

Cuantiles. Es el menor valor c tal que $\Pr[X \leq c] \geq p$ (cola inferior o izquierda) o que $\Pr[X > c] \leq p$ (cola superior o derecha).



Probabilidades. En variables discretas, da los valores de la función de probabilidad, es decir, $\Pr[X = k]$ para cierto k. En variables continuas, da la probabilidad acumulada (véase la siguiente entrada).

Probabilidades acumuladas. Dado un valor k, calcula la probabilidad $\Pr[X \leq k]$ (cola izquierda) o $\Pr[X > k]$ (cola derecha).

Gráfica. Dibuja la gráfica de la función de densidad (para variables continuas) o de probabilidad (para discretas), o bien la función de distribución.

Muestra. Permite generar un nuevo conjunto de datos aleatorio indicando los parámetros de la distribución y las cantidades de filas y columnas deseadas. Dos advertencias:

a) En algunas versiones de R-Commander, en *Introducir el nombre del conjunto de datos* aparece por omisión un nombre con un espacio en blanco. En general, los objetos de R-Commander no permiten nombres con espacios, por lo que si el usuario no lo cambia se producirá un error.

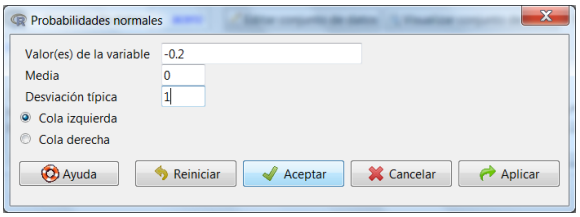
b) Según la ventana de diálogo, las filas corresponden a muestras y las columnas a observaciones. El usuario puede estar interesado en hacerlo al revés, de forma que las filas se correspondan con los individuos de cada muestra (la interpretación habitual). En tal caso, simplemente olvídese de las etiquetas *muestras* y *observaciones* y piense sólo en filas y columnas.

2.1. Modelos de distribuciones continuas

Ejemplo 2.1. *Sea Z la distribución normal $N(0,1)$. Halle $P(Z \leq -0'2)$.*

Solución: Siga la ruta:

↪ Distribuciones
↪ Distribuciones continuas
↪ Distribución normal
↪ Probabilidades normales...
↪ Valor(es) de la variable `-0.2`
↪ Media `0` (por omisión)
↪ Desviación típica `1` (p.o.)
↪ Cola izquierda (elegida)
↪ Aceptar



El resultado es $0'4207403$.

□

Ejemplo 2.2. *Sea X una distribución normal $N(5,2)$. Calcule:*

a) $P(X \leq 7)$

b) $P(X = 7)$

c) $P(X < 7)$

d) $P(X > 7)$

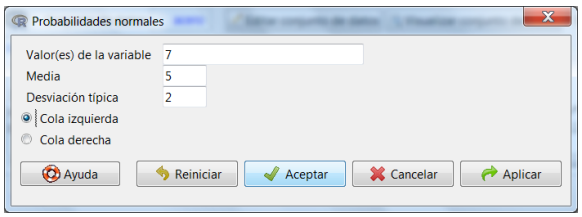
e) $P(X \geq 7)$

f) $P(4 < X < 7)$

g) $P(4 \leq X \leq 7)$

Solución: a) Siga la ruta:

↪ Distribuciones
↪ Distribuciones continuas
↪ Distribución normal
↪ Probabilidades normales...
↪ Valor(es) de la variable `7`
↪ Media `5`
↪ Desviación típica `2`
↪ Cola izquierda (elegida)
↪ Aceptar



El resultado es $0'8413447$.

b) En una distribución continua cada punto aislado tiene probabilidad cero, con lo que, sin necesidad del ordenador se puede concluir que $P(X = 7) = 0$.

c) En este caso el resultado es el mismo que en el primer apartado, puesto que en una distribución continua, como acabamos de comentar, no influye el añadir o no un punto.

d) El único cambio respecto al primer apartado es que habría que poner *Cola derecha*. Con esto se obtendría que la probabilidad pedida es 0′1586553. También se podría obtener restando el resultado del primer apartado a 1, puesto que $P(X > 7) = 1 - P(X \leq 7)$.

e) Coincide con la probabilidad del apartado anterior, al tratarse de una distribución continua, puesto que $P(X > 7) = P(X \geq 7)$.

f) En este caso tenemos varias formas de obtener la probabilidad pedida. Una de ellas seria considerar que

$$P(4 < X < 7) = P(X < 7) - P(X \leq 4).$$

El primer valor ya se obtuvo en el primer apartado y de forma similar (cambiando el 7 por un 4) se obtendría que $P(X \leq 4) = 0'3085375$, con lo que

$$P(4 < X < 7) = 0'8413447 - 0'3085375 = 0'5328072$$

Esta operación se puede hacer directamente en `R`, sin más que teclear

```
0.8413447-0.3085375
```

en la ventana `R Script` y pinchar en el botón `Ejecutar`.

g) De nuevo, al tratarse de una distribución continua se tiene que

$$P(4 \leq X \leq 7) = P(4 < X < 7)$$

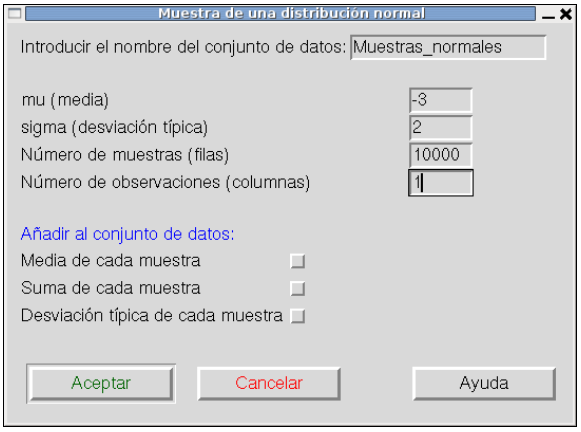
valor que ya ha sido calculado previamente.

□

Ejemplo 2.3. *Dibuja un histograma de una muestra aleatoria de* 10 000 *valores extraídos de una población normal* ($\mu = -3; \sigma = 2$). *Para el histograma, usa aproximadamente* 50 *barras.*

Solución: Siga la ruta:

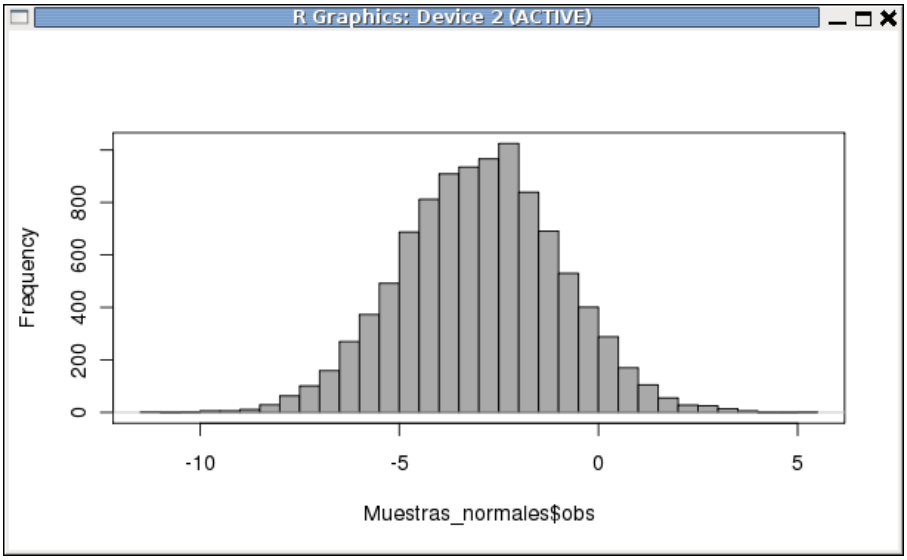
↪ Distribuciones
↪ Distribuciones continuas
↪ Distribución normal
↪ Muestra de una distribución normal...
↪ Dejar el nombre del conjunto de datos: `Muestrasnormales`
↪ mu (media): `-3`
↪ sigma (desviación típica): `2`
↪ Número de muestras (filas): `10000`
↪ Número de observaciones (columnas): `1`
↪ Media de cada muestra: desactívese
↪ Suma de cada muestra: desactivada
↪ Desviación típica de cada muestra: desactivada
↪ Aceptar



Ahora queda realizar el histograma:

↪ Gráficas
↪ Histograma...
↪ Variable (eligir una): `obs`
↪ Número de clases: `50`
↪ Aceptar

Aparecerá un gráfico similar a lo siguiente,

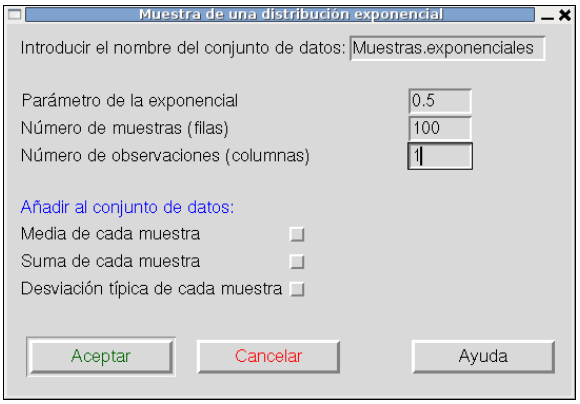


pero no idéntico, dado que se tratan de números pseudoaleatorios. □

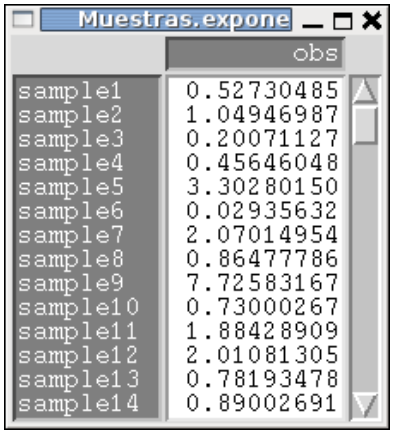
Ejemplo 2.4. *Genera* 100 *valores aleatorios de una distribución exponencial con media* 2.

Solución: Siga la ruta:

↪ Distribuciones
↪ Distribuciones continuas
↪ Distribución exponencial
↪ Muestra de una distribución exponencial...
↪ Dejar el nombre del conjunto de datos: `MuestrasExponenciales`
↪ Parámetro de la exponencial: `0.5`
↪ Número de muestras (filas): `100`
↪ Número de observaciones (columnas): `1`
↪ Media de cada muestra: desactívese
↪ Suma de cada muestra: desactivada
↪ Desviación típica de cada muestra: desactivada
↪ Aceptar



Si pulsa ahora *Visualizar conjunto de datos* obtendrá algo como lo siguiente,



pero no idéntico, dado que se tratan de números pseudoaleatorios. □

Ejemplo 2.5. *Sea X una distribución exponencial exp(0′1). Calcule:*

a) $P(X \leq 7)$

b) $P(X = 7)$

c) $P(X < 7)$

d) $P(X > 7)$

e) $P(X \geq 7)$

f) $P(4 < X < 7)$

g) $P(4 \leq X \leq 7)$

Solución: a) Siga la ruta:

↪ Distribuciones
↪ Distribuciones continuas
↪ Distribución exponencial
↪ Probabilidades exponenciales...
↪ Valor(es) de la variable `7`
↪ Parámetro de la exponencial `0.1`
↪ Cola izquierda (elegida)
↪ Aceptar

El resultado es 0′5034147.

b) Al ser la exponencial una distribución continua se tiene que $P(X = 7) = 0$.

c) Al ser la exponencial una distribución continua se tiene que $P(X \leq 7) = P(X < 7)$, con lo que $P(X < 7) = 0'5034147$.

d) Bien eligiendo la opción de *Cola derecha* o bien aplicando que $P(X > 7) = 1 - P(X \leq 7)$, se obtendría que $P(X > 7) = 0'4965853$.

e) Al tratarse de una distribución continua, se tiene que $P(X \geq 7) = P(X > 7) = 0'4965853$.

f) Puesto que $P(4 < X < 7) = P(X < 7) - P(X \leq 4)$, $P(X < 7) = 0'5034147$ y se puede obtener que $P(X \leq 4) = 0'32968$, con lo que $P(4 < X < 7) = 0'1737347$.

g) De nuevo, al tratarse de una distribución continua se tiene que $P(4 \leq X \leq 7) = P(4 < X < 7) = 0'1737347$.

□

Ejemplo 2.6. *De forma análoga podría trabajarse en el caso de una distribución Weibull* $W(2, 3)$. *En ese caso se obtendría que*

$$P(X \leq 7) = P(X < 7) = 0'9956798$$
$$P(X > 7) = P(X \geq 7) = 0'004320239$$
$$P(4 < X < 7) = P(4 \leq X \leq 7) = P(4 < X \leq 7) = P(4 \leq X < 7) = 0'1646931$$

2.2. Modelos de distribuciones discretas

Ejemplo 2.7. *Sea* X *una distribución binomial con parámetros* $n = 10$ *y* $p = 0'4$. *Calcule:*

a) $P(X \leq 7)$

b) $P(X = 7)$

c) $P(X < 7)$

d) $P(X > 7)$

e) $P(X \geq 7)$

f) $P(4 < X < 7)$

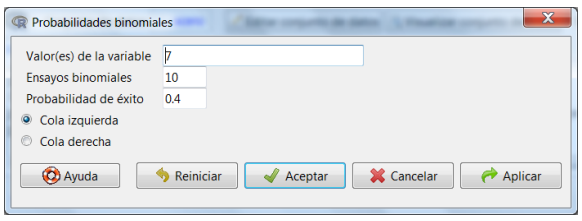
g) $P(4 \leq X \leq 7)$

h) $P(X = 2'3)$

i) $P(X = 25)$

Solución: a) Siga la ruta:

↪ Distribuciones
↪ Distribuciones discretas
↪ Distribución binomial
↪ Probabilidades binomiales acumuladas...
↪ Valor(es) de la variable `7`
↪ Ensayos binomiales `10`
↪ Probabilidad de éxito `0.4`
↪ Aceptar



El resultado es $P(X \leq 7) = 0'9877054$.

b) En este caso no se trata de una distribución continua, con lo que la probabilidad en un punto no es siempre cero. Una posible forma de obtener la probabilidad $P(X = 7)$ seria seguir los siguientes pasos:

↪ Distribuciones
↪ Distribuciones discretas
↪ Distribución binomial
↪ Probabilidades binomiales...
↪ Ensayos binomiale: `10`
↪ Probabilidad de éxito `.4`
↪ Aceptar

y observar como la probabilidad en el 7 es $0'0424673280$.

c) Como acabamos de ver, no se puede afirmar en general que $P(X < 7) = P(X \leq 7)$. Para calcular $P(X < 7)$ debemos observar que el el caso de la binomial lo que se tiene es que $P(X < 7) = P(X \leq 6)$, con lo que reemplazando en el primer apartado lo que sigue

↪ Valor(es) de la variable `6`

obtenemos que $P(X < 7) = 0'9452381$.

d) Para calcular $P(X > 7)$ simplemente tenemos que elegir la opción `Cola derecha` y obtenemos que $P(X > 7) = 0'01229455$.

e) De nuevo no se tiene la igualdad del caso continuo, es decir, $P(X \geq 7) \neq P(X > 7)$. Una forma de calcular $P(X \geq 7)$ sería observando que en la binomial se tiene que $P(X \geq 7) = P(X > 6) = 0'05476188$.

f) La probabilidad $P(4 < X < 7)$ la podríamos calcular como $P(4 < X < 7) = P(X \leq 6) - P(X \leq 4) = 0'3121348$.

g) En este caso, $P(4 \leq X \leq 7) = P(X \leq 7) - P(X \leq 3) = 0'6054248$.

h) En el caso de la binomial $B(10, 0'4)$ la variable solo puede tomar los valores $0, 1, 2, \ldots, 9, 10$, con lo que $P(X = 2'3) = 0$.

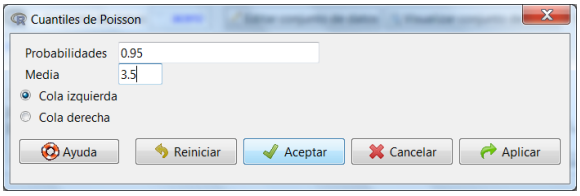
i) Por el mismo razonamiento del apartado anterior se llegaría a que $P(X = 25) = 0$.

□

Ejemplo 2.8. *Halle el percentil* 95 *de una distribución de Poisson de parámetro* $\lambda = 3'5$.

Solución: Siga la ruta:

↪ Distribuciones
↪ Distribuciones discretas
↪ Distribución de Poisson
↪ Cuantiles de Poisson...
↪ Probabilidades: `0.95`
↪ Media: `3.5`
↪ Cola izquierda (elegida)
↪ Aceptar



El resultado debe ser 7.

□

2.3. Ejercicios propuestos

Ejercicio 2.1. *Los errores de medida de una máquina siguen una distribución normal $N(0, 2)$. Se pide:*

a) Calcular la probabilidad de que el error sea menor de 1.

b) Calcular la probabilidad de que el error esté entre -2 y 2.

c) Calcular el valor del error tal que el 30 % de las veces el error es menor o igual que él y el resto de las veces es mayor o igual.

d) Calcular el valor del error tal que el 20 % de las veces el error es mayor o igual que él y el resto de las veces es menor o igual.

Ejercicio 2.2. *Si la duración en años de una pieza es una exponencial de media 2 años, se pide:*

a) Calcular la probabilidad de que dure al menos 5 años.

b) Calcular la probabilidad de que dure como mucho 6 años.

c) Calcular la probabilidad de que dure entre 5 y 6 años.

d) Obtener el tiempo de garantía que se tiene que dar a dicha pieza para que como mucho el 40 % de las piezas sean reparadas durante el periodo de garantía.

Ejercicio 2.3. *Si la duración en años de una pieza es una Weibull con parámetro de forma $k = 2$ y parámetro de escala $\lambda = 3$, calcula la probabilidad de que dure más de 5 años.*

Ejercicio 2.4. *Se ha hecho un estudio con cojinetes de rodillos y se ha obtenido que su tiempo de vida (en cientos de horas) sigue una distribución Weibull de parámetros de forma $k = 0'4$ y escala $\lambda = 4$.*

a) Estimar la probabilidad de que los cojinetes fallen antes de 160 horas.

b) Dado un lote de 10 cojinetes elegidos de forma independiente, ¿cuál es la probabilidad de que ninguno falle antes de 160 horas? ¿y de que falle como mucho uno?

Ejercicio 2.5. *Si el número de averías en una empresa en un turno de 8 horas sigue una distribución de Poisson de parámetro $\lambda = 16$.*

a) ¿Cuál es la probabilidad de que en un turno haya más de 20 averías?

b) ¿Cuál es la probabilidad de que el tiempo entre dos averías sea mayor de 1 hora?

2.4. Soluciones de los ejercicios propuestos

Ejercicio 2.1. *Si denotamos por X la variable aleatoria "error de medida de la máquina" se tiene que $X \equiv N(0,2)$. Con esto la solución a los distintos apartados es:*

a) $P(X < 1) = 0'6914625$.

b) $P(-2 < X < 2) = 0'6826894$.

c) $-1'048801$.

d) $1'683242$.

Ejercicio 2.2. *Si denotamos por X la variable aleatoria "duración en años de la pieza" se tiene que $X \equiv exp(0'5)$, puesto que se tiene como dato que $E(X) = 2$. Con esto se tiene que la solución a los distintos apartados es:*

a) $P(X \geq 5) = 0'082085$.

b) $P(X \leq 6) = 0'9502129$.

c) $P(5 \leq X \leq 6) = P(X \leq 6) - P(X < 5) = P(X \leq 6) - (1 - P(X \geq 5)) = 0'9502129 - (1 - 0'082085) = 0'0322979$.

d) *El punto c tal que $P(X \leq c) = 0'4$ es $c = 1'021651$.*

Ejercicio 2.3. *Si denotamos por X la variable aleatoria "duración en años de la pieza" se tiene que $X \equiv W(2,3)$ y la probabilidad pedida es $P(X > 5) = 0'06217652$.*

Ejercicio 2.4. *Si denotamos por X la variable aleatoria "duración en cientos de horas de un cojinete" se tiene que $X \equiv W(0'4, 4)$.*

a) *La probabilidad pedida es: $P(X < 1'6) = P(X \leq 1'6) = 0'4999988$.*

b) *Si $Y =$"n^o de cojinetes que han fallado antes de las 160 horas de los 10 elegidos de forma independiente", entonces $Y \equiv B(10, 0'4999988)$. Las probabilidades pedidas son:*

- $P(Y = 0) = 0'0009765859$.
- $P(Y \leq 1) = 0'0107424$.

Ejercicio 2.5. *a) Si el número de averías en una empresa en un turno de 8 horas sigue una distribución de Poisson de parámetro $\lambda = 16$ y denotamos dicha variable por X, lo que nos piden es $P(X > 20)$, que se podría calcular con la opción* `Cola derecha` *de una Poisson de parámetro* `16` *en el valor* `20`*, con lo que se obtiene que $P(X > 20) = 0'131832$.*

b) Si el número de averías en 8 h es una $P(16)$, el tiempo, en horas, entre dos averías sigue una distribución exponencial $exp(2)$, con lo que sin más que considerar la cola derecha de una exponencial de parámetro `2` *en el valor* `1`*, obtendríamos que la probabilidad de que el tiempo entre dos averías sea mayor de 1 hora es $0'1353353$.*

Práctica 3

Contrastes para una muestra

3.1. Introducción al contraste de hipótesis

Los métodos descriptivos proporcionan una idea de cómo es la muestra. Para obtener conclusiones relativas a la población necesitamos utilizar técnicas de inferencia estadística (contraste de hipótesis). Una *hipótesis* es una afirmación sobre las características estadísticas de un proceso. Una hipótesis es una conjetura. Por ejemplo: si un técnico observa el consumo de energía durante varias horas, sabrá el consumo medio de las horas que observó. Puede avanzar un paso más y conjeturar que el consumo medio de todas las horas de trabajo en esa fábrica es de 120. El proceso científico consiste entonces en probar su hipótesis contra una hipótesis alternativa.

$$\begin{array}{rrlcl} \text{Hipótesis nula} & H_0\text{:} & \text{consumo medio} & = & 120 \\ \text{Hipótesis alternativa} & H_1\text{:} & \text{consumo medio} & \neq & 120 \end{array}$$

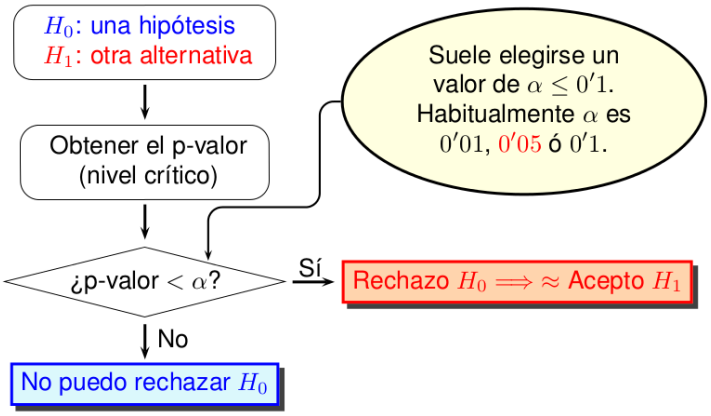
Un *contraste* consiste en un procedimiento estadístico para determinar la validez de una hipótesis (la hipótesis nula). Si los datos de la muestra resultan poco creíbles de obtenerse en caso de ser cierta dicha hipótesis, nuestra razón nos obligará a *rechazarla*. En caso contrario, no hay base suficiente para rechazarla, y entonces se *mantiene* o *considera admisible* la hipótesis nula.

