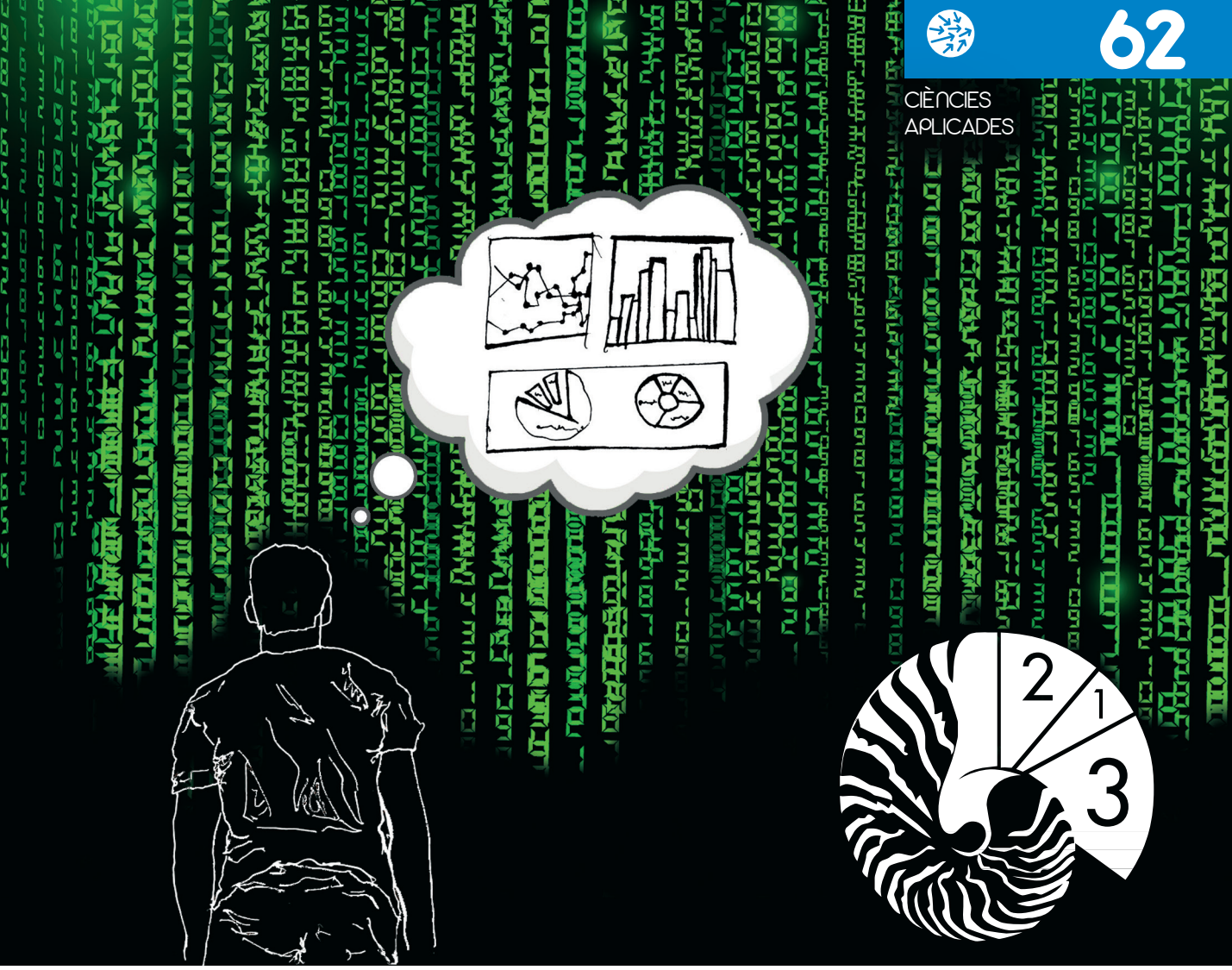
62

CIÈNCIES
APLICADES



50 ANYS
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
sense limits

UPCGRAU

Prácticas de estadística utilizando R
Aplicaciones en problemas de ingeniería

Luis Eduardo Mújica Delgado
Magda L. Ruiz Ordoñez

62

50 ANYS UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH sense límits

iniciativa digital politècnica
Publicacions Acadèmiques UPC

UPCGRAU

Prácticas de estadística utilizando R
Aplicaciones en problemas de ingeniería

Luis Eduardo Mújica Delgado
Magda L. Ruiz Ordoñez

Primera edición: octubre de 2021

Oficina de Publicacions Acadèmiques Digitals de la UPC
Jordi Girona 31,
Edifici Torre Girona, Plant 1, 08034 Barcelona
Tel.: 934 015 885
www.upc.edu/idp
E-mail: info.idp@upc.edu

DL: B 17069-2021
ISBN:978-84-9880-945-9

Índice

Introducción

No hace falta resaltar la importancia que tiene la estadística en las diferentes áreas profesionales (ingeniería, salud, educación, comercio, etc) donde se puede recolectar y agrupar información para construir informes que den idea y permita inferir desde el punto de vista cuantitativo y cualitativo de las poblaciones, procesos, sistemas, fallos, daños, estados de las condiciones normales y/o anormales, entre otros. Además, siendo la estadística parte de ciencia que envuelve teoría y práctica, resulta imperante aplicar en su enseñanza herramientas tecnológicas como elemento indispensable en la formación de nuestros futuros profesionales.

De igual manera, somos consientes de cómo la tecnología está en continuo desarrollo, que no parará porque nuestra curiosidad no tiene límites y por tanto seguiremos investigando en busca de conocimiento. Por ello, es imprescindible que conozcamos y manejemos las herramientas punteras que existen y facilitan los cálculos. En este aspecto, existen distintos paquetes de software estadístico como el SPSS, Minitab, S-Plus, entre otros. Estos softwares están muy bien desarrollados y son muy buenos satisfaciendo las necesidades de cualquier usuario. Sin embargo, es necesario el pago de una licencia para su uso. En contraste, ha surgido con fuerza una potente herramienta OPEN SOURCE llamada ***R***. Su uso en general es sencillo y permite diferentes