Para presentar los resultados de un contraste de hipótesis utilizamos el *P-valor* o *nivel crítico*. El P-valor es el nivel de significación menor que llevaría al rechazo de la hipótesis nula H_0. Una vez que se conoce el P-valor, el responsable de tomar las decisiones puede determinar si rechazar o no la hipótesis nula comparándolo con un nivel de significación α (que puede venir impuesto o que se decide en el momento).

REGLA DE DECISIÓN

P-valor $< \alpha$	$\Longrightarrow$	**Rechazo** H_0	
P-valor $\geq \alpha$	$\Longrightarrow$	**No rechazo** H_0	

Generalmente se considera $\alpha = 0'05$***.***



Algunos ejemplos de los principales contrastes de hipótesis con los que nos encontramos son:

- **Promedio de una población:** ¿El consumo medio es de 120?
- **Proporción poblacional:** ¿El porcentaje de horas con consumo alto es menor del 1 %?
- **Comparación de promedios:** ¿El consumo medio es distinto en la línea A y en la B?
- **Comparación de proporciones:** ¿El porcentaje de horas con consumo alto es distinto en la línea A y en la B?
- **Comparación de varianzas:** ¿La variabilidad del consumo es distinta en la línea A y en la B?

3.2. Contrastes para el promedio

En general, para realizar un contraste consideramos las siguientes etapas:

1. seleccionar el contraste adecuado a la muestra,
2. establecer quiénes son H_0 y H_1 en ese contraste e
3. interpretar el P-valor.

En particular para un contraste sobre un promedio, para seleccionar el contraste adecuado a nuestra muestra hemos de tener en cuenta si los datos siguen aproximadamente una distribución normal o no, y, según sea el resultado, decidir qué contraste realizamos (Tabla 3.1)[1].

Para comprobar dicha hipótesis, si se puede considerar que los datos provienen de una población normal o no, haremos un contraste previo de normalidad, antes de comenzar a contrastar las hipótesis sobre el promedio. Para contrastar la normalidad de los datos utilizaremos el **contraste de Shapiro-Wilk**, salvo que los datos vengan dados en intervalos, en cuyo caso realizamos el contraste de χ^2 (véase práctica 5). Para este tipo de contrastes, las hipótesis a contrastar son:

[1]El contraste de Wilcoxon para una muestra sólo es recomendable cuando la distribución es simétrica. Si los datos muestrales no avalan esta premisa, se pueden utilizar otros tests que están fuera del alcance de este curso.

Tabla 3.1: Contrastes para un promedio.

Contraste para la	¿Distribución aprox. normal?	Tipo de contraste
Media (μ)	Sí	Contraste t para una muestra
Mediana (Me)	No	Contraste de Wilcoxon para una muestra

CONTRASTE DE BONDAD DE AJUSTE A LA NORMAL

H_0: los datos provienen de una población normal
H_1: los datos NO provienen de una población normal

REGLA DE DECISIÓN

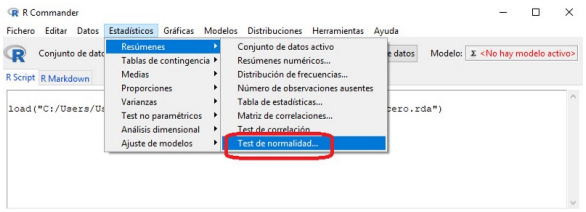
P-valor $< \alpha \Longrightarrow$ Rechazo H_0 (la distribución no es normal)
P-valor $\geq \alpha \Longrightarrow$ No rechazo H_0 (se puede admitir la normalidad)

Generalmente se considera $\alpha = 0'05$.

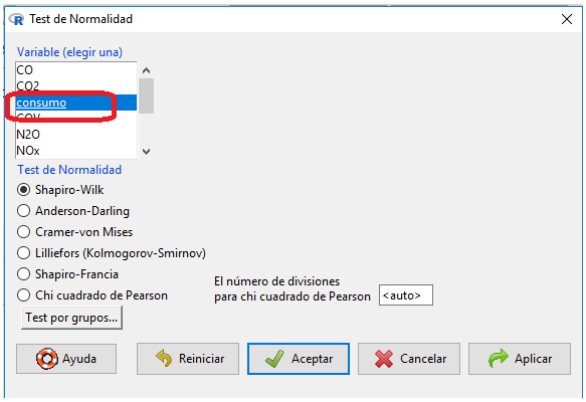
Ejemplo 3.1. *Estudie la normalidad de la variable* `consumo`.

Solución: Utilizamos el contraste de Shapiro y Wilk:

Estadísticos
➥Resúmenes
➥Test de normalidad de Shapiro...



Seleccionar `consumo`
➥Aceptar



y obtenemos

```
  Shapiro-Wilk normality test

data:  consumo
W = 0.9884, p-value = 0.4207
```

Como el P-valor (0′4207) es mayor que α ($\alpha = 0'05$ por omisión) no se rechaza la hipótesis nula, y podemos decir que los datos siguen una distribución normal. □

Como el consumo sigue una distribución normal, estamos en condición de realizar un contraste para la media. El contraste adecuado en este caso es el **contraste t para una muestra**, cuyas hipótesis a contrastar pueden ser de tres tipos, tal como se describe en el siguiente ejemplo:

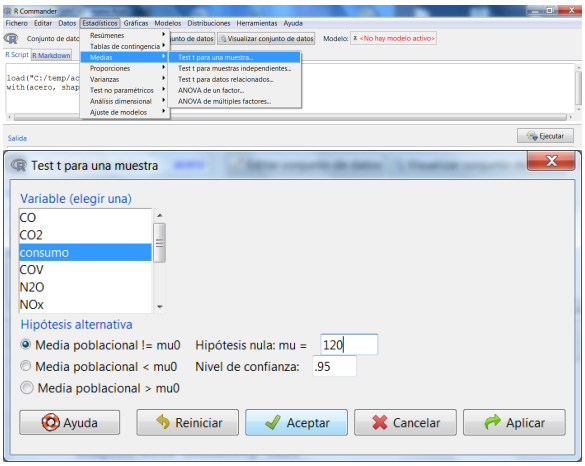
$H_0 : \mu = 120$	$H_0 : \mu \geq 120$	$H_0 : \mu \leq 120$
$H_1 : \mu \neq 120$	$H_1 : \mu < 120$	$H_1 : \mu > 120$

Ejemplo 3.2. *¿Es el consumo medio distinto de 120?*

Solución: En este caso se tiene:

H_0:	el consumo medio es de 120
H_1:	el consumo medio no es de 120

Estadísticos
➥Medias
➥Test t para una muestra



Seleccionar la variable `consumo`
➥Poner 120 en la hipótesis nula
➥Aceptar

```
        One Sample t-test

data:  consumo
t = 2.9798, df = 116, p-value = 0.003514
alternative hypothesis: true mean is not equal to 120
95 percent confidence interval:
 125.2568 146.0974
sample estimates:
mean of x
 135.6771
```

Si consultamos la regla de decisión:

P-valor $< \alpha$	$\Longrightarrow$	Rechazo H_0 (consumo medio $\neq$ 120)
P-valor $\geq \alpha$	$\Longrightarrow$	No rechazo H_0 (consumo medio $=$ 120)

Generalmente se considera $\alpha = 0'05$.

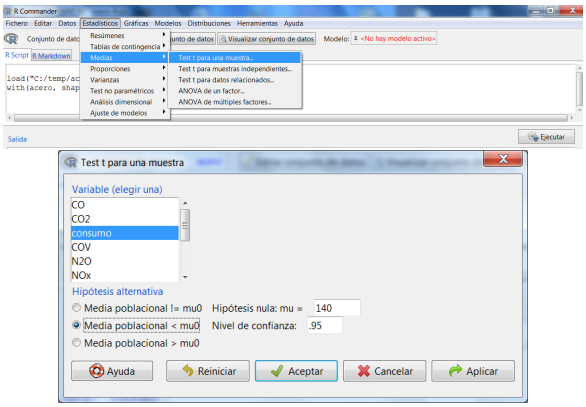
En este caso el P-valor (0′003514) es menor que α y se rechaza la hipótesis nula (H_0); por tanto, la conclusión es que la media es distinta de 120. □

Ejemplo 3.3. *¿El consumo medio es menor de 140?*

Solución: En este caso, como vuelve a tratarse de la variable `consumo`, ya sabemos que se comporta de manera normal y puede utilizarse el test t para la media. Así, un test adecuado para responder a esta pregunta sería:

H_0:	el consumo medio es mayor o igual de 140
H_1:	el consumo medio es menor de 140

Estadísticos
➥Medias
➥Test t para una muestra



Seleccionar la variable `consumo`
➥Poner 140 en la hipótesis nula
➥Marcar Media poblacional < mu0
➥Aceptar

```
    One Sample t-test

data:  consumo
t = -0.8217, df = 116, p-value = 0.2065
alternative hypothesis: true mean is less than 140
95 percent confidence interval:
     -Inf 144.4005
sample estimates:
mean of x
 135.6771
```

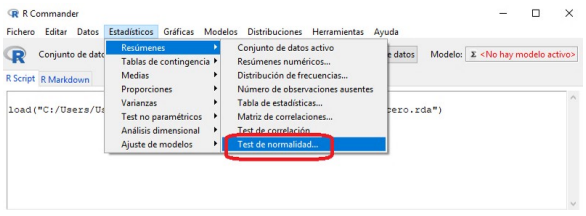
Como el p-valor (0′2065) supera a α, no se rechaza la hipótesis nula. No hay evidencia suficiente en la muestra para suponer que la media poblacional es menor de 140.

□

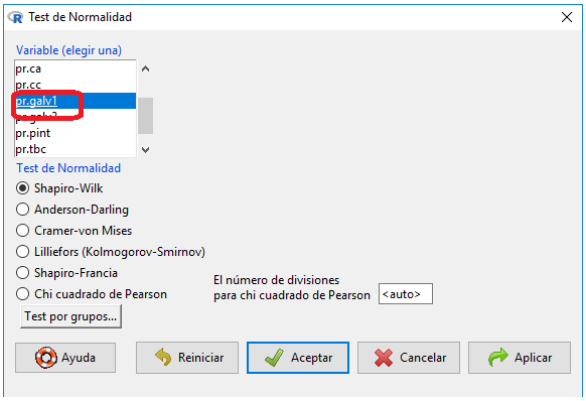
Ejemplo 3.4. *Se quiere hacer un contraste sobre el promedio de producción de galvanizado de tipo I. Para elegir un procedimiento adecuado es necesario contestar previamente a la siguiente pregunta: ¿la variable* `pr.galv1` *sigue una distribución normal?*

Solución: Examinaremos la normalidad de la variable `pr.galv1` mediante un contraste de hipótesis, tal como hemos hecho previamente.

Estadísticos
➥Resúmenes
➥Test de normalidad de Shapiro...



Seleccionar `pr.galv1`
➥Aceptar



Con lo que se obtiene:

```
    Shapiro-Wilk normality test

data:  pr.galv1
W = 0.9697, p-value = 0.00957
```

Como el P-valor (0′00957) es menor que α, se rechaza la hipótesis nula; por lo tanto, no hay normalidad.

□

Ejemplo 3.5. *¿La producción promedio de galvanizado 1 es menor de 400?*

Solución: Al no haber normalidad tenemos que realizar el **contraste de Wilcoxon para una muestra**. Para éste los distintos tipos de contrastes de hipótesis para la mediana son:

$H_0 : Me = 400$ $H_1 : Me \neq 400$	$H_0 : Me \geq 400$ $H_1 : Me < 400$	$H_0 : Me \leq 400$ $H_1 : Me > 400$
`two.side`	`less`	`greater`

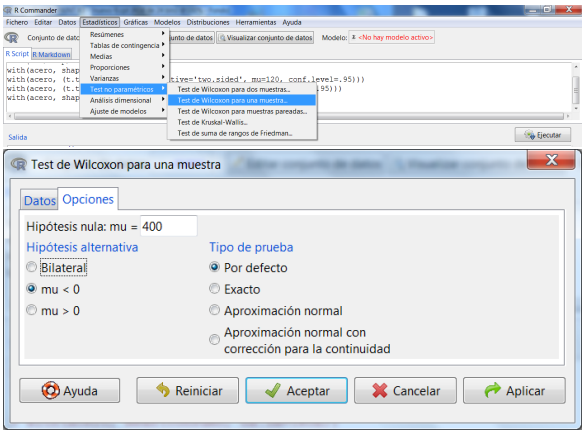
La hipótesis que nos interesa es:

¿La producción promedio es menor de 400?

$$\begin{array}{|c|}\hline H_0 : Me \geq 400 \\ H_1 : Me < 400 \\ \hline \end{array}$$

Para hacer esto seguimos los siguientes pasos:

Estadísticos
➥Test no paramétricos
➥Test de Wilcoxon para una muestra



Seleccionar la variable `pr.galv1`
➥Poner 400 en `Hipótesis nula: mu =`
➥Marcar `mu < 0` en `Hip.alternativa`
➥Aceptar

De los cuales se obtiene como resultado:

```
    Wilcoxon signed rank test with continuity correction

data:  pr.galv1
V = 4003.5, p-value = 0.9538
alternative hypothesis: true location is less than 400
```

Como el p-valor (0′9538) es mayor que α, no hay evidencia suficiente para rechazar la hipótesis nula; por tanto, hay que suponer que la producción promedio es al menos 400. □

3.3. Proporción poblacional

Es frecuente el interés por saber qué proporción de individuos de una población presentan una característica determinada, frente a los que no la presentan. Por ejemplo, queremos saber si el porcentaje de horas con avería es excesivo, considerándose excesivo si el porcentaje es mayor del 10 %.

Ejemplo 3.6. *Para nuestro ejemplo, ¿el porcentaje de horas con averías es significativamente mayor del 10 %?*

Solución: Procedemos a dar los pasos habituales en un contraste de hipótesis:

Seleccionar el contraste adecuado a la muestra

Para este problema el contraste de hipótesis es el **contraste de proporciones para una muestra**. R-Commander permite aplicar este contraste a las variables dicótomas (factores con exactamente dos niveles) del conjunto de datos.

Establecer quiénes son H_0 y H_1 en ese contraste

Los distintos tipos de contrastes de hipótesis para la proporción son del tipo:

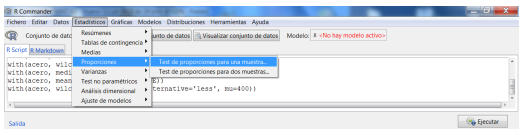
$H_0 : p = 0'1$ $H_1 : p \neq 0'1$	$H_0 : p \geq 0'1$ $H_1 : p < 0'1$	$H_0 : p \leq 0'1$ $H_1 : p > 0'1$
`two.side`	`less`	`greater`

Tendríamos en cuenta que R-Commander considera, por defecto, que la proporción p se refiere a la primera clase por orden alfabético, salvo que las categorías de la variable hayan sido previamente ordenadas por el investigador[2]; en este caso no es así y como la A va delante de la NoA, en nuestro ejemplo p se refiere a la proporción de la clase A, es decir, a la proporción de horas con averías. De este modo, consideramos las siguientes hipótesis a contrastar:

$$\begin{array}{|l|}\hline H_0 : p \leq 0'1 \text{ (proporción razonable de averías)} \\ H_1 : p > 0'1 \text{ (proporción excesiva de averías)} \\ \hline \end{array}$$

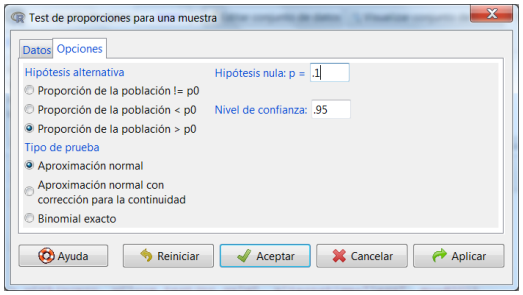
Ahora solo habría que hacer

Estadísticos
➥Proporciones
➥Test de proporciones para una muestra



[2]Esto se puede hacer con la opción del menú Datos ⟶ Modificar variables del conjunto de datos activo ⟶ Reordenar niveles del factor..., tal como veremos más adelante.

Seleccionar la variable `pres.aver`
➥Escribir `0.1` en `Hipótesis nula: p=`
➥Seleccionar `Proporción de la población > p0` en `Hipótesis alternativa`
➥Aceptar



Cuyo resultado es:

```
        1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.1
X-squared = 25.2317, df = 1, p-value = 2.542e-07
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
 0.1807938 1.0000000
sample estimates:
        p
0.2393162
```

Interpretar el p-valor
Como el P-valor $(2'542 \cdot 10^{-7})$ es menor que α se rechaza la hipótesis nula; se concluye que la proporción de averías es excesiva.

□

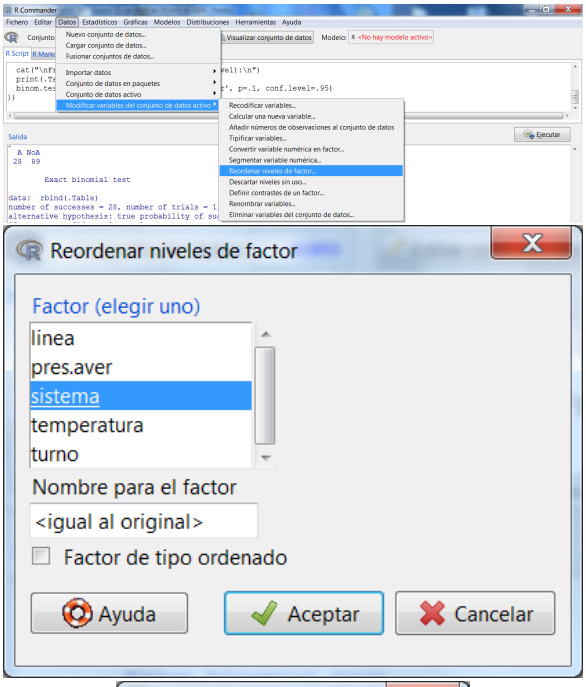
VARIACIONES

Binomial exacto. El contraste se ha realizado dejando la opción por omisión *Aproximación normal* en *Tipo de contraste*. Si los tamaños muestrales son pequeños, conviene hacer el contraste con la opción *Binomial exacto*. En este caso las diferencias son irrisorias, ya que el P-valor también es muy pequeño:

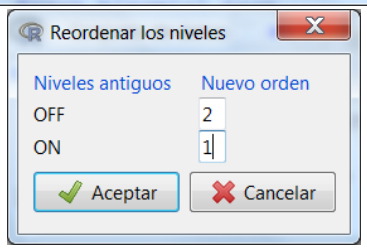
```
    Exact binomial test

data:  rbind(.Table)
number of successes = 28, number of trials = 117, p-value = 1.002e-05
alternative hypothesis: true probability of success is greater than 0.1
95 percent confidence interval:
 0.1757437 1.0000000
sample estimates:
probability of success
             0.2393162
```

Reordenar niveles del factor. Si el orden de las categorías no es el que queremos, podemos reescribir las hipótesis en función del valor que representa p para `R` o reordenar los niveles de factor. Si queremos, por ejemplo, como primer factor **ON** en la variable `sistema` deberíamos proceder como sigue:

Datos
➥Modificar variables
➥Reordenar niveles de factor



Seleccionar la variable `sistema`
➥Aceptar



Reordenar de la forma deseada
➥Aceptar

De esta manera, si nos piden contrastar si la proporción de veces que el sistema está encendido es menor del 40 %, las hipótesis del test serían:

$$H_0 : p \geq 0'4 \text{ (sistema encendido al menos el 40 \% de las veces)}$$
$$H_1 : p < 0'4 \text{ (sistema encendido menos del 40 \% de las veces)}$$

3.4. Intervalos de confianza

Finalmente, nótese que en la salida de estos tests también se muestra el intervalo de confianza para el parámetro (la media poblacional en el caso del test t para una muestra y la proporción poblacional en el caso del test de proporciones para una muestra). Por defecto, se obtiene un intervalo de confianza al 95 %. Esta opción se puede cambiar entre las opciones del propio test. Nótese también que, para obtener un intervalo acotado como el visto en clase, debemos seleccionar la opción bilateral representada con el símbolo "! =".

3.5. Ejercicios propuestos

Ejercicio 3.1. *a) Obtener un intervalo de confianza del consumo medio a los niveles de confianza $\alpha = 90\,\%$, $\alpha = 95\,\%$ y $\alpha = 99\,\%$, respectivamente.*

b) Obtener un intervalo de confianza de la proporción de veces que se usa la línea A a los niveles de confianza $\alpha = 90\,\%$, $\alpha = 95\,\%$ y $\alpha = 99\,\%$, respectivamente.

Ejercicio 3.2. *Responda razonadamente a las siguientes cuestiones:*

a) ¿Cuál ha sido el consumo medio de estos 117 datos? ¿y su desviación típica?

b) Realiza el histograma del consumo. Dicho gráfico ¿hace sospechar que los datos proceden de una distribución normal?

c) Realiza un contraste de hipótesis para contrastar la normalidad del consumo. ¿Cuál es su p-valor? En función de dicho p-valor, ¿puede admitirse la normalidad de los datos?

d) Admitiendo como cierto el resultado obtenido en el apartado anterior y considerando que los valores de media y desviación típica de la variable aleatoria consumo coinciden con las estimaciones puntuales obtenidas en el apartado a),

- *¿qué porcentaje esperado de horas laborales de esta empresa se estima que habrá consumos mayores de 265 megavarios-hora? ¿y menores de 99 megavatios-hora? ¿y entre 99 y 265 megavatios-hora?*
- *Si la empresa quiere instalar un sistema energético, ¿cuál se estima que debería ser la capacidad energética mínima de dicho sistema, para que sólo se quede sin abastecimiento como mucho un 2 % de las veces?*

Ejercicio 3.3. *¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora?*

Ejercicio 3.4. *a) ¿Apoyan estos datos la hipótesis de que el consumo promedio es menor de 130 megavatios-hora en aquellas horas en las que la temperatura es alta?*

b) Realice un diagrama de cajas de la variable consumo para cada una de las temperaturas consideradas y comente los resultados.