modos de trabajo ya sea que el usuario entienda y sepa generar líneas de instrucciones o también generar líneas de código por medio de menús. De esta manera, no se necesita ser un experto estadístico ni un experto programador para entender los procedimientos necesarios para obtener una solución, sino que, a la vez, el usuario puede enfocarse en el análisis de los resultados.

Al ser código abierto, ***R*** está en permanentemente actualización y la comunidad estadística mejora su alcance y reconfigura sus fallos. Como ventajas adicionales de ***R*** se puede enumerar: 1. Disponibilidad para Linux, Windows y Mac. 2. Confiabilidad, robustez y estabilidad. 3. Calidad y facilidad de creación de gráficas. 4. Existe la opción de trabajarlo en la red sin necesidad de instalarlo. Facilitando su uso y acceso.

De esta manera, nace la principal motivación al diseñar estas guías: Ayudar a comprender y entender a nuestros estudiantes o aprendices conceptos estadísticos al tiempo que son aplicados y resueltos utilizando ***R***. Como dato interesante y sin que nos lo propusiéra-

mos al diseñar las guías, nos cruzamos con tres paradigmas propios de la enseñanza de temas científicos: "Paradigma de la enseñanza por transmisión", establece los trabajos prácticos como actividades de descubrimiento de hechos y conceptos, en este caso estadísticos, mediante la utilización de ***R***. Irremediablemente nos conduce al siguiente paradigma, "descubrimiento guiado y del descubrimiento autónomo", se basa en el concepto: los trabajos prácticos son actividades encaminadas a aprender por medio de la observación, clasificación, emisión de hipótesis, realización, etc. Finalmente, el "paradigma de la ciencia de los procesos", donde los trabajos prácticos se utilizan para la adquisición de habilidades y para poner a los estudiantes en situación de resolver problemas prácticos. Como resultado, el objetivo al diseñar estas guías es el de encaminar el trabajo de los aprendices de ***R***, proporcionando la información básica y necesaria para el entendimiento del tema. De manera que observarán como sus conocimiento y capacidades de resolución van en evolución al resolver los retos propuestos (ejercicios propuestos) en cada una de ellas.