Ejercicio 3.5. *¿Es admisible considerar que la producción promedio del convertidor de acero es menor de 250 toneladas? ¿y distinta de 250 toneladas? ¿y distinta de 240? ¿y distinta de 180? Obtenga el correspondiente valor promedio de la producción del convertidor de acero en la muestra y comenta los resultados en relación con las preguntas contestadas anteriormente.*

Ejercicio 3.6. *¿Apoyan estos datos la hipótesis de que el porcentaje de veces que se usa la línea A es mayor del 20 %?*

Ejercicio 3.7. *Responda razonadamente a las siguientes cuestiones:*

a) Realiza un gráfico para representar adecuadamente el número de horas en la muestra en las que el sistema de sobrecalentamiento estaba encendido y el número de horas en las que estaba apagado. ¿Cuál ha sido el porcentaje de horas en las que ha estado encendido?

b) La compra del sistema de sobrecalentamiento no ha sido rentable si, en general, éste es usado menos del 40 % de las veces. Considerando estos datos como una muestra aleatoria del comportamiento de esta empresa respecto a las variables en estudio, ¿permiten concluir que la compra del sistema no ha sido rentable?

c) Un estudio sobre este sistema consiste en elegir 25 horas al azar de la producción de un mes y analizar si estaba encendido o no el sistema en cada una de ellas. Considerando que la proporción poblacional coincide con la estimación puntual obtenida en el apartado a), ¿en cuánto se estima la probabilidad de que exactamente 9 horas de las 25 el sistema estuviese encendido? ¿y como mucho 12 horas? ¿y al menos 10 horas? ¿y entre 10 y 12 horas, ambas inclusive? ¿y entre 9'5 y 12'5 horas? ¿y más de 9 y menos de 13 horas?

3.6. Soluciones de los ejercicios propuestos

Ejercicio 3.1. *a) Intervalos de confianza del consumo medio:*

Nivel de confianza	*Intervalo*
$\alpha = 90\,\%$	$(126'9537, 144'4005)$
$\alpha = 95\,\%$	$(125'2568, 146'0974)$
$\alpha = 99\,\%$	$(121'8989, 149'4553)$

b) Intervalos de confianza de la proporción de veces que se usa la línea A:

Nivel de confianza	*Intervalo*
$\alpha = 90\,\%$	$(0'4457825, 0'5959867)$
$\alpha = 95\,\%$	$(0'4316194, 0'6097571)$
$\alpha = 99\,\%$	$(0'4044922, 0'6359495)$

Ejercicio 3.2. *Las respuestas razonadas a las cuestiones planteadas pueden verse a continuación:*

a) Tal como se obtiene con la opción del menú Estadísticos → Resúmenes → Resúmenes numéricos:

```
     mean       sd   IQR   0%   25%   50%    75%   100%   n
 135.6771 56.90756 83.39 17.5 99.09 135.1 182.48 290.72 117
```

la media es $135'6771$ *megavatios-hora y la desviación típica* $56'90756$ *megavatios-hora.*

b) Sí, a priori parece que se podrían ajustar los datos a una distribución normal, puesto que el histograma se asemeja a la forma de su función densidad (campana).

c) El p-valor del test de normalidad de Shapiro-Wilk aplicado a estos datos es $0'4207$*. Como es claramente mayor que el nivel de significación (habitualmente* $\alpha = 0'05$ *y en general como máximo* $\alpha = 0'1$*), no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer la normalidad de la variable "consumo".*

d) Puesto que, según el apartado anterior, podemos suponer que la variable sigue una distribución normal y como parámetros consideraremos las estimaciones puntuales de la media y desviación típica poblacional obtenidas en el apartado a), estamos admitiendo que $X =$*"consumo"*$\equiv N(135'677, 56'908)$ *(redondeando los valores de los parámetros a tres decimales). En función de esto y utilizando las opciones de* Distribuciones → Distribuciones continuas → Distribución normal *podemos contestar a las cuestiones planteadas como sigue:*

- *Puesto que* $P(X > 265) = 0'01152839$*, el porcentaje esperado de las horas laborales de esta empresa en las que habrá consumos mayores de 265 megavatios-hora se estima que es del* $1'15\,\%$.

  De forma análoga, puesto que $P(X < 99) = 0'2596268$*, se estima que el porcentaje de horas laborales con menos de 99 megavatios-hora de consumo estará alrededor del* $25'96\,\%$.[3]

  De lo anterior se deduce que el porcentaje esperado de veces que el consumo está entre 99 y 265 megavatios-hora se estima que es $(100 - 1'15 - 25'96) = 72'89\,\%$.

- *Si denotamos por c la capacidad energética mínima del sistema, buscamos el valor c tal que* $P(X > c) = 0'02$*. Con la función* Cuantiles normales *se obtiene que* $c = 252'5517$*, con lo que se verificarían las condiciones exigidas si el sistema energético tuviese una capacidad de* $252'5517$ *megavatios-hora, puesto que se espera que sólo un* $2\,\%$ *de las veces la empresa se quedaría sin abastecimiento energético.*

Ejercicio 3.3. *Para contestar a esta pregunta podemos plantearnos el contraste* $H_0 : \mu \geq 130$ *frente a la alternativa* $H_1 : \mu < 130$*, puesto que como hemos visto en el ejercicio anterior se puede asumir la normalidad de la variable "consumo" y, por tanto, se puede usar el test t para una muestra para realizar el contraste anterior. Los resultados de dicho procedimiento con R son:*

```
    One Sample t-test

data:  consumo
t = 1.0791, df = 116, p-value = 0.8586
alternative hypothesis: true mean is less than 130
95 percent confidence interval:
     -Inf 144.4005
sample estimates:
mean of x
 135.6771
```

y puesto que el p-valor es $0'8586$*, no se rechaza la hipótesis nula, es decir, no hay evidencias en contra de suponer que el consumo medio sea mayor o igual de 130 megavatios-hora.*

Ejercicio 3.4. *a) En este caso es necesario comenzar filtrando los datos y creando un nuevo conjunto de datos que llamaremos, por ejemplo,* `acero_tempAlta`*. Esto ya ha sido hecho en la sección 1.7 de la primera práctica, con lo que se puede consultar allí los pasos a seguir. Una vez hecho esto, deberíamos contrastar la normalidad de la variable consumo con temperaturas altas. Los resultado del test de normalidad de Shapiro-Wilk son:*

```
    Shapiro-Wilk normality test

data:  consumo
W = 0.9448, p-value = 0.02965
```

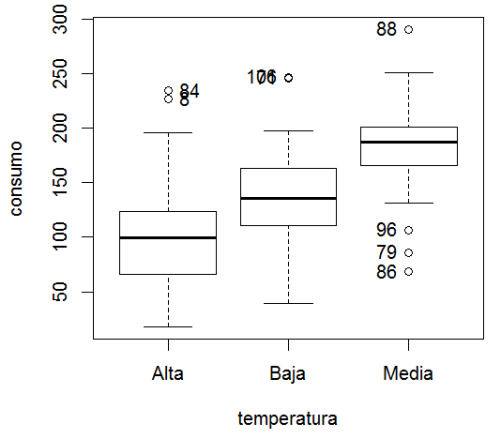
[3] Obsérvese la diferencia entre la probabilidad (porcentaje esperado) y el porcentaje en la muestra. Si nos fijamos en el apartado a) vemos como $99'09$ es el primer cuartil muestral, mientras que su probabilidad es sensiblemente superior al $0'25$.

con lo que no se puede suponer la normalidad y, por tanto, utilizaremos el test de Wilcoxon para una muestra. Los resultados de dicho test son:

```
    Wilcoxon signed rank test

data:  consumo
V = 245, p-value = 0.000464
alternative hypothesis: true location is less than 130
```

por lo tanto se rechaza la hipótesis nula y se concluye que el consumo promedio a temperaturas altas es menor de 130 megavatios-hora.

b) *Lo primero de todo es seleccionar la base de datos* `acero`*, ya que ahora está activa la base de datos* `acero_tempAlta`*. Hecho esto y el correspondiente diagrama de cajas para la variable* `consumo` *agrupada por* `temperatura` *obtenemos el gráfico:*



donde se observa claramente como los consumos a temperaturas altas han sido los más bajos. De hecho, si observamos la media global del consumo es de 135′6771 *megavatios-hora (ver, por ejemplo, el ejercicio anterior) y el consumo medio a temperatura alta ha sido de* 103′5239 *megavatios-hora. Este hecho explicaría que en general no se pueda concluir que el consumo es menor de 130, pero si nos restringimos a las horas con temperatura alta sí.*

Ejercicio 3.5. *Comenzamos contrastando la normalidad de los datos correspondientes a la variable* `pr.ca`*. El resultado del test de normalidad de Shapiro-Wilk es:*

```
    Shapiro-Wilk normality test

data:  pr.ca
W = 0.9349, p-value = 2.49e-05
```

por lo que se rechaza la normalidad de los datos. Para hacer test sobre la producción promedio usaremos pues el test de Wilcoxon para una muestra. Para ello, vamos a la opción del menú Estadísticos → Test no paramétricos → Test de Wilcoxon para una muestra *y la rellenamos como corresponde, con lo que se obtiene el resultado*

```
    Wilcoxon signed rank test with continuity correction

data:  pr.ca
V = 3071.5, p-value = 0.151
alternative hypothesis: true location is less than 250
```

con lo que no rechazamos H_0 y, por tanto, no hay evidencias que nos permitan concluir que la producción promedio del convertidor de acero es menor de 250 toneladas.

Para contrastar si la mediana es distinta de 250, iríamos a la opción del menú Estadísticos → Test no paramétricos → Test de Wilcoxon para una muestra *y la rellenaríamos como corresponde (*`Hipótesis alternativa: Bilateral, Hipótesis nula: mu = 250`*), obteniéndose un p-valor de $0'302$, lo que también nos lleva a considerar admisible que la mediana es de 250 toneladas.*

Si contrastamos $H_0 : Me = 240$ frente a la alternativa $H_1 : Me \neq 240$, el p-valor obtenido es $0'6244$, de nuevo admisible H_0.

Sin embargo, si contrastamos $H_0 : Me = 180$ frente a la alternativa $H_1 : Me \neq 180$, el p-valor obtenido es $0'001068$, por lo que se llega a la conclusión de que la mediana es distinta de 180.

Si obtenemos la mediana muestral de estos 117 datos, vemos que es 225. Esto nos ha llevado a no rechazar medianas poblacionales de 250, 240 o mayores o iguales a 250, pero sí a rechazar una mediana poblacional de 180.

```
             mean        sd  IQR 0%  25%  50%  75%  100%   n
pr.ca    244.9231  167.5311  234 13   99  225  333   677 117
```

Ejercicio 3.6. *Sí, porque el p-valor asociado al contraste $H_0 : p \leq 0'2$ frente a la alternativa $H_1 : p > 0'2$, donde p representa la proporción de veces que se usa la línea A, es menor de $2'2 \cdot 10^{-16}$, con lo que se rechaza H_0 y se concluye que existen evidencias significativas de que el porcentaje de veces que se usa la línea A es mayor del 20 %.*

Ejercicio 3.7. *a) Un gráfico adecuado para representar el número de horas en la muestra en las que que el sistema de sobrecalentamiento estuvo encendido sería hacer un diagrama de barras de la variable* `sistema`*. En la tabla de frecuencias vemos que un $49'57\,\%$ de las horas ha estado encendido.*

b) Contrastar $H_0 : p \leq 0'6$ frente a la alternativa $H_1 : p > 0'6$, puesto que $OFF < ON$ y, por tanto, $p = P(OFF)$.

El p-valor es $0'9827$, por lo que no hay evidencias de que $P(OFF) > 0'6$ o, lo que es lo mismo, que $P(ON) < 0'4$. No hay nada que indique que el sistema no es rentable.

c) Si consideramos la variable $X =$"nº de horas que el sistema está encendido de las 25 elegidas al azar" y suponemos que la proporción poblacional coincide con la estimación puntual obtenida en el apartado a), se tiene que $X \equiv B(25, 0'4957)$. Así,

- $P(X = 9) = 6'460361 \cdot 10^{-2} = 0'06460361$, *con lo que se estima que la probabilidad de que exactamente 9 horas de las 25 el sistema estuviese encendido es* $0'06460361$.
- $P(X \leq 12) = 0'5173218$, *con lo que se estima que la probabilidad de que como mucho 12 horas de las 25 el sistema estuviese encendido es* $0'5173218$.
- $P(X \geq 10) = P(X > 9) = 0'8766434$, *con lo que se estima que la probabilidad de que al menos 10 horas de las 25 el sistema estuviese encendido es* $0'8766434$.
- $P(10 \leq X \leq 12) = P(X \leq 12) - P(X < 10) = P(X \leq 12) - (1 - P(X \geq 10) = 0'5173218 - (1 - 0'8766434) = 0'3939652$, *con lo que se estima que la probabilidad de que entre 10 y 12 horas, ambas inclusive, de las 25 el sistema estuviese encendido es* $0'3939652$.
- $P(9'5 \leq X \leq 12'5) = P(10 \leq X \leq 12) = 0'3939652$, *con lo que se estima que la probabilidad de que entre 9'5 y 12'5 horas de las 25 el sistema estuviese encendido es* $0'393913$.
- $P(9 < X < 13) = P(10 \leq X \leq 12) = 0'3939652$, *con lo que se estima que la probabilidad de que, de las 25, el sistema estuviese encendido más de 9 y menos de 13 es* $0'393913$.

Práctica 4

Contrastes para dos muestras

En la práctica anterior estudiamos los contrastes de hipótesis estadísticas en los casos en que se trabaja solamente con una muestra. A modo de resumen, los contrastes vistos fueron:

- contrastes sobre la proporción
  - contraste de proporción para una muestra
- contrastes sobre el promedio
  a) realizar contraste de Shapiro y Wilk
  b) ¿hay normalidad?
     sí $\longrightarrow$ contraste t para una muestra
     no $\longrightarrow$ contraste de Wilcoxon para una muestra

En esta práctica estudiaremos los contrastes en los que se comparan dos grupos. El esquema es:

- contrastes sobre proporciones
  - contraste de igualdad de proporciones para dos muestras
- contrastes sobre el promedio
  a) realizar test de normalidad de Shapiro-Wilk
  b) ¿hay normalidad en ambas?
     - sí; ¿cómo son las muestras?
       independientes $\longrightarrow$ contraste t para muestras independientes
       pareadas $\longrightarrow$ contraste t para datos relacionados
     - no; ¿cómo son las muestras?
       independientes $\longrightarrow$ contraste de Wilcoxon para dos muestras
       pareadas $\longrightarrow$ contraste de Wilcoxon para muestras pareadas

4.1. Comparación de proporciones

En muchas ocasiones es adecuado comparar dos grupos a través de la proporción (por ejemplo, la proporción de fumadores entre las mujeres y la proporción de fumadores entre los hombres), planteándose preguntas del tipo: ¿la proporción de los que tienen determinada característica es la misma en los dos grupos? El contraste de proporciones para dos muestras permite decidir si las diferencias observadas en las proporciones muestrales obtenidas son reales o simplemente son debidas a las fluctuaciones muestrales. Para poder realizar este test es necesario que los tamaños muestrales de ambas muestras sean razonablemente altos.

Ejemplo 4.1. *¿El porcentaje de horas en las que hay averías es menor en la línea A que en la B?*

Solución: Vamos a seguir los pasos habituales para la resolución de un contraste de hipótesis.

Identificar el contraste adecuado al problema

En este problema se pregunta por si un porcentaje es menor en una situación que en otra, por lo que estamos comparando dos proporciones y el contraste de hipótesis pertinente es el **contraste de proporciones para dos muestras**.

Establecer quiénes son H_0 y H_1 en ese contraste

Los distintos tipos de contrastes de hipótesis para comparar proporciones son:

$H_0 : p_A = p_B$ $H_1 : p_A \neq p_B$	$H_0 : p_A \geq p_B$ $H_1 : p_A < p_B$	$H_0 : p_A \leq p_B$ $H_1 : p_A > p_B$
`two.sided`	`less`	`greater`

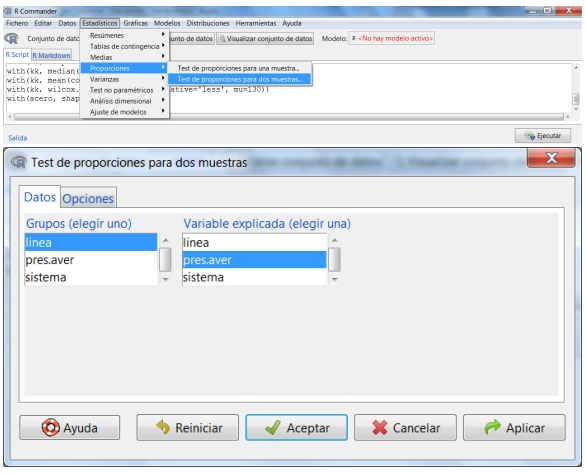
donde p_A y p_B son las dos proporciones en las poblaciones A y B, respectivamente. En nuestro caso, la muestra A estará constituida por aquellos individuos que cumplan `linea=="A"` y, la B, por los que cumplan `linea=="B"` (podría ser al revés).

Hemos de tener en cuenta que R-Commander considera, por defecto, que las proporciones p_A y p_B están asociadas a la primera clase por orden alfabético, en este caso la clase `A` de la variable `linea`. De este modo nuestras hipótesis son:

$H_0 : p_A \geq p_B$ (prop. de averías igual o mayor en A) $H_1 : p_A < p_B$ (prop. de averías menor en A)

Este constante se realizaría mediante:

Estadísticos
➥Proporciones
➥Test de proporciones para dos muestras



Seleccionar las variables `linea` y `pres.aver`
➥Marcar: Diferencia < 0
➥Aceptar

con lo que se obtiene

```
2-sample test for equality of proportions without continuity correction

data:  .Table
X-squared = 0.0673, df = 1, p-value = 0.3976
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000  0.1095155
sample estimates:
   prop 1    prop 2
0.2295082 0.2500000
```

Interpretar el p-valor
Como el p-valor ($0'3976$) es mayor que α no se rechaza la hipótesis nula; no hay evidencias de que vaya mejor la línea A que la B, en cuanto a la aparición de averías. □

Para aplicar el test de proporciones para dos muestras debemos tener en cuenta tres aspectos importantes:

- Identificar correctamente la variable explicada y la variable grupo. La variable explicada es el objeto de nuestro estudio y su comportamiento es comparado dentro de los grupos determinados por la variable grupo. Los valores de la variable grupo son los que solemos poner como subíndices en H_0 y H_1.
- Comprobar que la proporción representa aquello que queremos estudiar. Por defecto, R considera que p es la proporción del primer grupo (por orden alfabético). Si queremos estudiar la proporción del otro grupo, debemos reordenar los niveles del factor o, equivalentemente, expresar H_0 y H_1 en función de la proporción del primer grupo (por orden alfabético).
- El orden de los grupos en la variable grupo cuando se selecciona la hipótesis alternativa.

4.2. Comparación de varianzas

Un paso previo en un contraste de igualdad de medias es determinar si las varianzas en ambas poblaciones son o no iguales.[1] Para ello, R-Commander proporciona tres contrastes distintos: F para dos varianzas, Barlett y Levene. En estas prácticas usaremos el primero, el **contraste F para dos varianzas**, puesto que siempre que nos planteemos este contraste será dentro de un proceso en el que habremos previamente comprobado la normalidad.

Ejemplo 4.2. *¿Son distintas las varianzas de los consumos de las líneas A y B? (suponiendo normalidad)*

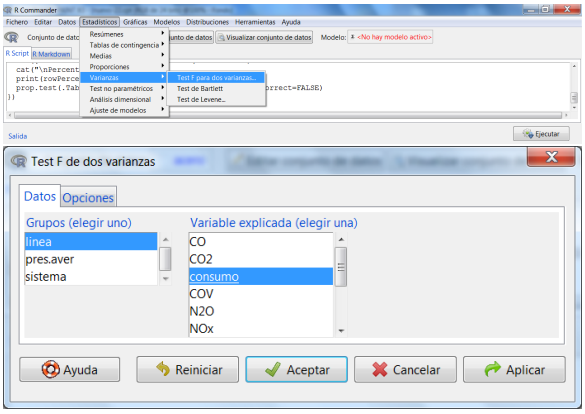
Solución: Las hipótesis para el contraste son

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ (varianzas iguales)}$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2 \text{ (varianzas distintas)}$$

[1] Se habla de *homoscedasticidad* si las dos varianzas son iguales y de *heteroscedasticidad* si no lo son.

por lo que tendremos que hacer

Estadísticos
➥Varianzas
➥Test F para dos varianzas



Seleccionar las variables `linea` y `consumo`
➥Aceptar

Los resultados de este procedimiento son:

```
       A        B
1431.355 2034.651


    F test to compare two variances

data:  consumo by linea
F = 0.7035, num df = 60, denom df = 55, p-value = 0.1834
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4158963 1.1828332
sample estimates:
ratio of variances
         0.7034893
```

Como el p-valor (0′1834) es mayor que α no se rechaza la hipótesis nula. Podemos suponer que no hay diferencias significativas entre las varianzas del consumo de la línea A y de la B, es decir, que se puede admitir que las dos muestras provienen de poblaciones con la misma varianza.

□

Este test se utiliza habitualmente como un contraste auxiliar en uno de los contrastes de igualdad de medias (el contraste t para dos muestras independientes que veremos más adelante). No obstante, en el ámbito de la ingeniería también tiene importancia en sí mismo, puesto que una estrategia básica para la mejora de la calidad pasa por la identificación de las causas que producen la variabilidad, para intentar reducirla. Por esta razón en muchas ocasiones se realizan test de comparación de varianzas (normalmente con hipótesis alternativas del tipo $H_1 : \sigma_1^2 > \sigma_2^2$ o $H_1 : \sigma_1^2 < \sigma_2^2$) para analizar si las medidas consideradas para reducir la variabilidad han surtido efecto.

4.3. Comparación de promedios: medias

Los contrastes para promedios nos permitirán comparar los promedios de dos variables aleatorias, a partir de los datos muestrales obtenidos. Para tal comparación hay que tener en

cuenta dos cuestiones fundamentales:

a) La relación que hay entre las dos muestras. Éstas pueden ser:

independientes: Se trata de dos muestras de forma que los individuos que pertenecen a una de ellas no pertenecen a la otra. En R-Commander, dos muestras independientes del mismo conjunto de datos requieren un factor dicótomo (de sólo dos niveles) para distinguirlas; la variable estadística (cuantitativa) bajo estudio estaría en una sola columna. Por ejemplo, imaginemos que se quiere estudiar el comportamiento de los consumos según la línea de producción utilizada; entonces los valores del consumo estarían en la columna `consumo` y la pertenencia de cada dato a una muestra u otra vendría dada por la variable `linea`.

pareadas: En este caso, cada individuo tendría un valor asociado en cada una de las dos muestras. En R-Commander, los datos correspondientes estarían dispuestos en dos columnas (cuantitativas)[2]. Se daría, por ejemplo, si queremos comparar, en promedio, las producciones de galvanizado tipo I y galvanizado tipo II.

b) Si se puede admitir o no la normalidad.

La siguiente tabla resume los diferentes contrastes de comparación de promedios que vamos a ver en las prácticas de laboratorio de esta asignatura:

Tabla 4.1: Contrastes para comparar dos promedios.

Comparación	¿Independientes?	¿Distribuciones aproximadamente normales?	Tipo de contraste
Diferencia de medias	Sí	Sí	Contraste t para muestras independientes
Media de la diferencia	No	Sí	Contraste t para datos relacionados
Diferencia de medianas[3]	Sí	No	Contraste de Wilcoxon para dos muestras
Mediana de la diferencia[3]	No	No	Contraste de Wilcoxon para muestras pareadas

Consideraremos en las dos siguientes subsecciones el caso de poblaciones normales y, en las dos últimas, el de poblaciones cualesquiera. En el caso de poblaciones normales se realizan contrastes sobre la media de la diferencia o, lo que es lo mismo, se compararán las medias de ambos grupos.