Hemos diseñado las guías de manera atractiva, comenzando con una pregunta que en ese momento no se podría resolver, pero al final de ella, obtendremos todas las respuestas. Los temas se van presentando de manera secuencial, comenzando por el uso básico como cualquier otro software, pasando por la simulación de experimentos aleatorios, hasta finalmente llegar a facilitar el análisis a temas mucho más especializados como lo es el contraste de hipótesis.

En todo momento, no esperamos que estas guías sean sustitución del profesor o profesora. Es primordial que, en el aprendizaje de temas científicos, todas las situaciones sean guiadas por un experto o experta para favorecer el desarrollo y familiarización de la tecnología en los aprendices. Nuestro deseo es que estas guías favorezcan el aprendizaje activo promoviendo el contacto entre unos y otros con la realimentación del conocimiento como etapa esencial en la asimilación de los conceptos. Como resultado, al final, nuestros estudiantes o aprendices encontrarán sentido a los conceptos explicados basándose en la experiencia, aplicación, análisis y finalmente inferencia de los retos.


2
1
3

Introducción al R

1.1. Introducción y objetivos

En esta sesión, se hace una introducción a la herramienta informática para el análisis estadístico *R*. *R* es un lenguaje de programación en código abierto (*free software*) y gratuito (*freeware*) que últimamente está suscitando el interés de la academia, de la investigación e incluso de la industria.

Primero, se explica brevemente qué es el *R* y sus diversos entornos de uso. Además, se ofrecen las indicaciones para descargar e instalar el programa, sus paquetes y el entorno de trabajo. A continuación, se detallan algunas funciones básicas, así como el procedimiento para crear y manipular tablas de datos (data.frame). También se describe cómo se guardan y recuperan los comandos ejecutados en una sesión y el espacio de trabajo. Finalmente, se muestra cómo se crea y guarda un script o guion para poder abrirlo y ejecutarlo desde otro ordenador o en otra sesión. Al finalizar esta sesión, el estudiante ha de ser capaz de:

- Descargar e instalar los ficheros de *R*, el paquete *R-Commander* y el entorno de trabajo *Rstudio*.
- Identificar las principales características de la *consola de R*, el paquete *R-Commander* y el entorno de trabajo *Rstudio*.
- Iniciar una sesión de trabajo en la *consola de R*, en el *R-Commander* y en *Rstudio*.
- Crear y manipular una tabla de datos.
- Guardar y recuperar el histórico de una sesión trabajada y su espacio de trabajo.
- Crear, guardar y recuperar un script o guion con una secuencia de comandos.

1.2. R, R-Commander y Rstudio

1.2.1. ¿Qué son R, R-Commander y Rstudio?

Aunque, para algunas personas, *R* es un software, este se puede considerar un lenguaje de programación enfocado al análisis estadístico de datos y su representación gráfica.

Puede ejecutarse en cualquier ordenador y tiene un soporte *online* muy amigable y activo (https://www.r-project.org). Proporciona una gran cantidad de herramientas con la capacidad de llamar a otras funciones y de desarrollar nuevas funciones muy sencillas de manejar Además, su gran capacidad de visualización de los datos permite generar gráficos muy variados y de extraordinaria calidad y flexibilidad. Permite su integración con diferentes bases de datos y con otros lenguajes de programación, como Matlab, Maple, Mathematica, Python, Perl, SPSS, etc. Además, como es un proyecto abierto y colaborativo, existe un repositorio oficial de paquetes (https://cran.r-project.org/web/packages/).

R permite trabajar con una ventana de interacción con usuario, *R-Console*. En su entorno básico, *R* no tiene una interfaz tipo ventana. Para obtener los resultados deseados, sus funciones se ejecutan por medio de comandos en su propio lenguaje. Sin embargo, *R* dispone de un módulo adicional (o paquete) llamado *R-Commander*, que proporciona una serie de menús que facilitan el uso inicial del programa, sin tener que escribir los comandos, es decir, con el uso del ratón.

R-Commander es una interfaz gráfica de usuario básica (*graphical user interface*, GUI). Sus menús permiten ejecutar muchas de las funciones básicas para el análisis estadístico de datos (pero no todas) y crear gráficas sin escribir los comandos; es más, genera el código en lenguaje *R* para que después pueda ejecutarse luego desde la *R-Console*, si así se desea. Toda la información, los ficheros, ayudas y manuales se pueden consultar en su página web (http://www.rcommander.com/)

Por otra parte, existe un entorno para el desarrollo integrado (*integrated development environment*, IDE) llamado *Rstudio*, que es básicamente una agradable interfaz que incluye una consola, un editor más completo y funcional, una ventana de gráficos y la visualización de las variables en el espacio de trabajo, entre otras cosas. Está completamente integrado al *R* y al *R-Commander* y permite ejecutar el código directamente del editor, gestionar múltiples directorios y ficheros.

1.2.2. Instalación

Como *R* es gratuito, en internet se pueden encontrar muchos sitios para descargar los ficheros necesarios para su instalación. Sin embargo, el sitio oficial de *R* tiene a disposición la última versión para Windows, Linux y Mac (OS X). CRAN (Comprehensive R Archive Network) es una red de servidores web y FTP por todo el mundo que almacena las versiones más actualizadas de código y documentación de *R* (https://cran.r-project.org/).

La instalación suele ser un tanto complicada, dependiendo del sistema operativo y naturalmente de su versión. Como *R* está constantemente en desarrollo, es difícil definir en este documento los pasos exactos y definitivos para su instalación. Sin embargo, las instrucciones sencillas que se ofrecen a continuación son la base para una instalación correcta en Windows. Si se quiere instalar en otra plataforma o surge algún inconveniente, lo mejor es consultar los foros y las preguntas frecuentes del CRAN.

La versión más actualizada de *R* se puede descargar desde (https://cran.r-project.org/) siguiendo la ruta: Download R for Windows >base >Download R 3.x.x for Windows. Se

ejecuta el fichero dejando todas las opciones de instalación por defecto. En la figura 1.1, se muestra el icono de acceso directo que aparece después de la instalación.

Fig. 1.1
Icono de acceso a R

Rstudio se descarga desde (http://rstudio.org/download/desktop). Se ejecuta el fichero de instalación seleccionando de nuevo todas las opciones de configuración por defecto. El icono de acceso directo se muestra en la figura 3.1.

Fig. 1.2
Icono de acceso a Rstudio

La instalación de *R-Commander* se puede realizar desde *R* o desde *Rstudio*. Si se quiere hacer desde *R*, se ejecuta el programa y en la consola se ejecuta la sentencia:

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

Se selecciona el servidor deseado e inmediatamente comienza la descarga y la instalación de todas las librerías necesarias para su ejecución.

Ahora, si se quiere instalar *R-Commander* desde *Rstudio*, se ejecuta el programa y se selecciona la pestaña Packages. Se hace clic en Install Packages (figura 3.2).

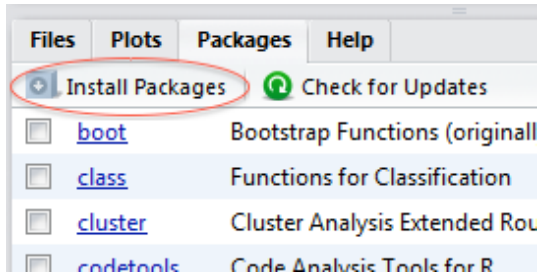


Fig. 1.3
Instalación de un paquete (o librería) desde Rstudio

En la nueva ventana, se comienza a escribir Rcmdr en el espacio de Packages asegurándose que la opción Install dependencies esté seleccionada. Finalmente, se pulsa Install, tal como se observa en la figura 3.3.