4.3.1. Muestras independientes con normalidad

Ejemplo 4.3. *¿El consumo promedio es menor en la línea A que en la B?*

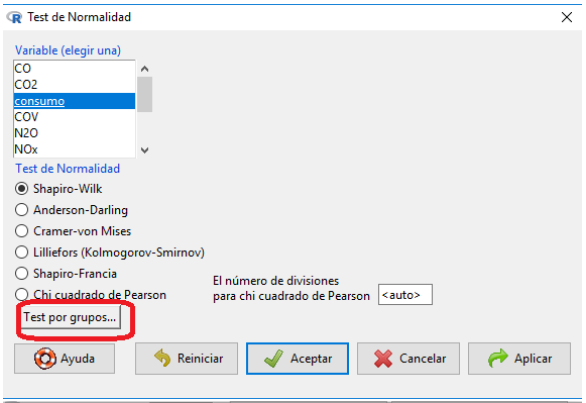
Demostración. La primera pregunta que nos hemos de hacer es si los datos son o no normales. Debemos contrastar, para cada uno de los dos casos, si las variables son normales utilizando el test de Shapiro-Wilk.

La opción más rápida consiste en aplicar el test de Shapiro-Wik contrastando la normalidad por grupos:

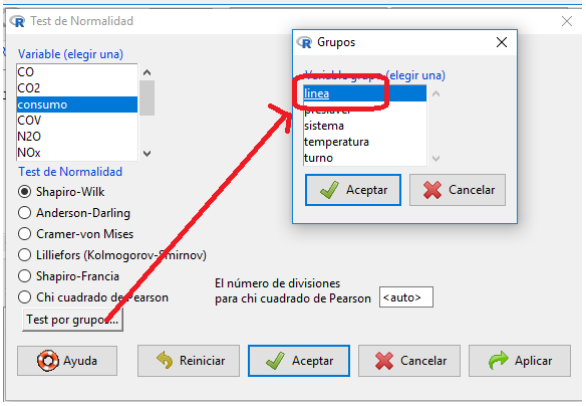
[2]Además de como muestras pareadas, también se las conoce como muestras apareadas, emparejadas o relacionadas.

[3]Bajo hipótesis de simetría.

Estadísticos
➥Resúmenes
➥Test de normalidad



Seleccionar la variable `linea`
➥Aceptamos



Como esta opción únicamente aparece en las últimas versiones de R, vemos a continuación otras formas de realizar el contraste. Una de ellas consiste en teclear en la *Ventana de instrucciones* la instrucción correspondiente, tal como sigue:

```
with(acero,by(consumo,linea,shapiro.test))
```

Otra manera de hacerlo sería crear dos conjuntos nuevos de datos mediante filtros, según la línea (A o B) utilizada, para después realizar el contraste de Shapiro-Wilk en cada uno de ellos para la variable `consumo`, a través de los menús *Estadísticos* → *Resúmenes* → *Test de normalidad de Shapiro-Wilk*.

Los resultados obtenidos son:

```
linea: A

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9708, p-value = 0.1534

------------------------------------------------------------------------
linea: B

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9746, p-value = 0.2841
```

Para los datos del consumo en la línea A el p-valor es 0′1534 y para los de la línea B es 0′2841. En ambos casos suficientemente grande como para no rechazar la hipótesis nula (se puede admitir la normalidad de los datos). Las poblaciones son independientes. Las varianzas son iguales en ambas poblaciones, como vimos en el Ejemplo 4.2. En estas circunstancias estamos en condiciones de aplicar el **contraste t para muestras independientes**, suponiendo las varianzas iguales.

Establecer quiénes son H_0 y H_1 en ese contraste

Para el **contraste t para muestras independientes** los distintos tipos de contrastes de hipótesis para la media son:

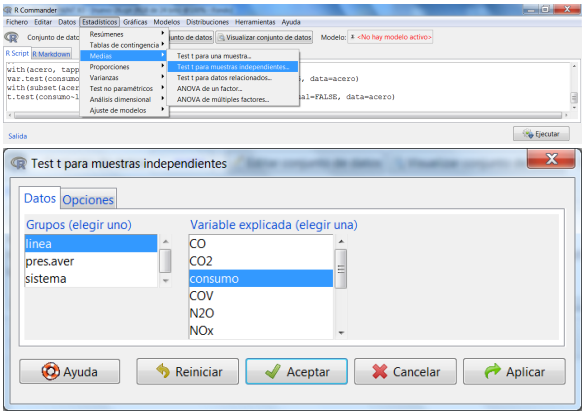
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : \mu_1 \geq \mu_2$ $H_1 : \mu_1 < \mu_2$	$H_0 : \mu_1 \leq \mu_2$ $H_1 : \mu_1 > \mu_2$

Para este ejemplo vamos a considerar:

$H_0 : \mu_A \geq \mu_B$ (consumo mayor o igual en la línea A)
$H_1 : \mu_A < \mu_B$ (consumo menor en la línea A)

y para calcularlo procedemos de la siguiente forma:

Estadísticos
➥Medias
➥Test t para muestras independientes



Seleccionar las variables `linea` y `consumo`
➥Marcar: Diferencias < 0
➥Varianzas iguales
➥Aceptamos

Los resultados son:

```
    Two Sample t-test

data:  consumo by linea
t = -10.1697, df = 115, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -65.32647
sample estimates:
mean in group A mean in group B
        98.3182         176.3716
```

Interpretar el p-valor

Como el p-valor ($< 2'2 \cdot 10^{-16}$) es menor que α se rechaza la hipótesis nula, es decir, hay evidencias significativas de que el consumo promedio es menor en la línea A que en la B.

□

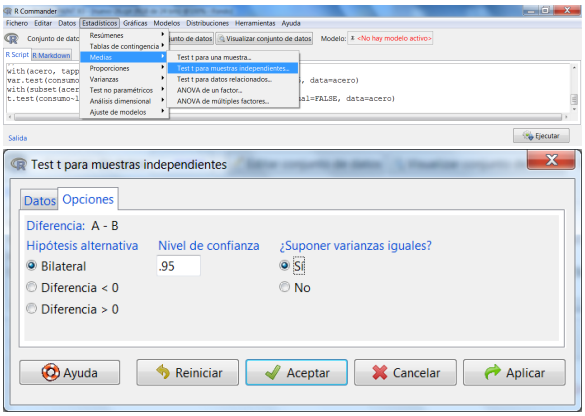
Realicemos ahora un contraste bilateral (el de igualdad de medias) en la misma situación.

Ejemplo 4.4. *¿El consumo promedio es distinto en ambas líneas de producción (A y B)?*

Solución: Las hipótesis en este caso serán

$$\begin{array}{|l|}\hline H_0 : \mu_A = \mu_B \text{ (consumo medio igual en las dos líneas)} \\ H_1 : \mu_A \neq \mu_B \text{ (consumos medios distintos)} \\ \hline\end{array}$$

Estadísticos
➥Medias
➥Test t para muestras independientes



Seleccionar las variables `linea` y `consumo`
➥Marcar: Bilateral
➥Varianzas iguales
➥Aceptar

Con lo que obtenemos:

```
    Two Sample t-test

data:  consumo by linea
t = -10.1697, df = 115, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -93.25631 -62.85051
sample estimates:
mean in group A mean in group B
        98.3182         176.3716
```

Como el p-valor ($< 2'2 \cdot 10^{-16}$) es de nuevo menor que α se rechaza la hipótesis nula, luego podemos considerar que hay diferencias significativas en el consumo medio de ambas líneas.

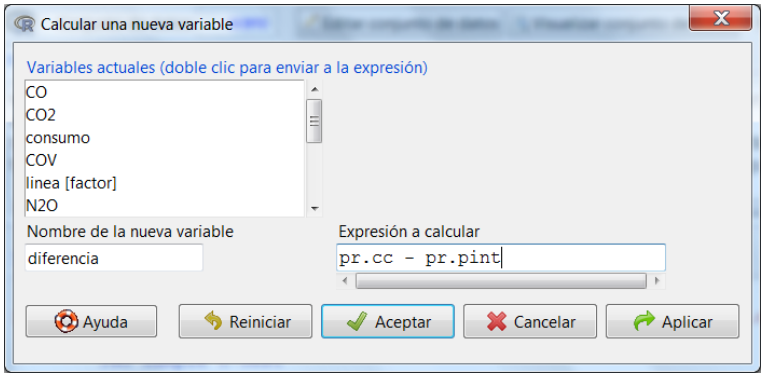
Evidentemente, que iba a quedar esta alternativa es lógico puesto que si hay evidencias de que la línea A consume en media menos que la B, también las tiene que haber de que ambos consumos son distintos.

□

4.3.2. Muestras dependientes con normalidad

Ejemplo 4.5. *Compárense, en promedio, la producción de colada continua y la producción de chapa pintada.*

Solución: En primer lugar, obtendremos la variable diferencia mediante la opción del menú *Datos* → *Modificar variables del conjunto de datos activo* → *Calcular una nueva variable...*. Rellenamos la ventana que se abre tal como puede verse a continuación:



con lo que generamos una nueva variable a la que le vamos a hacer la prueba de normalidad de Shapiro-Wilk. Para ello seguimos el procedimiento habitual: *Estadísticos* → *Resúmenes* → *Test de normalidad de Shapiro-Wilk*. Una vez en la ventana *Shapiro-Wilk Test for Normality* seleccionamos la variable *diferencia* y obtenemos el siguiente resultado:

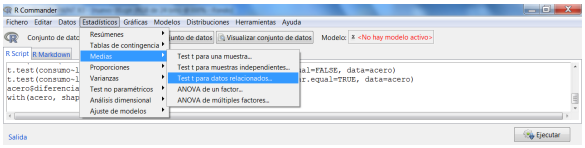
```
    Shapiro-Wilk normality test

data:  diferencia
W = 0.988, p-value = 0.3948
```

Como el p-valor es $0'3948$, mayor que α, podemos admitir la hipótesis de normalidad. Por tanto, podemos realizar el **contraste t para muestras relacionadas**, con el que contrastamos las hipótesis

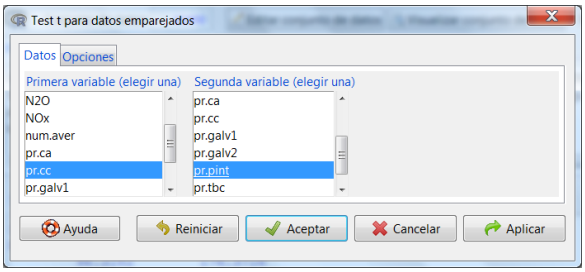
$$\boxed{\begin{array}{l} H_0 : \mu_{cc-pint} = 0 \\ H_1 : \mu_{cc-int} \neq 0 \end{array}}$$

Para realizar dicho test con `R` seguimos los siguientes pasos:

Estadísticos
➥Medias
➥Test t para muestras relacionadas



Primera variable (elegir una)
➥Elegir `pr.cc`
➥Segunda variable (elegir una)
➥Elegir `pr.pint`
➥Hipótesis alternativa
➥Bilateral



El resultado es

```
        Paired t-test

data:  pr.cc and pr.pint
t = 2.5405, df = 116, p-value = 0.01239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  18.56348 149.91515
sample estimates:
mean of the differences
               84.23932
```

en donde se aprecia un p-valor de $0'01239$, menor que $\alpha = 0'05$, por lo que las medias de `pr.cc` y `pr.pint` son significativamente distintas.

□

4.3.3. Muestras independientes sin normalidad

Supongamos que estamos interesados en estudiar el promedio de la diferencia de dos variables (dos características de los individuos de una población o una característica examinada en dos muestras procedentes de sendas poblaciones) y que la hipótesis de normalidad no es admisible. En ese caso, podemos utilizar el test de Wilcoxon.

Ejemplo 4.6. *Estudiemos el comportamiento de la producción de galvanizado 2 en función de la línea de producción.*

Solución: En primer lugar efectuaremos la prueba de normalidad de Shapiro-Wilk a cada muestra de la variable `pr.galv2` según el valor de `linea`. Para ello, seleccionamos en primer lugar aquellos valores de producción con la línea A (Figura 4.1).

Seguidamente, realizamos el contraste de Shapiro y Wilk, para el que se obtiene el siguiente resultado:

```
    Shapiro-Wilk normality test

data:  pr.galv2
W = 0.8955, p-value = 7.985e-05
```

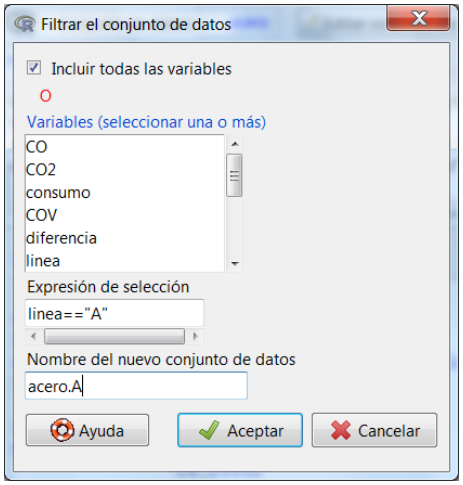


Figura 4.1: Filtro para seleccionar los datos de la línea A.

En este caso, no se puede admitir la normalidad de los datos, es decir, hay evidencias en contra de suponer que la producción de galvanizado tipo 2 en la línea A es normal.

Como para usar el test t ambas variables deben ser normales, ya podemos concluir que vamos a trabajar con el test de Wilcoxon acerca de la mediana de la diferencia.

Antes de nada, debemos volver a considerar el conjunto de datos completo:

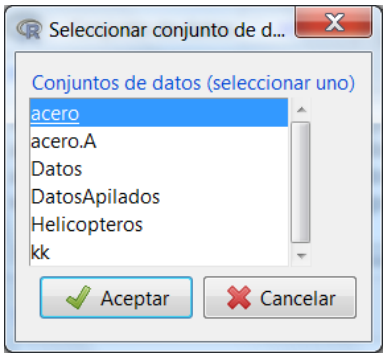


Figura 4.2: Volver al conjunto de datos con todos los valores.

Antes de continuar, obsérvese que otra forma más rápida de realizar el contraste anterior, sin necesidad de hacer los sucesivos filtros, sería escribiendo en la ventana de instrucciones los siguientes comandos:

```
with(acero,by(pr.galv2,linea,shapiro.test))
```

En general, se ha optado por el trabajo mediante menús, para recurrir lo menos posible a la programación, pero en este caso, la opción de la programación sería considerablemente más rápida.

Supongamos que estamos interesados en un contraste para comparar los promedios:

$H_0 : Me_A - Me_B = 0$ (la producción es igual, en promedio)
$H_1 : Me_A - Me_B \neq 0$ (la producción difiere en ambas líneas, en promedio)

Si realizamos el **contraste de Wilcoxon para dos muestras** (*Estadísticos* → *Test no paramétricos* → *Test de Wilcoxon para dos muestras*, Figura 4.3).

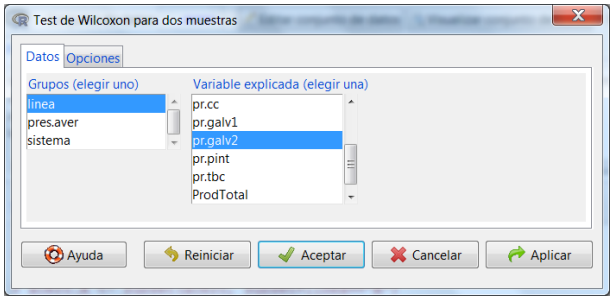


Figura 4.3: Test de Wilcoxon para dos muestras independientes

El resultado es:

```
        Wilcoxon rank sum test with continuity correction

data:  pr.galv2 by linea
W = 1431, p-value = 0.1314
alternative hypothesis: true location shift is not equal to 0
```

Así pues, no hay evidencia de que el promedio de la producción de galvanizado de tipo 2 varíe, en promedio, en función de la línea de producción, pues el p-valor (0′1314) es mayor que cualquier nivel de significación α razonable.

□

4.3.4. Muestras dependientes sin normalidad

Ejemplo 4.7. *Compárense, en promedio, la producción de galvanizado de tipo 1 y la de tipo 2.*

Solución: En primer lugar, obtenemos la variable diferencia mediante la opción del menú *Datos* → *Modificar variables del conjunto de datos activo* → *Calcular una nueva variable...*, tal como vimos en el Ejemplo 4.5 (página 61). A la nueva variable la llamaremos `dif`.

Los resultados de test de normalidad de Shapiro-Wilk para la muestra de la variable `dif` son:

```
        Shapiro-Wilk normality test

data:  dif
W = 0.9671, p-value = 0.005665
```

Se rechaza la normalidad suponiendo $\alpha = 0'05$, no habiendo lugar a dudas para ningún nivel de significación α de los habituales, puesto que el p-valor sale muy pequeño. Por lo tanto, no vamos a realizar un contraste t para muestras relacionadas, sino que vamos a recurrir al **test de Wilcoxon para muestras pareadas**. En concreto, vamos a resolver el contraste de comparación de promedios siguiente:

$H_0 : Me_{X_1-X_2} = 0$ (la producción de ambos galvanizados coincide, en promedio)
$H_1 : Me_{X_1-X_2} \neq 0$ (la producción de ambos galvanizados difiere, en promedio)

Procedemos siguiendo los menús *Estadísticos* → *Test no paramétricos* → *Test de Wilcoxon para muestras pareadas* y rellenando la ventana que se abre, tal como puede verse en la Figura 4.4.

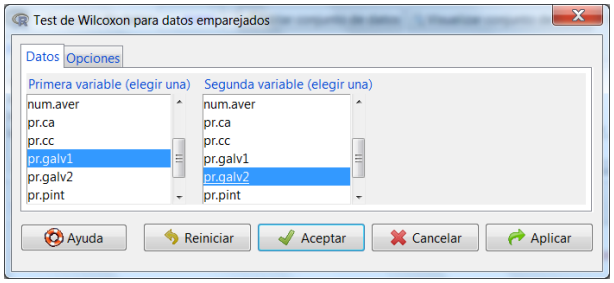


Figura 4.4: Test de Wilcoxon para muestras pareadas.

El resultado de este procedimiento es:

```
        Wilcoxon signed rank test with continuity correction

data:  pr.galv1 and pr.galv2
V = 249, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

El p-valor del contraste es $< 2'2 \cdot 10^{-16} \approx 0$, menor que cualquier nivel de significación α razonable, por lo que se concluye que la producción de ambos tipos de galvanizado es significativamente distinta, en promedio.

□

4.4. Ejercicios propuestos

Ejercicio 4.1. *Responda razonadamente a las siguientes cuestiones:*

a) Realiza un contraste para determinar si el porcentaje de horas que el sistema de detección de sobrecalentamiento está apagado es mayor en la línea A que en la B. ¿Cuál es su p-valor? En función de dicho p-valor, ¿qué se concluye?

b) Realiza un contraste para determinar si el porcentaje de horas que el sistema de detección de sobrecalentamiento está encendido es menor en la línea A que en la B. ¿Cuál es su p-valor? En función de dicho p-valor, ¿qué se concluye?

c) En la línea A, ¿qué porcentaje de horas el sistema de detección de sobrecalentamiento estuvo apagado?

d) En la línea B, ¿qué porcentaje de ellas el sistema de detección de sobrecalentamiento estuvo apagado?

e) *¿Cuál de los dos porcentajes anteriores (apartados b) y c)) es mayor? Comenta los resultados.*

Ejercicio 4.2. *Se quiere comparar el consumo promedio con el sistema de detección de sobrecalentamiento encendido y apagado.*

a) *¿Qué test sería adecuado para ello? Detalla los procedimientos que has llevado a cabo para contestar a esta pregunta.*

b) *Si la hipótesis alternativa en el contraste es que el consumo medio con el sistema apagado es mayor que con el sistema encendido ¿cuál es el p-valor asociado a dicho test? ¿qué se concluye?*

c) *Realiza alguna representación gráfica que permita ilustrar las conclusiones obtenidas en el apartado anterior.*

Ejercicio 4.3. *Se quiere comparar la producción promedio de colada continua y del convertidor de acero.*

a) *¿Qué test sería el adecuado para comparar ambas producciones promedio? Detalla los procedimiento que has llevado a cabo para contestar a esta pregunta.*

b) *Para esta muestra, ¿cuánto vale la producción media de colada continua? ¿y la del convertidor de acero?*

c) *Si la hipótesis alternativa en el contraste es que la producción media de colada continua es mayor que la producción media del convertidor de acero ¿cuál es el p-valor asociado a dicho test? ¿qué se concluye?*

Ejercicio 4.4. a) *¿La producción del convertidor de acero es mayor, en promedio, cuando el sistema de detección de sobrecalentamiento está encendido que cuando está apagado?*

b) *¿Cuánto vale la mediana muestral de la producción del convertidor de acero cuando el sistema está apagado? ¿y cuándo está encendido?*

c) *Realiza un gráfico en el que puedan compararse la producción del convertidor de acero con el sistema apagado y encendido, en la muestra.*

Ejercicio 4.5. a) *¿La producción del convertidor de acero es menor, en promedio, que la del tren de bandas calientes?*

b) *¿Cuánto vale la mediana muestral de la producción del convertidor de acero? ¿y la del tren de bandas calientes?*

4.5. Soluciones de los ejercicios propuestos

Ejercicio 4.1. a) *Haciendo un test de igualdad de proporciones ($H_0 : p_A \leq p_B$ frente a la alternativa $H_1 : p_A > p_B$) se obtiene que:*

```
     sistema
linea  OFF   ON Total Count
    A 50.8 49.2   100    61
    B 50.0 50.0   100    56


2-sample test for equality of proportions without continuity correction

data:  .Table
X-squared = 0.0078, df = 1, p-value = 0.4647
alternative hypothesis: greater
95 percent confidence interval:
 -0.1439993  1.0000000
sample estimates:
   prop 1    prop 2
0.5081967 0.5000000
```

El p-valor es $0'4647$ *con lo que no se puede asegurar que el % sea significativamente mayor en la línea A, las diferencias muestrales pueden ser debidas al azar.*

b) *La respuesta es la misma que en el apartado anterior. Para llegar a esta respuesta se deben reordenar los niveles del factor o expresar las hipótesis nula y alternativa en función del porcentaje de horas que el sistema de detección de sobrecalentamiento está apagado.*

c) *Según vemos en la tabla del apartado anterior, el porcentaje de veces que el sistema está apagado en la línea A es* $50'8\,\%$.

d) *De forma análoga obtenemos que el porcentaje de veces que el sistema está apagado en la línea B es* $50\,\%$.

e) *El % de veces en la muestra que el sistema está apagado es mayor para la línea A que para la B, no obstante las diferencias no son lo suficientemente grandes como para poder concluir que esto pasa en general en la población.*

Ejercicio 4.2. *a*)

- *Son muestras independientes.*
- *¿Normalidad?*

```
> with(acero,by(consumo,sistema,shapiro.test))
sistema: OFF

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9798, p-value = 0.4319

------------------------------------------------------------
sistema: ON

    Shapiro-Wilk normality test
```

```
data:  dd[x, ]
W = 0.9757, p-value = 0.2958
```

Las dos se pueden suponer poblaciones normales.

- *¿Varianzas iguales?*

```
   F test to compare two variances

data:  consumo by sistema
F = 0.9125, num df = 58, denom df = 57, p-value = 0.7292
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5410492 1.5372658
sample estimates:
ratio of variances
         0.9125423
```

Se pueden suponer varianzas iguales.

De todo lo anterior se deduce que el test adecuado es el test t para muestras independientes, con la opción varianzas iguales activada.

b) `Two Sample t-test`