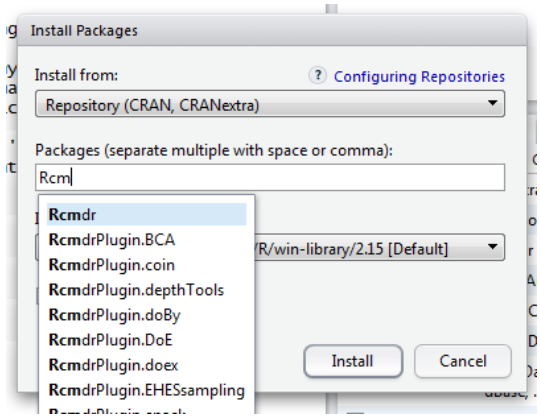


Fig. 1.4
Instalación de R-Commander desde Rstudio

Como *R-Commander* es un paquete de *R*, no se puede ejecutar de forma independiente y, por tanto, no genera ningún icono de acceso directo. Para ejecutar el *R-Commander* (desde *R* o *Rstudio*), hay que ejecutar desde la consola la siguiente instrucción:

```
[]library(Rcmdr)
```

1.2.3. Primeras impresiones

R-Console

Una vez iniciado el programa, se puede observar que se abre una ventana de trabajo denominada *Consola R*, tal como se muestra en la siguiente figura 3.5.

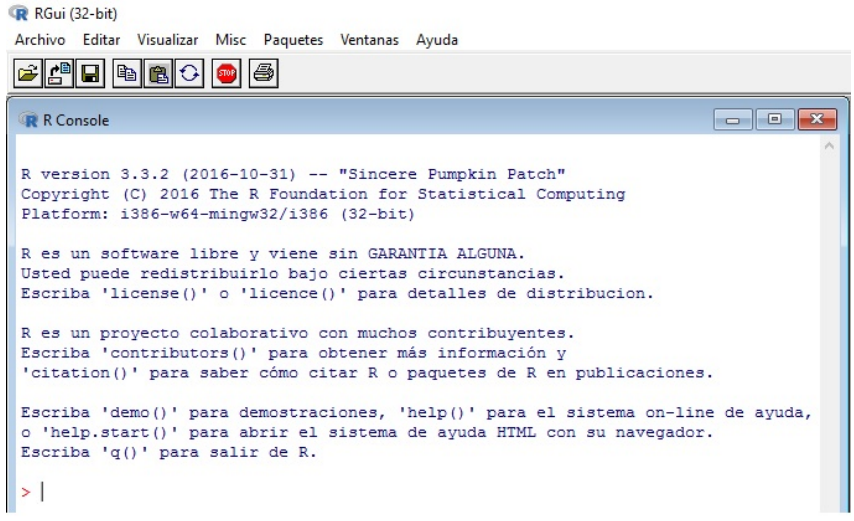


Fig. 1.5
Vista de la consola de R.

El cursor >| indica que el programa está preparado para aceptar órdenes y efectuar los cálculos correspondientes. Dichas órdenes deben darse en forma de comandos, operadores y funciones. Los más importantes se irán introduciendo progresivamente a lo largo de las distintas sesiones de prácticas. Adicionalmente, contiene una barra de menú principal con diferentes opciones, tales como las típicas de cualquier programa en el entorno Windows y la configuración de paquetes y ventanas.