```
data:  consumo by sistema
t = 2.1912, df = 115, p-value = 0.01523
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.518213      Inf


sample estimates:
mean in group OFF  mean in group ON
         146.9241          124.2362
```

El p-valor es $0'01523$ *con lo que se rechaza la hipótesis nula al nivel* $\alpha = 0'05$*, es decir, el consumo medio con el sistema apagado es significativamente mayor que con el sistema encendido.*

c) *Aunque otros gráficos serían más adecuados (diagrama de barras de error, por ejemplo) por tratarse de un test que compara medias, vamos a hacer el diagrama de cajas para OFF y ON, puesto que es el más rápido de los dos que hemos visto en la primera práctica para variables continuas.*

Ejercicio 4.3. *a) Son muestras relacionadas, así que comenzamos obteniendo la variable diferencia que llamaremos* `difccca` *a la que pasamos el test de normalidad.*

Creación de la variable diferencia:

```
> acero$difccca <- with(acero, pr.cc-pr.ca)
```

Test de normalidad:

```
    Shapiro-Wilk normality test

data:  difccca
W = 0.979, p-value = 0.06339
```

Como sale normal (p-valor $= 0'06339$*), utilizaremos el test t para datos relacionados.*

b) Repitiendo el procedimiento del ejercicio 1.4 de la página 23, se obtiene que la producción media muestral de cc es $433'93$ *toneladas y la de ca es de* $244'92$ *toneladas.*

c) Los resultados obtenidos son:

```
        Paired t-test

data:  pr.cc and pr.ca
t = 5.6404, df = 116, p-value = 6.079e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 133.4459      Inf
sample estimates:
mean of the differences
               189.0085
```

Con un p-valor $6'079 \cdot 10^{-8}$*, está claro que hay evidencias significativas de que la producción media de cc es mayor que la de ca.*

Ejercicio 4.4. *a) El p-valor para el test de normalidad aplicado a los datos de producción del convertidor de acero con el sistema apagado es* $0'002512$*, con lo que se rechaza la normalidad de una de las variables. Como hemos considerado el criterio de usar el test t sólo si hay normalidad en ambas, tenemos que usar el test de Wilcoxon. En este caso el test de Wilcoxon para dos muestras.*

Si consideramos las hipótesis:

H_0:	*la producción con el sistema apagado es mayor o igual que cuando está encendido, en promedio*
H_1:	*la producción con el sistema apagado es menor que cuando está encendido, en promedio*

El p-valor asociado a este contraste es $0'07351$*, por lo que no hay evidencias significativas que indiquen que la producción, en promedio, es mayor con el sistema encendido que con él apagado.*

b) Cuando está apagado 179 *y cuando está encendido* 241*.*

c) *Por ejemplo un diagrama de cajas para la variable* `pr.ca` *con variable de agrupación el* `sistema`.

Ejercicio 4.5. *a*) *El p-valor del test de normalidad para la diferencia (*`pr.ca-pr.tbc`*) es* $1'892 \cdot 10^{-7}$*, con lo que rechazamos la normalidad. El p-valor del test de Wilcoxon para muestras pareadas es menor de* $2'2 \cdot 10^{-16}$*, con lo que hay evidencias suficientes para afirmar que la producción del convertidor de acero es significativamente menor que la producción del tren de bandas calientes, en promedio.*

b) *La mediana de las producciones del convertidor de acero en la muestra es de 225T y la del tren de bandas calientes es 8062 toneladas.*

Práctica 5

Contrastes de independencia y correlación lineal

Muchas veces nos interesa analizar si existe o no relación entre dos variables, como por ejemplo si el salario inicial depende del tipo de estudios realizados o si el porcentaje de piezas defectuosas depende de la línea de producción utilizada. Para contestar a esta pregunta existen una serie de contrastes, dos de los cuales van a ser presentados a continuación. Evidentemente estos contrastes no son los únicos que existen para realizar este tipo de análisis, pero tienen la virtud de ilustrar los principios estadísticos de contrastes similares. Los que vamos a estudiar en este curso son:

- **Test de independencia Chi-cuadrado**: contrasta la relación estadística genérica, es decir, si hay dependencia o independencia entre las dos variables. En general, se utiliza cuando las dos variables son cualitativas o categóricas (*factores* en R-Commander).
- **Test de correlaciones de Pearson**: contrasta la existencia o no de relación lineal entre las dos variables. Para poder aplicarse ambas variables tienen que ser cuantitativas.

Téngase especial cuidado al escoger el contraste para el caso de dos variables cuantitativas discretas, pues podría interesar considerarlas como factores. Las variables cuantitativas continuas requerirían agrupamiento en intervalos para poder usarse como factores, por lo que no es habitual someterlas al contraste de independencia chi-cuadrado.

En el caso de querer estudiar la relación entre una variable cualitativa y otra cuantitativa, téngase en cuenta el número de categorías (niveles) de la variable cualitativa. Si es dicótoma (con dos niveles) entonces el contraste correspondiente sería el de promedio nulo de la diferencia en dos muestras independientes: la hipótesis nula de igualdad correspondería al caso de no relación; la hipótesis alternativa de desigualdad correspondería al caso de relación entre las variables. Si no es dicótoma, hay que recurrir a métodos estadísticos que no entran dentro de los contenidos mínimos de este curso, como el análisis de varianza (ANOVA) o el contraste de Kruskal y Wallis.

De hecho, muchos de los contrastes analizados con anterioridad pueden englobarse dentro de una estructura común de análisis de relaciones entre variables. Así, si denominamos **variable explicada** a la variable cuyo comportamiento se quiere comprender y que puede estar asociada a otra variable y **variable explicativa** o **grupo** a la variable que asume una cierta "influencia" sobre la variable explicada o, por lo menos, se asocia a cambios en dicha variable, en la Tabla 5.1 se presentan dentro de este esquema general varios tipos de contrastes.

VARIABLE EXPLICATIVA O GRUPO	VARIABLE EXPLICADA	TIPO DE TEST E HIPÓTESIS NULA
CUALITATIVA O CATEGÓRICA CON SOLO DOS CATEGORÍAS	CUANTITATIVA O NUMÉRICA	Contraste para el promedio de la diferencia (t y Wilcoxon) H_0: el promedio de la diferencia es cero. Esto puede considerarse equivalente a decir que no hay relación entre las variables, es decir, la variable explicativa no influye en la variable explicada.
Ejemplo: Estudio sobre si el peso medio es igual en hombres y mujeres.		
CUALITATIVA O CATEGÓRICA CON MÁS DE DOS CATEGORÍAS	CUANTITATIVA O NUMÉRICA	ANOVA de un factor o Kruskal-Wallis H_0: hay el mismo comportamiento en todos los grupos, en promedio. Esto puede considerarse equivalente a decir que no hay relación entre las variables, es decir, la variable explicativa no influye en la variable explicada.
Ejemplo: Estudio sobre si el peso medio es igual en la gente de España, USA y Japón.		
CUALITATIVA O CATEGÓRICA CON DOS CATEGORÍAS	CUALITATIVA O CATEGÓRICA CON DOS CATEGORÍAS	Contraste de proporciones para dos muestras H_0: las proporciones de la variable explicada en cada uno de los dos grupos o categorías de la variable explicativa son iguales. Esto puede considerarse equivalente a decir que no hay relación entre las variables, es decir, la variable explicativa no influye en la variable explicada.
Ejemplo: Estudio sobre si el porcentaje de fumadores es igual en hombres y en mujeres.		
CUALITATIVA O CATEGÓRICA CON VARIAS CATEGORÍAS	CUALITATIVA O CATEGÓRICA CON VARIAS CATEGORÍAS	Contraste de independencia chi-cuadrado H_0: la distribución de probabilidad de la variable en estudio es igual en cada grupo o categoría de la variable explicativa. Esto puede considerarse equivalente a decir que no hay relación entre las variables, es decir, la variable explicativa no influye en la variable explicada.
Ejemplo: Estudio sobre si el porcentaje de fumadores es igual en España, USA y Japón.		
CUANTITATIVA O NUMÉRICA	CUANTITATIVA O NUMÉRICA	Contraste de correlación de Pearson H_0: No hay relación lineal entre ambas variables.
Ejemplo: Estudio sobre si existe relación lineal entre el peso y la estatura.		

Tabla 5.1: Algunos contrastes para el análisis de relación entre variables.

En líneas generales y tal como ya hemos visto anteriormente, podemos establecer un protocolo en tres pasos, para realizar un contraste a partir de los datos muestrales. En el caso particular de los dos contrastes tratados en esta práctica los pasos a seguir serían:

1. **Seleccionar el contraste adecuado a la muestra.**

   Si contrastamos la existencia o no de independencia, es decir, si existe algún tipo de relación estadística general o no entre las dos variables consideradas, entonces seleccionamos el test de independencia Chi-cuadrado.

   Si nos interesa estudiar si hay relación lineal entre dos variables cuantitativas, entonces seleccionamos el test de correlación de Pearson.

2. **Establecer quiénes son H_0 y H_1 en ese contraste.**

   Las hipótesis por contrastar son siempre del tipo:

H_0:	no existe relación entre las variables
H_1:	sí existe relación entre las variables

En particular, en el caso del contraste de independencia chi-cuadrado, estas hipótesis pueden concretarse como sigue:

H_0:	hay independencia estadística entre las dos variables
H_1:	hay dependencia estadística entre las dos variables

En el caso del contraste de correlación, las hipótesis análogas serían:

H_0:	hay independencia lineal entre las dos variables (la correlación lineal entre las variables es nula)
H_1:	hay dependencia lineal entre las dos variables (la correlación lineal entre las variables no es nula)

No obstante, el R-Commander también permite plantear en este caso contrastes donde la hipótesis alternativa no es sólo de correlación distinta de cero, sino también de correlación positiva (valores altos de una de las variables hacen esperar valores altos de la otra) y de correlación negativa (valores altos de una de las variables hacen esperar valores bajos de la otra).

3. **Interpretar el p-valor.**

   Así pues, un p-valor claramente menor del nivel de significación α indicará que existe relación entre las variables. En caso contrario se concluye que los datos no nos proporcionan evidencias estadísticas de dicha relación.

5.1. Independencia

El test de independencia Chi-cuadrado, nos permite determinar si existe relación estadística entre dos variables categóricas. Es necesario resaltar que esta prueba nos indica si existe o no una relación entre las variables, pero no indica el grado o el tipo de relación; es decir, no indica el porcentaje de influencia de una variable sobre la otra o la variable que causa la influencia.

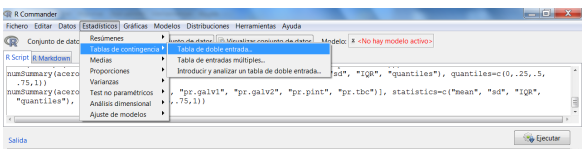
Este procedimiento puede verse como una generalización a más de dos muestras del contraste de igualdad de proporciones.

Vamos a explicar el funcionamiento de este contraste a través de un ejemplo.

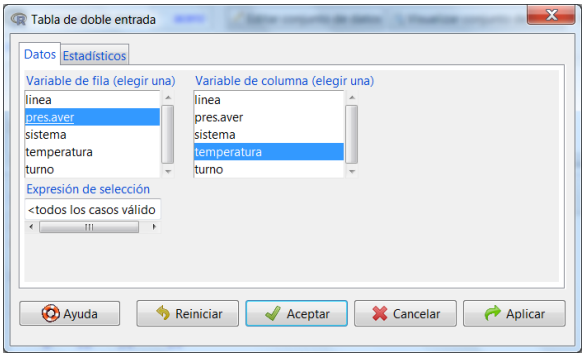
Ejemplo 5.1. *¿Existe relación entre que haya o no averías y la temperatura? o, dicho de otro modo, ¿la proporción de averías depende de la temperatura?*

Solución: Como las variables son cualitativas vamos a utilizar el **Test de independencia Chi-cuadrado**. Para hacer esto vamos a

Estadísticos
➥Tablas de contingencia
➥Tabla de doble entrada...



Seleccionar las variables `pres.aver` y `temperatura`
➥Aceptar



Con esto obtenemos las siguientes salidas:

```
Frequency table:
         temperatura
pres.aver Alta Baja Media
      A      8   14     6
      NoA   38   24    27

   Pearson's Chi-squared test

data:  .Table
X-squared = 5.1595, df = 2, p-value = 0.07579
```

En esta salida vemos que el coeficiente chi-cuadrado, que mide la relación entre ambas variables en la muestra, toma el valor 5′1595. Como el p-valor (0′07579) es mayor que el nivel de significación habitual ($\alpha = 0'05$), no se rechaza la hipótesis nula. Por lo tanto concluimos que no hay evidencias estadísticas de que la temperatura afecte a que haya o no averías.

□

A la hora de obtener el p-valor para este contraste, se utiliza una aproximación a la distribución chi-cuadrado. Para que dicha aproximación sea buena, es necesario que se cumplan ciertas condiciones; entre ellas, que las frecuencias esperadas (las que tendría que haber habido en cada grupo en caso de que fuese cierta la hipótesis de independencia (H_0)) no sean demasiado pequeñas. Suele asumirse que si existen frecuencias esperadas menores que 5, éstas no pueden superar el 20 % del total de frecuencias en la tabla. En el caso de que dicha condición no se cumpla, existe la convención de proceder a agrupar categorías de las variables de la tabla hasta solventar el problema, en cuyo caso se vuelve a obtener el correspondiente p-valor. En caso contrario, el valor del p-valor debe ser interpretado con cautela.

Para comprobar si ciertas condiciones se están verificando es conveniente seleccionar la opción `Imprimir las frecuencias esperadas` en la ventana anterior: `Tabla de doble entrada`, con el fin de ver si realmente dichas frecuencias esperadas son o no menores de 5.

Ejemplo 5.2. *En el caso del ejemplo anterior, el resultado obtenido sería*

```
Expected Counts:
         temperatura
pres.aver     Alta      Baja     Media
      A   11.00855  9.094017  7.897436
      NoA 34.99145 28.905983 25.102564
```

Así, podemos ver como en este caso no hay ninguna casilla con frecuenca esperada menor de 5 y, por tanto, podemos utilizar el p-valor obtenido para sacar conclusiones.

Aparte de las frecuencias esperadas, también se puede pedir al ordenador, entre otras cosas, que calcule los porcentajes totales, por filas o por columnas de los datos de la muestra, de forma que nos dé información sobre el porcentaje de individuos en cada categoría dentro del total de individuos, dentro de los individuos que forman esa fila o esa columna.

Ejemplo 5.3. *En el ejemplo anterior si seleccionamos la opción* `Porcentajes totales` *obtendríamos además de los resultados vistos en el Ejemplo 5.1, la siguiente tabla:*

```
Total percentages:
      Alta Baja Media Total
A      6.8 12.0   5.1  23.9
NoA   32.5 20.5  23.1  76.1
Total 39.3 32.5  28.2 100.0
```

Dicha tabla nos informa, por ejemplo, de que:

- *el* $32'5\,\%$ *de las horas de la muestra no hubo averías y la temperatura era alta,*
- *el* $12\,\%$ *de las horas de la muestra sí hubo averías y la temperatura era baja,*
- *el* $76'1\,\%$ *de las horas de la muestra no hubo averías y*
- *el* $39'3\,\%$ *de las horas de la muestra la temperatura era alta.*

Si en lugar de marcar la opción de porcentajes totales, marcamos la opción `Porcentajes por filas`*, lo que obtendríamos sería:*

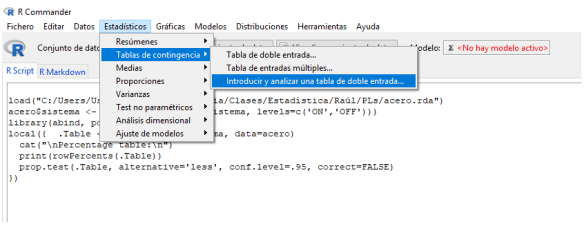
```
Row percentages:
         temperatura
pres.aver Alta Baja Media Total Count
      A   28.6   50  21.4   100    28
      NoA 42.7   27  30.3   100    89
```

que nos informa, por ejemplo, de que dentro de las 89 horas de la muestra en las que no hubo avería, el $42'7\,\%$ *de las veces la temperatura fue alta, el* $27\,\%$ *de las veces baja y el* $30'3\,\%$ *restante fue media.*

Por otro lado, con la opción `Porcentajes por columnas` *obtendríamos dentro de cada temperatura, qué porcentaje de veces hubo y no hubo averías.*

Nótese que se puede introducir la tabla de contingencia directamente en R y aplicar el test de independencia χ^2 a los datos allí representados. Para esto, debemos seleccionar la opción *Introducir y analizar una tabla de doble entrada.*

Estadísticos
➥Tablas de contingencia
➥Introducir y analizar una tabla de doble entrada



Determinamos el tamaño de la tabla eligiendo el número de filas (número de valores distintos que toma la primera variable) y el número de columnas (número de valores distintos que toma la segunda variable) e introducimos las frecuencias conjuntas en la tabla. El resto de opciones en el test son las opciones explicadas con anterioridad.

5.2. Correlación

El test de independencia chi-cuadrado se usa para comprobar la independencia de dos caracteres estadísticos. Bajo ciertas condiciones de normalidad (y solamente en ese caso) dos variables aleatorias son independientes si y sólo si ellas no están correlacionadas, es decir, si el coeficiente de correlación de Pearson (ρ) toma el valor cero. Sin embargo, en general, el coeficiente de correlación mide solo la relación de tipo lineal (línea recta). Dos variables pueden tener una relación curvilínea fuerte, a pesar de que su correlación sea pequeña.

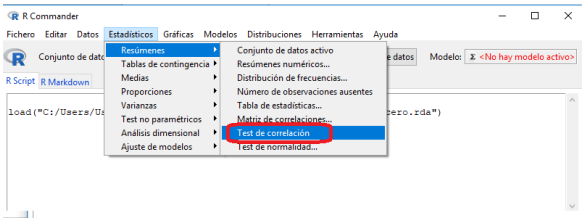
Así, se puede cuantificar la fuerza de la relación lineal entre dos variables cuantitativas por medio de la estimación del coeficiente de correlación de Pearson. Dicho coeficiente oscila entre -1 y $+1$. Un valor de ± 1 indica una relación lineal o línea recta perfecta (positiva en el caso $\rho = 1$ y negativa en el caso $\rho = -1$). Una correlación próxima a cero indica que no hay relación lineal entre las dos variables.

El problema aquí es decidir si una correlación observada entre dos variables, medidas en los mismos individuos, es o no significativa. La respuesta a esta cuestión se obtiene mediante el test de correlación de Pearson. Vamos a ver cómo realizar este procedimiento con R-Commander a través de un ejemplo.

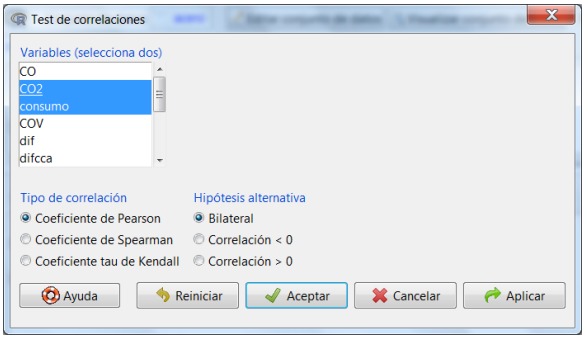
Ejemplo 5.4. *En el conjunto de datos* `acero.rda` *se encuentran los datos relativos a emisión de gases con efecto invernadero recogidos por otro equipo técnico de la misma empresa. En función de estos datos, ¿qué se puede decir sobre si existe o no relación lineal entre el consumo energético y la emisión de dióxido de carbono?*

Solución: Como las variables son cuantitativas continuas, podemos plantearnos la utilización del **test de correlación de Pearson**, para lo cual haremos:

Estadísticos
➥Resúmenes
➥Test de correlación



Seleccionar las variables `CO2` y `consumo`
➥Aceptar



El resultado de este procedimiento es:

```
Pearson's product-moment correlation
```

```
data:  CO2 and consumo
t = 35.1003, df = 115, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9376074 0.9695667
sample estimates:
      cor
0.9563613
```

Como el p-valor es menor que $2'2 \cdot 10^{-16}$, es menor que cualquier nivel de significación habitual α, por lo que se rechaza la hipótesis nula. Así, podemos concluir que hay evidencias estadísticas de relación lineal entre el consumo y la emisión de CO2.

Además de esto se puede observar que la estimación puntual del coeficiente de correlación entre ambas variables es $0'9563613$, lo que nos lleva a tener una confianza del 95 % de que el verdadero valor del coeficiente de correlación poblacional está entre $0'9376074$ y $0'9695667$ (información incorporada en la salida del procedimiento anterior, justo después de especificar la hipótesis alternativa: `95 percent confidence interval: 0.9376074 0.9695667`). De ahí también se puede observar que hay evidencias para rechazar la hipótesis de no relación lineal, puesto que ésta se traduce en un coeficiente de correlación nulo, el cual es un valor inadmisible de acuerdo con este intervalo de confianza.

Puesto que la estimación del coeficiente de correlación es $0'9563613$ es muy cercana a uno y el p-valor muy pequeño, se puede considerar que el grado de relación lineal entre ambas variables es muy alto. Además, puesto que dicha estimación es positiva y el p-valor asociado al contraste $H_0 : \rho \leq 0$ frente a la alternativa $H_1 : \rho > 0$ es de nuevo menor que $2'2 \cdot 10^{-16}$ (marcando la opción `Correlación > 0` en la ventana `Test de correlaciones`), se obtiene que la relación lineal entre ambas variables es positiva, es decir, consumos energéticos altos hacen esperar emisiones altas de de CO2 y consumos bajos emisiones bajas. □

Al calcular el p-valor asociado a este contraste se está suponiendo que la distribución conjunta de ambas variables es normal bivariada[1]. No hay una forma completamente satisfactoria de comprobar lo razonable de la suposición de normalidad bivariada. Una comprobación parcial consiste en realizar contrastes de normalidad para cada una de las dos variables, ya que la normalidad bivariada implica que las distribuciones de cada una de las dos variables sean normales. Si en cualquiera de los dos contrastes se rechaza la normalidad, los contrastes anteriores respecto al coeficiente de correlación de Pearson no pueden emplearse cuando el tamaño muestral n es pequeño. En tal caso se trabaja con un coeficiente de correlación no paramétrico, como por ejemplo el coeficiente de correlación de Spearman, que tiene el mismo significado que el coeficiente de correlación de Pearson, pero se calcula utilizando el rango de las observaciones.

Ejemplo 5.5. *En el ejemplo anterior tienen sentido las conclusiones obtenidas a partir de test de correlaciones, puesto que los resultados de la prueba de normalidad de Shapiro-Wilk para ambas variables son:*

```
   Shapiro-Wilk normality test

data:  consumo
W = 0.9884, p-value = 0.4207
```

[1] La normal bivariada o bidimensional es una distribución que se escapa de los objetivos de este curso.

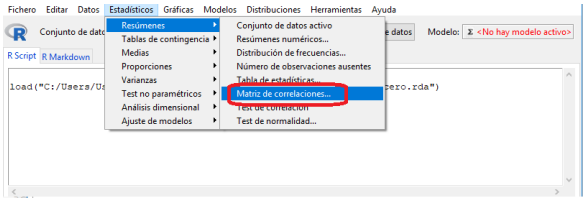
```
   Shapiro-Wilk normality test

data:  CO2
W = 0.9924, p-value = 0.771
```

por lo que en ambos casos es admisible la hipótesis de normalidad.

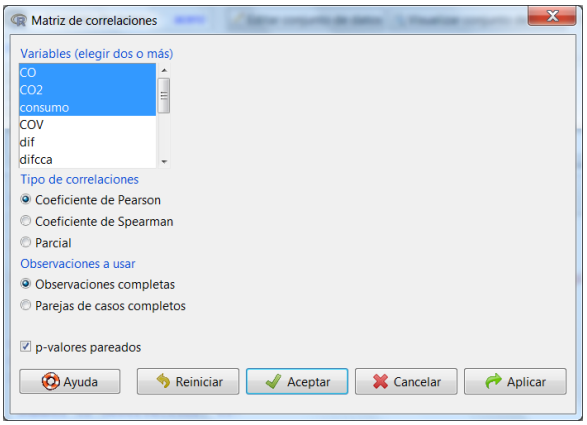
En muchas ocasiones nos interesa, dadas varias variables, identificar la que tiene más relación lineal con otra variable dada, es decir, mayor coeficiente de correlación. Para analizar esto, es muy habitual representar los coeficientes de correlación de Pearson de cada par de variables de forma matricial, mediante la matriz de correlaciones. La forma de obtener dicha matriz, así como los p-valores asociados a los coeficientes de correlación presentes en la misma, puede verse a través del siguiente ejemplo.

Ejemplo 5.6. *En el ejemplo anterior podemos querer analizar si el consumo energético tiene o no relación lineal con la emisión de* `CO`*,* `CO2` *y* `SO2`*. Con el objetivo de obtener los resultados necesarios (coeficientes de correlación y p-valores asociados) de una sola vez, deberíamos seguir los siguientes pasos:*

Estadísticos
➥*Resúmenes*
➥*Matriz de correlaciones...*



Seleccionamos `CO`*,* `CO2`*,* `consumo` *y* `SO2`*.*
➥*Coeficiente de Pearson*
➥*p-valores pareados*
➥*Aceptar*



Con lo que se obtienen las siguientes salidas:

```
 Pearson correlations:
            CO    CO2 consumo    SO2
CO       1.0000  0.9442  0.9198  0.0444
CO2      0.9442  1.0000  0.9564 -0.0286
consumo 0.9198  0.9564  1.0000 -0.0076
SO2      0.0444 -0.0286 -0.0076  1.0000

 Number of observations = 117


 Pairwise two-sided p-values:
```

```
         CO      CO2     consumo SO2
CO               <.0001 <.0001  0.6347
CO2      <.0001         <.0001  0.7599
consumo  <.0001 <.0001          0.9352
SO2      0.6347 0.7599 0.9352
```

Adjusted p-values (Holm's method)

```
          CO      CO2     consumo SO2
CO                <.0001 <.0001   1
CO2       <.0001         <.0001   1
consumo   <.0001 <.0001           1
SO2        1       1       1
```

De la segunda matriz (p-valores), podemos deducir que existen evidencias de relación lineal entre el consumo energético y la emisión de CO *o la de* CO2*, pero no así con la de* SO2[2]*. De las dos variables con relación lineal significativa, la mayor relación muestral es con la emisión de* CO2*, puesto que el coeficiente de correlación (primera matriz) es* $0'9564$ *en lugar de* $0'9198$*.*

El estudio de la correlación suele ser un primer paso para determinar la relación entre las variables y, en función de ella, predecir el valor de una variable dado un valor determinado de la otra. Dichas técnicas, conocidas como regresión, serán abordadas en la siguiente práctica.

5.3. Ejercicios propuestos

Responda razonadamente a las siguientes cuestiones, en función de los datos recogidos en el conjunto de datos `acero.rda`.

Ejercicio 5.1. *a) ¿Existe relación estadística entre la temperatura y que el sistema de detección de sobrecalentamiento esté encendido o apagado?*

b) Dentro de la muestra, ¿cuántas veces ha habido temperatura alta y el sistema ha estado apagado? ¿cuántas ha estado encendido y ha habido temperatura media?

c) Si realmente hubiese independencia estadística entre estas dos variables, ¿cuántas veces, de las 117 analizadas, se esperaría que hubiese habido temperatura alta y el sistema estuviese apagado?

d) Dentro de las horas en las que se ha trabajado con temperatura media, ¿qué porcentaje de ellas el sistema de detección de sobrecalentamiento estaba encendido?

Ejercicio 5.2. *a) Sin realizar ninguna operación con los datos, ¿qué test se podría utilizar para analizar la relación entre el consumo y la producción total, el de independencia chi-cuadrado o el de correlación de Pearson?*

b) ¿Se verifican los requerimientos mínimos para aplicar dicho test?

[2]No es rechazable la hipótesis de normalidad para ninguna de las cuatro variables implicadas, puesto que los p-valores del test de normalidad de Shapiro-Wilk para cada una de ellas, obtenidos a partir de los datos muestrales, son: p-valor(CO)=$0'1485$, p-valor(CO2)=$0'771$, p-valor(SO2)=$0'2773$ y p-valor(consumo) =$0'4207$. Como consecuencia de esto, tiene sentido interpretar los p-valores asociados a los contrastes de correlación de Pearson.

c) En función del p-valor, ¿qué se puede decir acerca de la relación lineal entre el consumo y la producción total?

d) ¿Cuánto vale la estimación puntual del coeficiente de correlación de Pearson (ρ)? En función de este valor, ¿qué se espera que ocurra con el consumo al aumentar la producción total, que aumente o que disminuya?

e) En función del intervalo de confianza de ρ obtenido, ¿sería admisible considerar que el verdadero coeficiente de correlación es $0'75$?

f) De todas las emisiones de gases, ¿cuáles tienen relación significativa con el consumo energético?, ¿cuál es la más relacionada con ella en la muestra?

Ejercicio 5.3. *a) ¿Es la presencia de averías independiente del turno?*

b) ¿Cuál es la proporción de averías que ocurrieron en el turno de noche? ¿y en los turnos de mañana y de tarde?

Ejercicio 5.4. *¿Existe una relación lineal entre las producciones de acero galvanizado de tipo 1 y de tipo 2?*

5.4. Solución de los ejercicios propuestos

Ejercicio 5.1. *a) El p-valor del test de independencia chi-cuadrado de $0'9471$, por lo que no hay evidencia estadística de que el sistema esté más veces encendido con unas temperaturas que con otras. Dicho p-valor es fiable, porque se verifican las condiciones necesarias para la aplicación del test de independencia chi-cuadrado, puesto que las frecuencias esperadas son:*

```
> .Test$expected # Expected Counts
            sistema
temperatura      OFF       ON
      Alta  23.19658 22.80342
      Baja  19.16239 18.83761
      Media 16.64103 16.35897
```

y, por tanto, todas ellas son mayores o iguales que 5.

b) Dentro de la muestra, 24 ha habido temperatura alta y el sistema ha estado apagado y 17 ha estado encendido y ha habido temperatura media.

c) Si realmente hubiese independencia estadística entre estas dos variables, se esperaría que hubiese habido $23'19658$ horas con temperatura alta y el sistema estuviese apagado, de las 117 analizadas.

d) Dentro de las horas en las que se ha trabajado con temperatura media, el $51'5\,\%$ de ellas el sistema de detección de sobrecalentamiento estaba encendido.

Ejercicio 5.2. *a) El test de correlación de Pearson, porque son dos variables continuas.*

b) *El p-valor del test de normalidad de Shapiro-Wilk para la variable* `consumo` *es* $0'4207$ *y para la variable* `producción total` *es* $0'8543$*, por lo que en ambos casos es admisible la hipótesis de normalidad, con lo que se verifican los requerimientos mínimos para aplicar el test de correlación de Pearson.*

c) *El p-valor del test de correlación de Pearson es menor que* $2'2 \cdot 10^{-16}$*, por lo que existen evidencias significativas de que* $\rho \neq 0$*, es decir, de existencia de relación lineal entre el consumo y la producción total.*

d) *La estimación puntual del coeficiente de correlación de Pearson es* $R = 0'9496154$*. Como este valor es positivo, se espera que el consumo aumente al aumentar la producción total.*

e) *Como el intervalo de confianza al 95 % de* ρ *es* $(0'9280690, 0'9648255)$*, un valor de* ρ *igual a* $0'75$ *no es admisible.*

f) *Todas tienen relación lineal significativa menos* `SO2`*, puesto que la matriz asociada de p-valores del test de correlación de Pearson es:*

```
         CO     CO2    consumo COV    N2O    NOx    SO2
CO               <.0001 <.0001  <.0001 <.0001 <.0001 0.6347
CO2      <.0001        <.0001  <.0001 <.0001 <.0001 0.7599
consumo  <.0001 <.0001          <.0001 <.0001 <.0001 0.9352
COV      <.0001 <.0001 <.0001          <.0001 <.0001 0.7414
N2O      <.0001 <.0001 <.0001  <.0001        <.0001 0.9398
NOx      <.0001 <.0001 <.0001  <.0001 <.0001        0.1753
SO2      0.6347 0.7599 0.9352  0.7414 0.9398 0.1753
```

y se dan las condiciones que hacen fiables dichos p-valores, puesto que los p-valores del test de normalidad de Shapiro-Wilk son:

```
   Shapiro-Wilk normality test

data:  CO
W = 0.9831, p-value = 0.1485
```

```
   Shapiro-Wilk normality test

data: CO2
W = 0.9924, p-value = 0.771
```

```
   Shapiro-Wilk normality test

data:  consumo
W = 0.9884, p-value = 0.4207
```

```
   Shapiro-Wilk normality test

data:  COV
W = 0.9944, p-value = 0.9229
```

```
   Shapiro-Wilk normality test

data:  N2O
W = 0.9922, p-value = 0.7518
```

```
   Shapiro-Wilk normality test

data:  NOx
W = 0.9797, p-value = 0.07302
```

```
   Shapiro-Wilk normality test

data:  SO2
W = 0.9862, p-value = 0.2773
```

lo que hace admisible la suposición de que todas las variables en estudio son normales. Además, simplemente el elevado número de datos ($n = 117$) en la muestra, ya permitiría obtener conclusiones fiables a partir de los contrastes de correlación de Pearson.

Con la que más relación tiene en la muestra es con `CO2`*, puesto que es la de mayor coeficiente de correlación en valor absoluto, puesto que la matriz de correlaciones asociada es:*

```
             CO     CO2 consumo     COV     N2O     NOx     SO2
CO       1.0000  0.9442  0.9198  0.9950  0.8196  0.5195  0.0444
CO2      0.9442  1.0000  0.9564  0.9650  0.8540  0.5685 -0.0286
consumo  0.9198  0.9564  1.0000  0.9334  0.8274  0.5384 -0.0076
COV      0.9950  0.9650  0.9334  1.0000  0.8359  0.5344  0.0308
N2O      0.8196  0.8540  0.8274  0.8359  1.0000  0.5317  0.0071
NOx      0.5195  0.5685  0.5384  0.5344  0.5317  1.0000 -0.1262
SO2      0.0444 -0.0286 -0.0076  0.0308  0.0071 -0.1262  1.0000
```

Ejercicio 5.3. *a) Puesto que tanto* `pres.aver` *como* `turno` *son variables cualitativas, tenemos que aplicar el test χ^2 de independencia. Obtenemos un p-valor de $1'527e-15$,*

por lo que concluimos que hay evidencias significativas de que las dos variables NO son independientes.

b) *Si la variable* `pres.aver` *aparece como fila, seleccionamos* `porcentajes por fila` *en las opciones del test* χ^2 *de independencia. Observamos que el* $100\,\%$ *de las averías ocurrieron en el turno de noche, y, por tanto, un* $0\,\%$ *de las averías ocurrieron en los turnos de mañana y de tarde.*

Ejercicio 5.4. *Puesto que tanto* `pr.galv1` *como* `pr.galv2` *son variables cuantitativas, tenemos que aplicar el test de correlación de Pearson. Para comprobar que podemos aplicar este test, primero tenemos que estudiar la normalidad de ambas variables. Obtenemos los p-valores 0.00957 para la variable* `pr.galv1` *y 0.0000003397 para la variable* `pr.galv2` *por lo que rechazamos la normalidad y, por tanto, no podemos aplicar el test de correlación de Pearson. Si aplicamos el test de Spearman, obtenemos un p-valor de 0.127. Puesto que este valor es mayor que el nivel de significación fijado (0.05), concluimos que no hay evidencias significativas de que las dos variables estén correladas.*

Práctica 6

Regresión lineal

Para muchas aplicaciones en ingeniería se necesita modelar las relaciones entre conjuntos de variables. Por ejemplo, el rendimiento de un proceso en función de la temperatura y la presión a las cuales se llevan a cabo las reacciones, la demanda máxima diaria de una planta generadora de energía eléctrica como función del número de clientes, el oxígeno disuelto en muestras de agua de un lago en función del contenido de algas, etc.

Hemos visto en la práctica anterior una forma de analizar dichas relaciones, mediante el contraste de correlación de Pearson. Ahora vamos más allá, nos interesa además desarrollar un método de predicción, es decir, un procedimiento para estimar el valor de una de las variables en función de los valores de la o las otras, a partir de la información experimental. El aspecto estadístico del problema se convierte, entonces, en lograr la mejor estimación de la relación entre las variables.

Así, en algunas situaciones podríamos conocer qué cambios deben hacerse en ciertas variables controlables, para obtener los mejores resultados posibles para otra variable no controlable directamente. Por ejemplo, podríamos analizar a qué temperatura y presión se obtiene el mejor rendimiento, etc.

El **análisis de regresión lineal** es una técnica estadística utilizada para estudiar y modelar la relación entre variables cuantitativas. En la mayoría de los casos la verdadera ecuación de predicción se desconoce y el investigador debe elegir una función adecuada para aproximarla. Decimos adecuada y no óptima, porque habitualmente no tendremos constancia de que el modelo no pueda ser mejorado. No obstante, además de buscar un modelo que se ajuste bien, también suele buscarse un modelo que sea lo más sencillo posible. Así, el principio de parsimonia (la navaja de Ockham) induce a optar por un modelo sencillo en vez de uno complicado. Dado un conjunto de posibles explicaciones igualmente buenas, según este criterio, la más sencilla se convierte en la mejor.

En cualquiera de los modelos considerados existe una distinción clara entre las variables en lo que respecta a su papel en el proceso experimental. Muy a menudo existe una sola **variable explicada**, dependiente o respuesta, que no se controla en el experimento. Esta respuesta depende de una o más **variables explicativas**, independientes o de regresión. El caso de una sola variable explicada y una sola variable explicativa, se denomina **regresión lineal simple** y es el que vamos a tratar en esta práctica. Otros modelos más complicados, como por ejemplo los modelos con varias variables explicativas (regresión lineal múltiple) se escapan de los objetivos de este curso.

En cualquiera de los casos, el procedimiento de análisis de regresión puede describirse en cuatro pasos:

1. Suponer la forma que tiene el promedio de la variable explicada en función de las variables explicativas.
2. Utilizar los datos de la muestra para estimar los parámetros desconocidos del modelo.
3. Comprobar estadísticamente la adecuación del modelo obtenido.
4. Usar el modelo para realizar predicciones, estimaciones, etc., cuando se haya llegado a la conclusión de que puede considerarse adecuado.

6.1. Paso 1: Búsqueda de un modelo

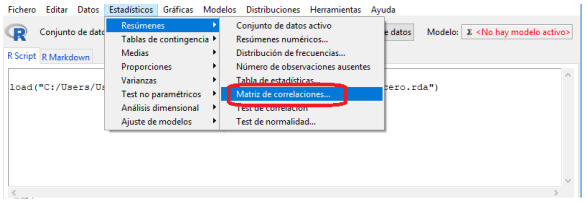
El principio de parsimonia indica que el modelo de regresión lineal simple se convierte en el primer candidato para explicar la relación entre las variables.

Entre las posibles candidatas a variable explicativa del modelo, debemos elegir la más adecuada. Para determinar esto, se suele comenzar estudiando la existencia o no de relación lineal significativa entre la variable explicada y cada una de las posibles variables explicativas, así como estimando los coeficientes de correlación lineal entre ellas. Una representación gráfica que suele acompañar a este estudio son los correspondientes diagramas de dispersión.

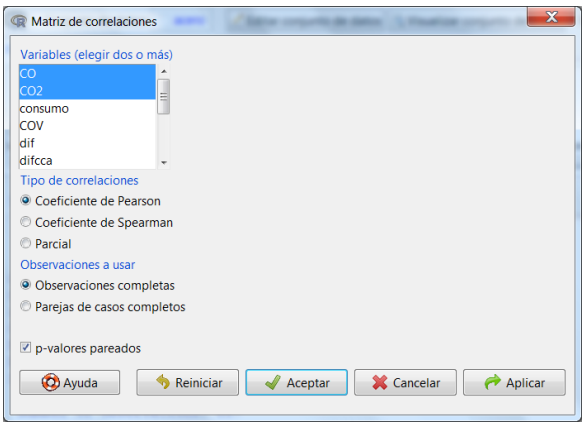
Ejemplo 6.1. *Supongamos que tenemos los datos relativos a emisión de sustancias contaminantes que aparecen en el archivo* `acero.rda` *y que estamos interesados en conocer una aproximación a la emisión de* `N2O` *conocida la emisión de alguno de estos otros gases:* `CO`*,* `CO2`*,* `NOx` *y* `SO2`*. Así pues, queremos predecir la emisión de óxido nitroso (*`N2O`*) en función de la emisión de óxidos de monóxido de carbono (*`CO`*), dióxido de carbono (*`CO2`*), mezcla de óxidos de nitrógeno (*`NOx`*) o dióxido de azufre (*`SO2`*), ¿cuál de estas cuatro variables es la mejor variable explicativa?*

Solución: Comenzamos obteniendo la correspondiente matriz de correlaciones:

Estadísticos
➥Resúmenes
➥Matriz de correlaciones...



Seleccionamos `CO`, `CO2`, `N2O`, `NOx` y `SO2`.
➥Coeficiente de Pearson
➥p-valores pareados
➥Aceptar



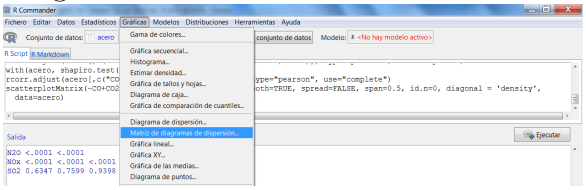
El resultado de este procedimiento es:

```
Pearson correlations:
        CO     CO2    N2O     NOx     SO2
CO  1.0000  0.9442 0.8196  0.5195  0.0444
CO2 0.9442  1.0000 0.8540  0.5685 -0.0286
N2O 0.8196  0.8540 1.0000  0.5317  0.0071
NOx 0.5195  0.5685 0.5317  1.0000 -0.1262
SO2 0.0444 -0.0286 0.0071 -0.1262  1.0000

 Number of observations: 117

 Pairwise two-sided p-values:
    CO     CO2    N2O    NOx    SO2
CO         <.0001 <.0001 <.0001 0.6347
CO2 <.0001        <.0001 <.0001 0.7599
N2O <.0001 <.0001        <.0001 0.9398
NOx <.0001 <.0001 <.0001        0.1753
SO2 0.6347 0.7599 0.9398 0.1753

Adjusted p-values (Holm's method)

    CO     CO2    N2O    NOx    SO2
CO         <.0001 <.0001 <.0001 1.0000
CO2 <.0001        <.0001 <.0001 1.0000
N2O <.0001 <.0001        <.0001 1.0000
NOx <.0001 <.0001 <.0001        0.7011
SO2 1.0000 1.0000 1.0000 0.7011
```

La tercera fila de la primera tabla muestra los coeficientes de correlación de N2O con las demás variables. Se observa como el coeficiente mayor es con CO2. En la siguiente tabla se muestran los p-valores del test de correlación de Pearson. En la tercera fila se obtienen todos los p-valores prácticamente nulos menos el último que es significativamente mayor que cualquiera de los niveles de significación α habituales, con lo cual se concluye que N2O tiene relación lineal significativa con CO, CO2 y NOx, pero no con SO2[1]. De las tres con las que tiene relación, con la que la estimación del coeficiente de correlación es mayor, en valor absoluto, es con CO2, con lo que consideramos ésta como variable explicativa de N2O.

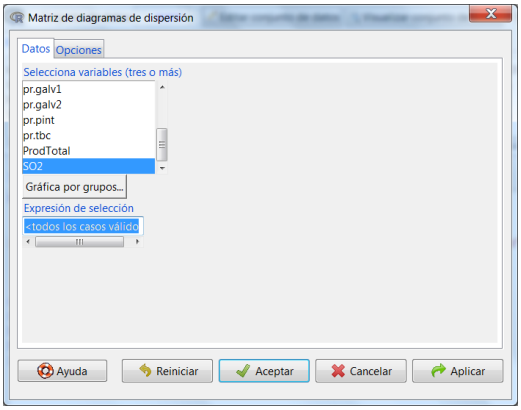
Gráficamente, estas conclusiones se pueden apoyar con la correspondiente matriz de diagrama de dispersión. Dicha matriz se obtiene como sigue:

Gráficas
➥Matriz de diagrama de dispersión...

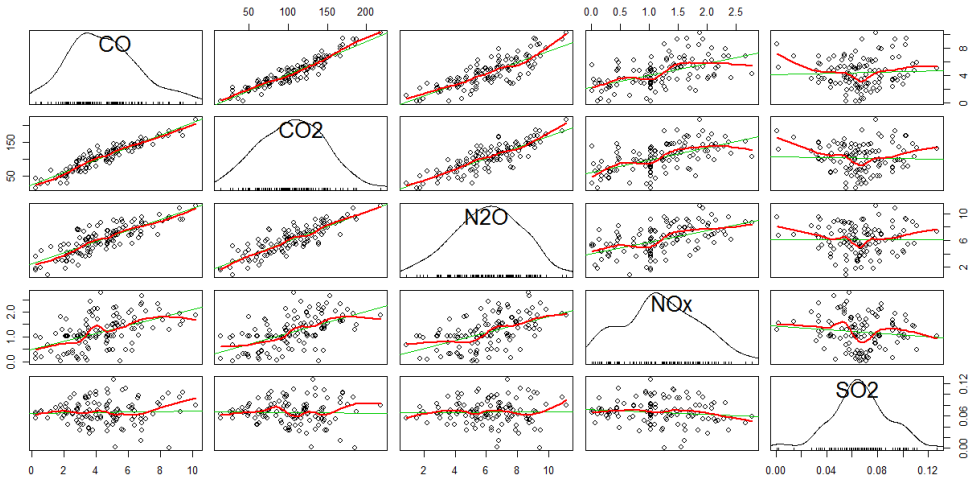


[1]No es rechazable la hipótesis de normalidad para ninguna de las cinco variables implicadas, puesto que los p-valores del test de normalidad de Shapiro-Wilk para cada una de ellas, obtenidos a partir de los datos muestrales, son: p-valor(CO)=0′1485, p-valor(CO2)=0′771, p-valor(N2O)=0′7518, , p-valor(NOx)=0′07302 y p-valor(SO2)=0′2773. Como consecuencia de esto, tiene sentido interpretar los p-valores asociados a los contrastes de correlación de Pearson.

Seleccionamos `CO`, `CO2`, `N2O`, `NOx` y `SO2`.
➥Aceptar



Los resultados de dicho procedimiento son:



De los diferentes gráficos que aparecen, los que más nos interesan en este caso se encuentran en la tercera fila, ya que la variable explicada o dependiente, el `N2O`, se suele representar en el eje de ordenadas, mientras que la explicativa o independiente, una de los otros tipos de emisiones, se suele representar en el eje de abscisas.

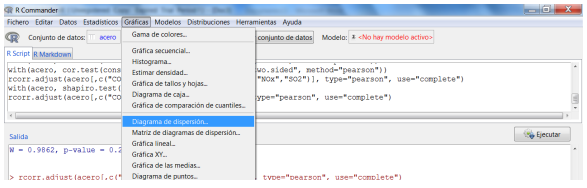
¿Qué diagrama de dispersión de la tercera fila muestra un patrón más claro de relación? Si bien usualmente no se puede responder de forma concluyente a esta pregunta a través de estos gráficos, se ve claramente en este caso cómo no parece haber relación con `SO2`, no hay una relación lineal clara con `NOx` y la relación lineal es fuerte con `CO` y `CO2`. □

Una vez determinada cuál va a ser la variable explicativa, se pasa a realizar un diagrama de dispersión ya sólo para estas dos variables, explicativa y explicada, con el fin de estudiar si realmente el plantearse un modelo de regresión lineal simple tiene sentido, es decir, si el ajuste por una recta es adecuado para estos datos.

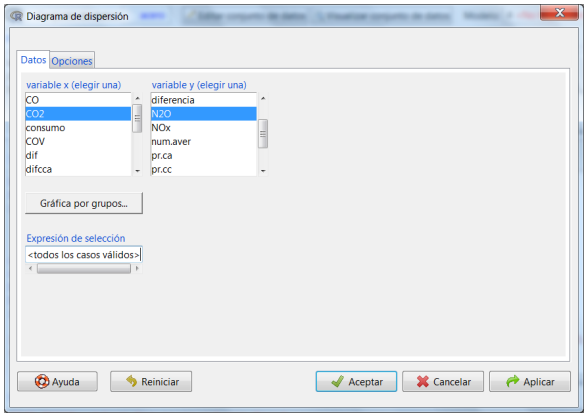
Ejemplo 6.2. *Dibuje el diagrama de dispersión con la emisión de óxido nitroso (`N2O`) en el eje de ordenadas y la emisión de dióxido de carbono (`CO2`) en el eje de abscisas.*

Solución: El gráfico se consigue de la siguiente forma:

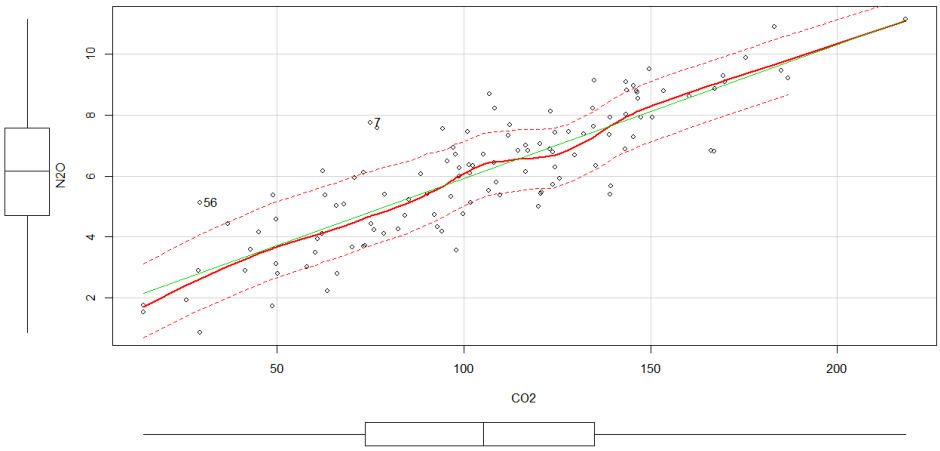
Gráficas
➥Diagrama de dispersión...



Seleccionamos: `CO2` y `N2O`
➥Aceptar



El gráfico obtenido con este procedimiento es:



El eje de abscisas muestra la emisión de `CO2` y el de ordenadas la de `N2O`. Se observa una relación creciente entre ambas magnitudes. En el gráfico aparecen dos líneas. Una es la recta de regresión (el modelo más simple) y la otra la línea de regresión no paramétrica (el mejor ajuste posible a los datos, respecto de mínimos cuadrados). Si ambas líneas son muy similares, el ajuste lineal resulta adecuado. En este caso la línea recta sigue muy bien el comportamiento de la línea no paramétrica, por lo que el modelo lineal parece ajustar bien los datos. □

6.2. Paso 2: Estimación del modelo

Una vez elegida una variable como variable explicativa, el siguiente paso sería la estimación de los parámetros del modelo de regresión (en este caso, recta de regresión). Dicha estimación se basa en el método de mínimos cuadrados: se seleccionan como estimadores de los parámetros aquellos que hacen mínima la suma de los errores elevados al cuadrado ($n-2$ veces el cuadrado del *residual standard error*), que se denotará por SSE. En el caso particular de regresión lineal simple, esto quiere decir que se eligen como estimadores de β_0 y β_1 aquellos valores que minimizan

$$SSE = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2,$$

donde n representa el tamaño de la muestra (número de pares de observaciones).

Vamos a ver la forma de obtener dichas estimaciones a través de un ejemplo.

Ejemplo 6.3. *Estime la emisión de* `N2O` *a partir de la emisión de* `CO2`.

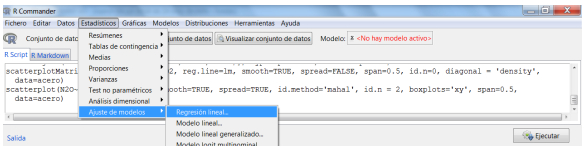
Solución: Puesto que el diagrama de dispersión obtenido en el Ejemplo 6.2 sugería que el modelo lineal se ajustaba bien a los datos, procedemos a construir un modelo lineal que cuantifica la relación entre `N2O` y `CO2`, permitiendo en algunos casos la estimación de la primera, conocido el valor de la segunda. Así pues, buscamos estimar los coeficientes β_0 y β_1 tales que

$$\texttt{N2O} = \beta_0 + \beta_1 \cdot \texttt{CO2} + \epsilon$$

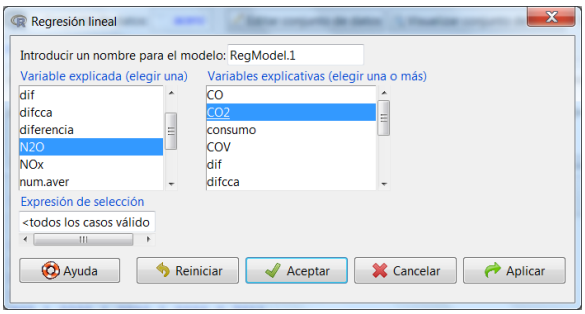
donde ϵ denota el error aleatorio.

Para ello los pasos a seguir son:

Estadísticos
➥Ajuste de modelos
➥Regresión lineal



Variable explicada: `N2O`
➥Variables explicativas: `CO2`
➥Aceptar



Adicionalmente le podemos poner un nombre al modelo resultante. Si se deja el nombre por defecto, los resultados que aparecen en la `Ventana de resultados` son los siguientes:

```
Call:
lm(formula = N2O ~ CO2, data = acero)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2585 -0.7287  0.0404  0.6511  2.9353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.526865   0.280149    5.45 2.91e-07 ***
CO2         0.043850   0.002491   17.60  < 2e-16 ***
---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 115 degrees of freedom
Multiple R-squared: 0.7293,Adjusted R-squared: 0.7269
F-statistic: 309.8 on 1 and 115 DF,  p-value: < 2.2e-16
```

La columna `Estimate` proporciona los valores de las estimaciones de los coeficientes, con lo que el modelo de regresión lineal simple que mejor se ajusta a estos datos es:

$$\texttt{N2O} = \ 1'526865 \ + \ 0'043850 \ \cdot \texttt{CO2} \tag{6.1}$$

Así pues, la estimación del valor del parámetro β_0 (*Intercept*) es $1'526865$, que se interpreta como la emisión estimada o el promedio de emisión de `N2O` si no hubiese emisión de `CO2`. Este coeficiente es significativo (distinto de 0), puesto que el p-valor asociado al contraste $H_0 : \beta_0 = 0$ frente a la alternativa $H_1 : \beta_0 \neq 0$ es $2'91 \cdot 10^{-7}$.

La estimación de β_1 es $0'043850$, coeficiente de nuevo significativo, puesto que el p-valor asociado al contraste $H_0 : \beta_1 = 0$ frente a la alternativa $H_1 : \beta_1 \neq 0$ es inferior a $2 \cdot 10^{-16}$. Así pues, por cada unidad que se incremente la emisión de `CO2`, la emisión promedio de `N2O` se espera que se incremente en $0'043850$ unidades. □

6.3. Paso 3: Adecuación del modelo

El ajuste de un modelo de regresión requiere de varias suposiciones. Por ejemplo, la estimación de los parámetros del modelo requiere la suposición de que los errores son variables aleatorias no correlacionadas con media cero y varianza constante; los contrastes de hipótesis ya vistos y las estimaciones por intervalo que veremos más adelante requieren que los errores se ajusten a una distribución normal; etc. Además se supone que el grado de ajuste es bueno.

El analista debe realizar siempre comprobaciones acerca de la adecuación del modelo antes de comenzar a utilizarlo para hacer predicciones o pronósticos. Una forma de medir dicha adecuación será:

- Mediante el coeficiente de determinación ajustado. Éste debe ser lo más cercano a uno posible.
- Mediante los niveles críticos de los contrastes asociados a los coeficientes, en este caso al coeficiente de regresión (β_1). Dicho coeficiente debería ser significativo, es decir, significativamente distinto de cero.
- Mediante el análisis de los residuos o errores. Como hemos comentado deben ser variables aleatorias normales no correlacionadas con media cero y varianza constante. Un conocimiento del cumplimiento/incumplimiento de dichas condiciones será imprescindible para poder interpretar los resultados obtenidos mediante la regresión en su justo término. No obstante un estudio completo de estas condiciones se escapa de los objetivos de este curso. Por esta razón, no vamos a detenernos con detalles al respecto, vamos a conformarnos con un estudio gráfico de la verificación o no de dichas hipótesis. Las opciones más básicas para ello están en *Modelos* → *Gráficos* → *Gráficas básicas de diagnóstico*.

  De las cuatro gráficas que resultan de este procedimiento, es muy importante la gráfica `Residuals vs Fitted`, es decir, **gráfica de residuos frente a valores pronosticados**. Se trata de una representación gráfica donde el eje de abscisas incluye los valores ajustados de la recta de regresión y el eje de ordenadas los residuos de dichos ajustes. La forma de analizar el cumplimiento de las hipótesis previas a partir de dicho gráfico sería:

  - MEDIA CERO: cuando la curva de medias coincida con el eje de abscisas, que es el valor cero de los residuos.
  - HOMOCEDASTICIDAD: esta hipótesis se estaría violando (heterocedasticidad) si al analizar el gráfico detectamos que el tamaño de los residuos aumenta o disminuye de forma sistemática para algunos valores pronosticados.
  - LINEALIDAD: cuando no se observe ningún patrón extraño no lineal.

Además de esto, los residuos tienen que seguir distribuciones normales, para poder calcular intervalos de confianza de los coeficientes de regresión o de las predicciones de la variable explicada. Cuando el tamaño muestral sea pequeño, la comprobación de esta condición será imprescindible para poder hacer los cálculos inferenciales. El análisis de la hipótesis de normalidad se hará mediante el gráfico `Normal Q-Q`:

- NORMALIDAD: si la distribución es normal, los puntos del gráfico `Normal Q-Q` están sobre recta; cuanto más se alejen de ella, menos creíble es esta hipótesis.

El gráfico `Residuals vs Leverage` permite visualizar la existencia de observaciones atípicas, las cuales pueden ser muy influyentes en la estimación de los parámetros.

Vamos a ver el estudio de la adecuación del modelo a través del ejemplo que estamos considerando en toda esta práctica.

Ejemplo 6.4. *Determine la adecuación del modelo de regresión lineal simple obtenido en el Ejemplo 6.3.*

Solución:

- Hemos obtenido que el coeficiente de determinación es: $R^2 = 72'93\,\%$, que estima el porcentaje de la variación de la variable dependiente que es explicado por la regresión. En este caso, el $72'93\,\%$ de la variación de la emisión de `N2O` puede ser explicada por la emisión de `CO2`.

  Asociado a este valor, tenemos el del coeficiente de determinación ajustado, que es, en este caso, $R^2_a = 0'7269$. Puesto que no es muy pequeño, el modelo puede considerarse adecuado, de momento.

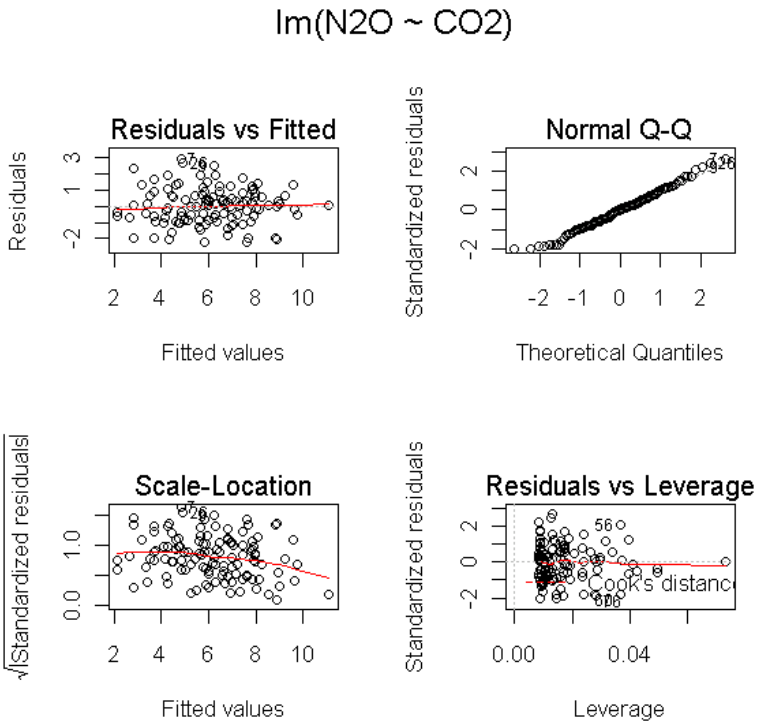
- Los p-valores asociados a los contrastes para los dos coeficientes ($2'91 \cdot 10^{-7}$ e inferior a $2 \cdot 10^{-16}$, respectivamente) son significativos, con lo que el modelo no es totalmente inadmisible, aporta algo en la predicción de la variable explicada.

  El p-valor de la tabla ANOVA (última línea de la salida) coincide en el caso de la regresión lineal simple con el p-valor de la variable explicativa, así que no merece un análisis separado.

- Para obtener los gráficos necesarios para analizar los residuos de este modelo de regresión debemos seguir los siguientes pasos:

  Modelos
  ➥Gráficas
  ➥Gráficas básicas de diagnóstico

  Cuyos resultados son:



En la gráfica de arriba a la izquierda (`Residual vs Fitted`) se observa como la hipótesis de media cero para los residuos parece admisible (la curva de medias se parece mucho al eje de abscisas), no hay evidencias claras de heterocesdasticidad y la hipótesis de linealidad no es rechazable.

En el gráfico de arriba a la derecha (`Normal Q-Q`) los puntos están casi todos sobre la recta $y = x$, con lo que la hipótesis de normalidad también parece admisible.

El gráfico de abajo a la derecha (`Residuals vs Leverage`) no muestra la existencia de ningún dato atípico.

Todo lo anterior nos lleva a considerar que el modelo propuesto es adecuado para predecir o pronosticar la emisión de `N2O` en función de la emisión de `CO2`. □

Como ya comentamos, aquí nos hemos concentrado en un estudio preliminar y básico sobre el análisis de los residuos. Basarnos en representaciones gráficas no nos permiten concluir con seguridad la bondad del ajuste. En la práctica estos gráficos deben ser complementadas con contrastes adecuados, como por ejemplo el test de Breusch-Pagan para estudiar la homocedasticidad del modelo, el test Reset de no linealidad para estudiar la linealidad de los residuos o el test de valores atípicos de Bonferroni para detectar la presencia de observaciones atípicas. Todos ellos se encuentran dentro de la opción del menú *Modelos* → *Diagnósticos numéricos*. También puede estudiarse la normalidad de los residuos mediante el test de normalidad de Shapiro-Wilk. No obstante la aplicación de todos estos tests se escapa de los objetivos de este curso y nosotros nos conformaremos con un diagnóstico gráfico.

6.4. Paso 4: Realización de pronósticos

Recordemos que los valores que proporciona la recta de regresión para un valor dado de la variable explicativa pueden interpretarse como predicciones del valor de la variable explicada o

como estimaciones de su media.

Tanto para estas predicciones como para estas estimaciones, podemos proporcionar intervalos de confianza al nivel que se considere apropiado, normalmente al 95 %.

Vamos a ver su funcionamiento a través de un ejemplo.

Ejemplo 6.5. *Utilice el modelo de regresión lineal simple obtenido en el Ejemplo 6.3 para estudiar los valores de emisión de* `N2O` *en las horas en las que se emiten* 110*t/h de* `CO2`*.*

Solución: Para hacer todo esto la forma más rápida es teclear en la `Ventana de instrucciones` el comando

```
predict(RegModel.1,data.frame(CO2=c(110)))
```

donde `RegModel.1` se refiere al nombre que le ha asignado al modelo la función `lm` (puede verse en la parte superior derecha de la pantalla) y a continuación especificamos el valor de la variable explicativa para el que queremos hacer las predicciones.

Las salidas de este procedimiento nos indican que la estimación puntual, obtenida a partir de este modelo de regresión lineal simple, de la emisión de `N2O` que se producirá en una hora en la que se hayan emitido 110t de `CO2` es de 6′350341.

Si además de la estimación puntual, queremos un intervalo de pronóstico, simplemente debemos completar la instrucción anterior como sigue:

predict(RegModel.1,data.frame(CO2=c(110)),interval='prediction')

Las salidas del procedimiento anterior son:

```
       fit      lwr      upr
1 6.350341 4.140992 8.559691
```

Lo que significa que se tiene una confianza del 95 % de que la emisión de `N2O` estará entre 4′140992t/h y 8′55969t/h para una hora en la que haya una emisión de 110t/h de `CO2`. El valor 6′350341 es la estimación puntual de la emisión de `N2O` que se obtiene a partir de este modelo, es decir, es el valor que se obtiene en la recta de regresión para `CO2`=110, como ya comentamos anteriormente.

Además se puede añadir una última opción que sea del tipo `level=0.99`, para modificar el nivel de confianza del intervalo, que por defecto es del 95 %, con lo que la instrucción quedaría: predict(RegModel.1,data.frame(CO2=c(110)),interval='prediction',level=0.99)

Con la opción `interval='prediction'` hemos especificado que queremos obtener un intervalo de pronóstico o predicción para un valor concreto. Si queremos obtener intervalos de confianza para el promedio estimado debemos reemplazar esta opción por `interval='confidence'`. Así, si tecleamos predict(RegModel.1,data.frame(CO2=c(110,100)),interval='confidence') obtenemos las salidas

```
       fit      lwr      upr
1 6.350341 6.145248 6.555434
2 5.911843 5.707193 6.116494
```

que nos permiten afirmar que tenemos una confianza del 95 % de que la emisión media de `N2O` aquellas horas en la que se emiten 110t/h de `CO2` está entre 6′145248 y 6′555434.

Aquí vemos además como se pueden hacer varias estimaciones a la vez, sin más que incluir la lista de valores en los que las queremos obtener.

□

6.5. Ejercicios propuestos

Ejercicio 6.1. *Responda razonadamente a las siguientes cuestiones, en función de los datos recogidos en el conjunto de datos* `acero.rda`.

a) *¿Existe relación lineal significativa entre el consumo y la emisión de* `SO2`*? ¿y entre el consumo y la emisión de* `N2O`*?*

b) *Si se quiere predecir el consumo a través de un modelo lineal simple, y las posibles variables explicativas son: la emisión de* `N2O`*, la emisión de* `SO2` *y la emisión de* `NOx`*, ¿cuál es la variable que debería considerarse?*

c) *Represente gráficamente el consumo frente a la variable seleccionada en el apartado anterior y comente dicho gráfico.*

d) *Ajuste un modelo de regresión lineal simple para explicar el consumo en función la variable elegida en el apartado b).*

e) *Calcule e interprete el coeficiente de determinación para dicho modelo.*

f) *¿Es significativo (significativamente distinto de cero) el coeficiente de regresión (coeficiente que multiplica a la variable explicativa)?*

g) *Analice la adecuación del modelo.*

h) *¿Cuánto se estima que aumenta el consumo por cada unidad de* `N2O` *emitida?*

i) *Realice una predicción del consumo energético en una hora en la que la emisión de* `N2O` *ha sido de 6 t/h, mediante un valor puntual y mediante un intervalo de predicción al 95 %.*

j) *Si la empresa fija la emisión por hora de* `N2O` *en 6 t/h, realice una estimación del consumo medio en esa empresa, mediante un intervalo de confianza al 95 %.*

Ejercicio 6.2. *Responda razonadamente a las siguientes cuestiones, en función de los datos recogidos en el conjunto de datos* `acero.rda`.

a) *Defina una nueva variable, a la que dé el nombre de* `Y`*, mediante la expresión*

```
Y=ProdTotal + 20 * ProdTotal^2
```

b) *Represente gráficamente las observaciones de la variable* `Y` *frente a la variable* `ProdTotal`.

c) *Ajuste un modelo de regresión lineal simple para explicar los valores de la variable* `Y` *en función de la variable* `ProdTotal`.

d) *Analice la adecuación del modelo.*

6.6. Solución de los ejercicios propuestos

Ejercicio 6.1. *a) Puesto que la matriz de correlaciones es*

```
 Pearson correlations:
        consumo    N2O     NOx     SO2
consumo  1.0000 0.8274  0.5384 -0.0076
N2O      0.8274 1.0000  0.5317  0.0071
NOx      0.5384 0.5317  1.0000 -0.1262
SO2     -0.0076 0.0071 -0.1262  1.0000
```

no se rechaza la normalidad de ninguna de las variables implicadas, puesto que los p-valores del test de normalidad de Shapiro-Wilk son:

```
data:  acero$consumo
W = 0.9884, p-value = 0.4207

data:  acero$N2O
W = 0.9922, p-value = 0.7518

data:  acero$NOx
W = 0.9797, p-value = 0.07302

data:  acero$SO2
W = 0.9862, p-value = 0.2773
```

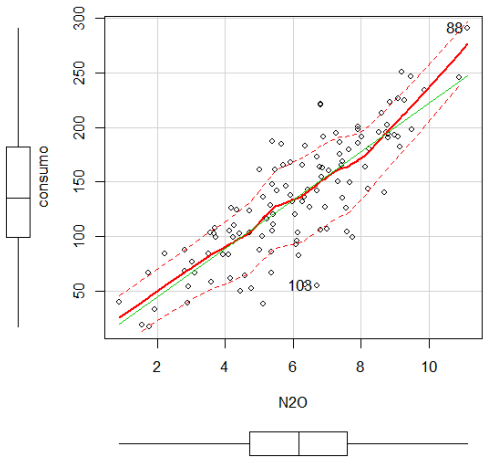
y los correspondientes p-valores sobre la nulidad del coeficiente de correlación son:

```
 Pairwise two-sided p-values:
        consumo N2O    NOx    SO2
consumo         <.0001 <.0001 0.9352
N2O     <.0001         <.0001 0.9398
NOx     <.0001  <.0001        0.1753
SO2     0.9352  0.9398 0.1753
```

se observa como no existe relación lineal significativa entre el consumo y la emisión de `SO2`*, pero sí entre el consumo y la emisión de* `N2O`*.*

b) El consumo tiene relación lineal con dos de las tres posibles variables explicativas (p-valor $= 0'000$*). Entre ellas, con la que la relación se estima que es más fuerte (estimación del coeficiente de correlación, en valor absoluto, mayor) es con la emisión de* `N2O`*.*

c) El diagrama de dispersión asociado es:



En él se puede observar que la consideración de un ajuste lineal parece bastante recomendable.

d) *El modelo de regresión lineal simple para explicar el consumo en función la emisión de* `N2O` *es:*

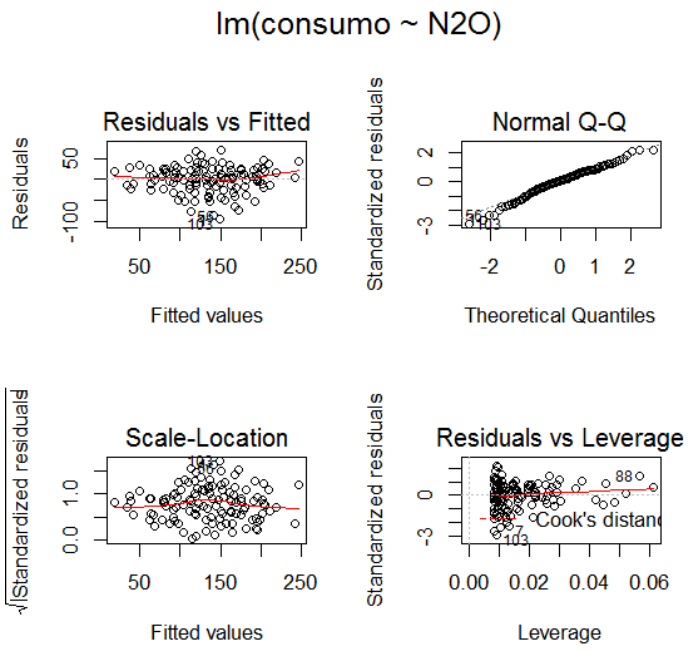
$$\texttt{consumo} = 0'2003 + 22'1556 \cdot \texttt{N2O}$$

e) *El coeficiente de determinación para dicho modelo es* $R^2 = 68'46\,\%$ *que puede interpretarse como que se estima que un* $68'46\,\%$ *de la variabilidad del consumo es explicada por la emisión de* `N2O`*. Dicho coeficiente no es demasiado cercano a cero, con lo que el ajuste de los datos a la recta de regresión no parece descartable en función de este valor.*

f) *Sí, el p-valor para el contraste* $H_0 : \beta_1 = 0$ *frente a la alternativa* $H_1 : \beta_1 \neq 0$ *es menor de* $2 \cdot 10^{-16}$*, con lo cual menor que cualquiera de los niveles de significación* α *habituales, lo que lleva a rechazar la hipótesis nula y concluir que el coeficiente de regresión es significativamente distinto de cero.*

g) *Según acabamos de ver en los dos puntos anteriores, simplemente nos quedaría analizar los residuos para determinar que el ajuste es bueno, puesto que tanto el coeficiente de determinación (*R^2 *era igual a* $0'6846$ *con lo que el coeficiente de determinación ajustado también es suficientemente alto; más en concreto se ha obtenido que* $R^2_a = 0'6819$*), como el contraste para el coeficiente de regresión no proporcionan indicios de que el modelo sea malo.*

Los gráficos asociados al análisis de residuos son:



En el gráfico `Residuals vs Fitted` *se observa que tanto la hipótesis de media nula, como la de linealidad y homocedasticidad (varianza constante) son admisibles. En el gráfico* `Normal Q-Q` *se observa además que la hipótesis de normalidad también es admisible.*

De lo anterior se deduce que el modelo de regresión lineal simple es adecuado para realizar con él pronósticos y estimaciones.

h) *Se estima que el consumo aumenta* 22′1556 *megavatios-hora por cada unidad extra de* `N2O` *emitida.*

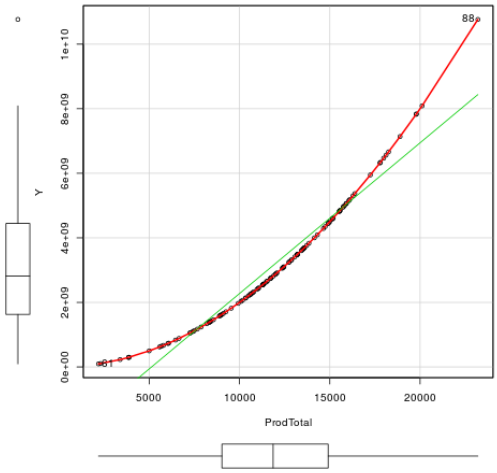
i) *Se estima que el consumo energético en una hora en la que la emisión de* `N2O` *ha sido de 6 t/h es de* 133′1339 *toneladas, lo que lleva además a afirmar que se tiene una confianza del* 95 % *de que el verdadero consumo esa hora está entre* 69′28326 *y* 196′9846 *toneladas.*

j) *Si la empresa fija la emisión por hora de* `N2O` *en 6 t/h, se tiene una confianza del* 95 % *de que el consumo medio en esa empresa está entre* 127′2474 *y* 139′0204 *toneladas.*

Ejercicio 6.2. *a*) *Para ello debemos ir a la opción del menú:* Datos → Modificar variables del conjunto de datos activo → Calcular una nueva variable.... *Una vez en la ventana* `Calcular una nueva variable`, *la rellenamos como sigue:*

- Nombre de la nueva variable*:* `Y`
- Expresión a calcular*:* `ProdTotal + 20 * ProdTotal^2`

b) *Representamos gráficamente la relación entre la variable* `Y` *y la variable* `ProdTotal` *mediante un diagrama de dispersión:*

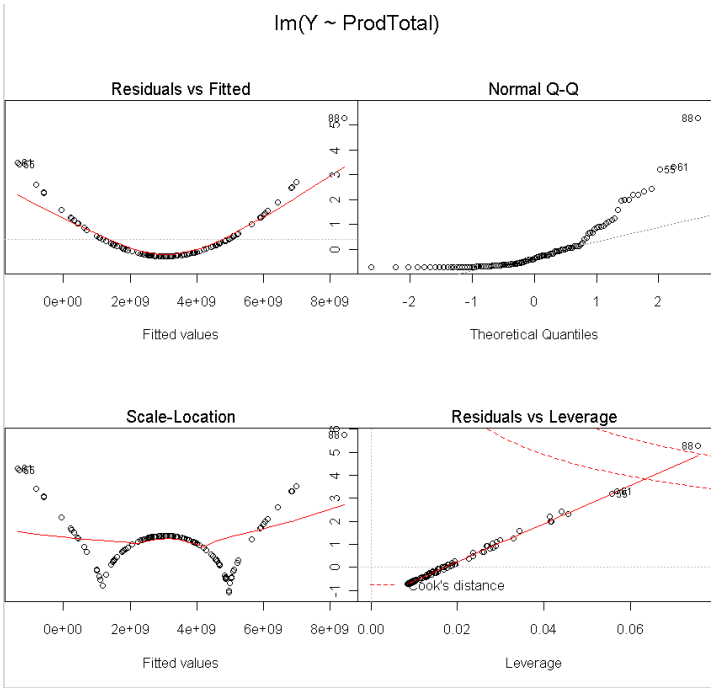


c) *El modelo es:* $\texttt{Y} = -2'394 \cdot 10^9 + 4'669 \cdot 10^5 \cdot \texttt{ProdTotal}$.

d) `Adjusted R-squared: 0.9452` *(alto).*

P-valor para el test sobre el coeficiente de regresión: `<2e-16 ***`*, con lo cual el coeficiente es significativamente distinto de cero.*

Gráficos de residuos:



En el gráfico `Normal Q-Q` *ya se observa una clara violación de la hipótesis de normalidad y la sugerencia del uso de un modelo cuadrático puede deducirse del gráfico* `Residuals vs Fitted`*. Realmente las dos comprobaciones anteriores (p-valor y coeficiente de determinación) no invalidaban el modelo y no es hasta este punto en el que se ve como el ajuste lineal no es el más adecuado y se sugieren otros mejores.*

Apéndice A

Esquema sobre los principales contrastes

- Bondad de ajuste.
  - Test de normalidad de Shapiro-Wilk, página 38
    `Estadísticos` → `Resúmenes` → `Test de normalidad de Shapiro-Wilk...`
- Promedio de una población.
  - Test t para una muestra (pobl. normal), página 40
    `Estadísticos` → `Medias` → `Test t para una muestra...`
  - Test de Wilcoxon para una muestra (pobl. NO normal), página 42
    `Estadísticos` → `Test no paramétricos` → `Test de Wilcoxon para una muestra...`
- Proporción de una población, página 43.
  `Estadísticos` → `Proporciones` → `Test de proporciones para una muestra`
- Comparación de proporciones, página 54.
  `Estadísticos` → `Proporciones` → `Test de proporciones para una muestra`
- Comparación de varianzas.
  - Test F (pobl. normal), página 55
    `Estadísticos` →`Varianzas` → `Test F para dos varianzas`
- Promedio de la diferencia.
  - Test t para muestras independientes (pobl. normal, muestras independientes), página 59
    `Estadísticos` → `Medias` → `Test t para muestras independientes ....`
    - Varianzas iguales
    - Varianzas distintas
  - Test t para datos relacionados (pobl. normal, datos apareados), página 61
    `Estadísticos` → `Medias` → `Test t para datos relacionados...`

  - Test de Wilcoxon para dos muestras (pobl. NO normal, muestras independientes), página 64

    `Estadísticos` → `Test no paramétricos` → `Test de Wilcoxon para dos muestras...`

  - Test de Wilcoxon para muestras pareadas (pobl. NO normal, datos apareados), página 64

    `Estadísticos` → `Test no paramétricos` → `Test de Wilcoxon para muestras pareadas...`

- Test de relación.

  - Test de independencia Chi-cuadrado, página 73

    `Estadísticos` → `Tablas de contingencia` → `Tabla de doble entrada...` → `Test de independencia Chi-cuadrado`

  - Test de correlación de Pearson, página 76

    `Estadísticos` → `Resúmenes` → `Test de correlación`

Apéndice B

Conjuntos de datos

B.1. Producción de acero

Con el fin de analizar el consumo energético de una empresa productora de acero se inspeccionaron varias horas de forma aleatoria, para las que se anotaron las variables más relevantes. En total se disponen de 117 mediciones (117 horas) recogidas en las siguientes variables:

`consumo` Consumo energético de la empresa por hora (megavatios).

`pr.tbc` Producción del tren de bandas calientes por hora (toneladas de acero).

`pr.cc` Producción de colada continua por hora (toneladas de acero).

`pr.ca` Producción del convertidor de acero por hora (toneladas de acero).

`pr.galv1` Producción de galvanizado de tipo I por hora (toneladas de acero).

`pr.galv2` Producción de galvanizado de tipo II por hora (toneladas de acero).

`pr.pint` Producción de chapa pintada por hora (toneladas de acero).

`linea` Línea de producción empleada: A o B.

`turno` Turno en la que se recogieron los datos: mañana (M), tarde (T) o noche (N).

`temperatura` Temperatura del sistema esa hora laborable: Alta, Media y Baja.

`pres.aver` Presencia de averías en esa hora: hubo averías (A) o no hubo averías (NoA).

`num.aver` Número de averías detectadas por hora.

`sistema` Activación de un sistema de detección de sobrecalentamiento en esa hora de trabajo: encendido (ON), apagado (OFF).

`ProdTotal` Producción total de la empresa por hora (toneladas de acero).

`NOx` Emisiones de mezcla de óxidos de nitrógeno por hora (toneladas).

`CO` Emisión de monóxido de carbono por hora (toneladas).

`COV` Emisión de compuestos orgánicos volátiles por hora (toneladas).

`SO2` Emisión de dióxido de azufre por hora (toneladas).

`CO2` Emisión de dióxido de carbono por hora (toneladas).

`N2O` Emisión de óxido nitroso por hora (toneladas).

Una muestra de la base de datos `acero` sería:

	consumo	pr.tbc	pr.cc	pr.ca	pr.galv1	pr.galv2	pr.pint	linea	turno	temperatura
1	55.31	3155	830	225	579	1401	120	A	N	Alta
2	84.08	443	903	58	611	1636	717	A	M	Alta
3	131.62	7270	572	36	982	1963	243	A	M	Baja
4	102.46	4931	694	122	96	1568	100	A	T	Baja
5	120.04	9365	1054	157	403	1300	180	B	M	Baja
6	103.68	9281	1003	172	605	1525	473	A	N	Baja

	pres.aver	num.aver	sistema	ProdTotal	NOx	CO	COV	SO2	CO2	N2O
1	A	1	OFF	11266	0.4900	3.5450	0.5450	0.038	101.5000	6.35
2	NoA	0	OFF	7251	0.0725	2.8950	0.4250	0.047	63.5650	2.23
3	NoA	0	OFF	11066	1.4900	5.0075	0.6900	0.062	98.8175	5.99
4	NoA	0	ON	8311	1.7150	2.1600	0.3600	0.066	70.1825	3.66
5	NoA	0	OFF	12459	0.4650	4.8450	0.6625	0.086	88.5300	6.06
6	A	1	OFF	13059	2.4175	3.6725	0.5750	0.056	116.5375	6.15

B.2. Datos sociales de alumnos de 1º de la EPI

Se les ha pasado el siguiente cuestionario a alumnos de 1º de Grado en la Escuela Politécnica de Ingeniería de Gijón:

1. Indica tu sexo:
   - Mujer
   - Hombre
2. Indica tu edad
3. Indica el número de hermanos que tienes (excluyéndote a ti mismo)
4. ¿Tienes carné de conducir?
   - Sí
   - No
5. ¿Utilizas habitualmente el transporte público para venir al Campus?

- Sí
- No

6. ¿Cuál es la distancia entre el Campus y la vivienda en que resides durante el curso?
7. Indica cuál es el tiempo, en minutos, que tardas en llegar al Campus desde la vivienda en la que resides durante el curso
8. Indica cuántas horas pasas en el Campus a la semana
9. Considera una semana intermedia del semestre, ¿cuántas horas semanales (máximo dos decimales) dedicadas, en conjunto, a participar en redes sociales (facebook, twitter, tuenti, etc.) o programas de mensajería (MSN, Yahoo Messenger, etc.)?
10. Considera una semana intermedia del semestre, ¿cuántas horas semanales (máx. dos decimales) dedicas, en conjunto, a ver la televisión o jugar con consolas de juegos, video-juegos, juegos por móvil, etc.?
11. Considera una semana intermedia del semestre, ¿cuántas horas dedicas al estudio (excluyendo las horas de clase presencial) de lunes a viernes?
12. Considera una semana intermedia del semestre, ¿cuántas horas dedicas al estudio en el fin de semana?
13. ¿Cuál ha sido la duración, en minutos, de la última llamada recibida en tu teléfono móvil?
14. ¿Cuál ha sido la duración, en minutos, de la última llamada enviada desde tu teléfono móvil?
15. Clasifícate respecto al uso que haces del tabaco:
    - No fumador/a (`No`)
    - Fumador/a ocasional o social (`Ocasional`)
    - Fumador/a habitual (`Habitual`)
16. ¿Practicas deporte?
    - Nunca (`Nunca`)
    - Sí, ocasionalmente (no todas las semanas) (`Ocasional`)
    - Sí, una o dos veces por semana (`1o2`)
    - Sí, al menos tres veces por semana (`Mas3`)
17. Indica la vía por la que accediste al Grado que estas cursando actualmente
    - Bachillerato+PAU (`PAU`)
    - Ciclo Formativo de Grado Superior (`Ciclo FP`)
    - Titulado/a universitario (`Titulado`)
    - Otros (`Otros`)

18. Indica la nota de acceso a la Universidad (si fue la PAU 2010 indica calificación sólo de la fase general, para años anteriores, la calificación de la PAU, si fue un Ciclo Formativo, indica la calificación media del Ciclo, si fue la prueba de Mayores de 25, Mayores de 40 o Mayores de 45 indica la calificación de la prueba, si fue una titulación universitaria, indica la nota media de la titulación en escala decimal, etc.)

El fichero `VidaEstudiantes.rda` recoge los resultados obtenidos para dicho cuestionario. Las variables de dicho fichero son:

`Grupo` Grupo de clase al que pertenece.

`Sexo` Sexo del encuestado: `Hombre` o `Mujer`.

`Edad` Edad, en años, del encuestado.

`Hermanos` Número de hermanos, excluyendo al encuestado.

`Carne` Carné de conducir: `Si` o `No`.

`TP` Utilización habitual del transporte público para venir al Campus: `Si` o `No`.

`Distancia` Distancia entre el Campus y la vivienda en que reside durante el curso.

`TiempoLlegar` Tiempo, en minutos, que tarda en llegar al Campus desde la vivienda en la que reside durante el curso.

`TiempoCampus` Horas en el Campus a la semana.

`Redes` Horas semanales dedicadas, en conjunto, a participar en redes sociales (facebook, twitter, tuenti, etc.) o programas de mensajería (MSN, Yahoo Messenger, etc.) en una semana intermedia del semestre.

`TV` Horas semanales dedicadas, en conjunto, a ver la televisión o jugar con consolas de juegos, video-juegos, juegos por móvil, etc. en una semana intermedia del semestre.

`EstudioLaV` Horas dedicas al estudio (excluyendo las horas de clase presencial) de lunes a viernes en una semana intermedia del semestre.

`EstudioFinde` Horas dedicas al estudio el fin de semana de una semana intermedia del semestre.

`LlamadaRecibida` Duración, en minutos, de la última llamada recibida en su teléfono móvil.

`LlamadaEmitida` Duración, en minutos, de la última llamada enviada desde su teléfono móvil.

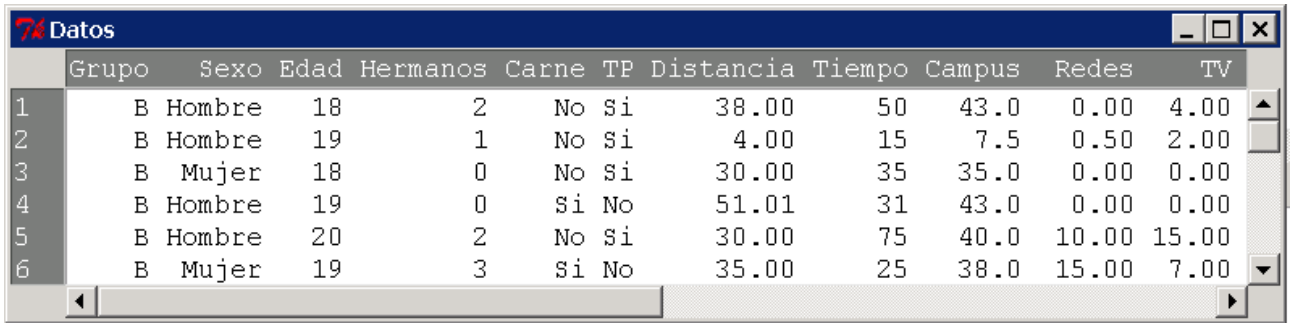
`Fumar` Uso del tabaco: No fumador/a (`No`), Fumador/a ocasional o social (`Ocasional`) y Fumador/a habitual (`Habitual`).

`Deporte` Práctica del deporte: Nunca (`Nunca`), Ocasionalmente (`Ocasional`), Una o dos veces por semana (`1o2`) y Al menos tres veces por semana (`Mas3`).

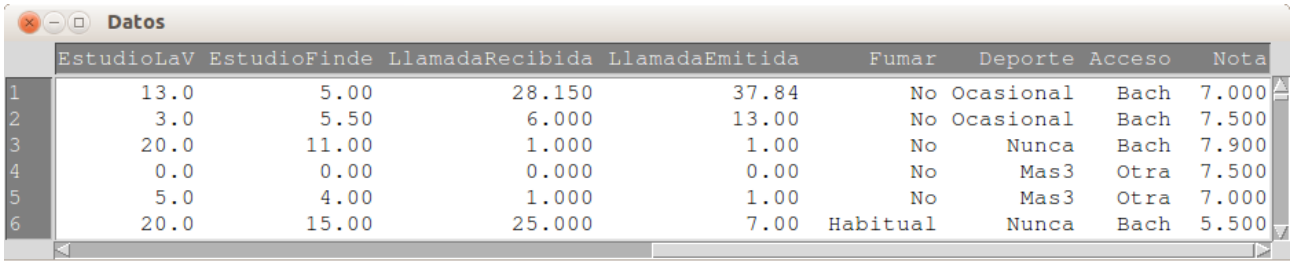
`Acceso` Vía de acceso al Grado: Bachillerato+PAU (`Bach`), Ciclo Formativo de Grado Superior (`Ciclo`), Mayores de 25 años (`Mas25`), Mayores de 40 años con experiencia laboral (`Mas40`), Mayores de 45 años (`Mas45`), Titulado/a universitario (`Titulo`) y Otra (`Otra`).

`Nota` Nota de acceso a la Universidad.

Una muestra de dicho fichero es:



	Grupo	Sexo	Edad	Hermanos	Carne	TP	Distancia	Tiempo	Campus	Redes	TV
1	B	Hombre	18	2	No	Si	38.00	50	43.0	0.00	4.00
2	B	Hombre	19	1	No	Si	4.00	15	7.5	0.50	2.00
3	B	Mujer	18	0	No	Si	30.00	35	35.0	0.00	0.00
4	B	Hombre	19	0	Si	No	51.01	31	43.0	0.00	0.00
5	B	Hombre	20	2	No	Si	30.00	75	40.0	10.00	15.00
6	B	Mujer	19	3	Si	No	35.00	25	38.0	15.00	7.00



	EstudioLaV	EstudioFinde	LlamadaRecibida	LlamadaEmitida	Fumar	Deporte	Acceso	Nota
1	13.0	5.00	28.150	37.84	No	Ocasional	Bach	7.000
2	3.0	5.50	6.000	13.00	No	Ocasional	Bach	7.500
3	20.0	11.00	1.000	1.00	No	Nunca	Bach	7.900
4	0.0	0.00	0.000	0.00	No	Mas3	Otra	7.500
5	5.0	4.00	1.000	1.00	No	Mas3	Otra	7.000
6	20.0	15.00	25.000	7.00	Habitual	Nunca	Bach	5.500