- **Archivo**: Para efectuar operaciones básicas con los ficheros (*scripts*, área de trabajo, histórico)
- **Editar**: Se trata del típico menú de edición (copiar, pegar, etc.). También se usa para limpiar la consola y editar los datos.
- **Visualizar**: Para visualizar u ocultar la barra de herramientas y la barra de estado.
- **Misc**: Para configurar opciones avanzadas.
- **Paquetes**: Gestiona los distintos paquetes que se pueden cargar en *R*.
- **Ventanas**: Para configurar las ventanas.
- **Ayuda**: Facilita información acerca del programa *R*.

Rstudio

Rstudio es un entorno libre y de código abierto para el desarrollo integrado (IDE) de *R*. Se puede ejecutar en el escritorio o incluso a través de internet, mediante el servidor *Rstudio*. Este programa aúna todos los entornos y asume la filosofía de las expresiones, pero aportando algunas 'ayudas' que hacen más llevadero el día a día. Está organizado en cuatro zonas de trabajo distintas, como se aprecia en la figura 3.6.

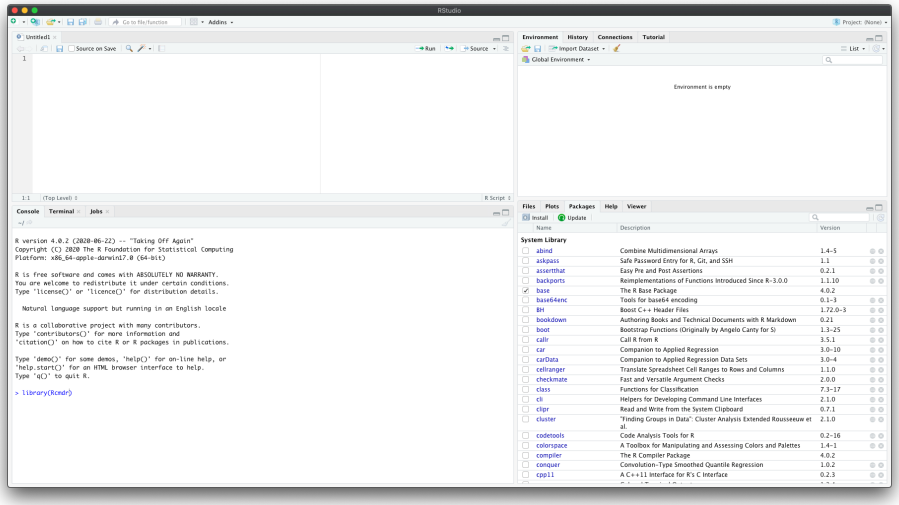


Fig. 1.6
Vista de Rstudio

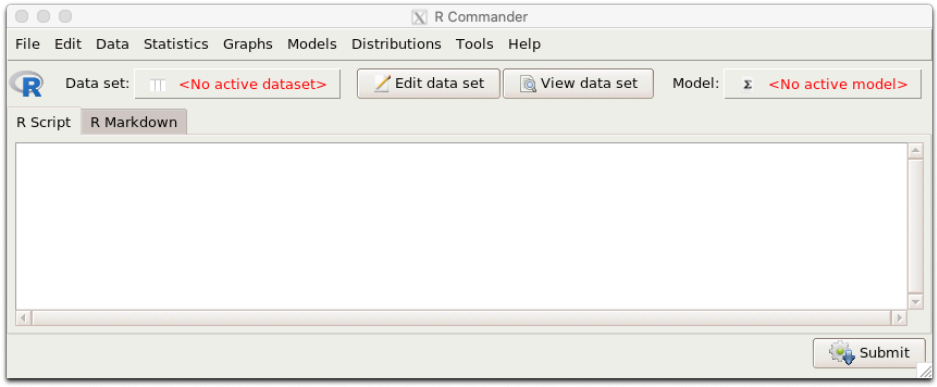
En la parte superior izquierda, pueden abrirse y editarse uno o varios ficheros con código *R* (*scripts*) a la vez. En la parte inferior izquierda, hay una consola de *R* en la cual pueden ejecutarse comandos de *R* individualmente. La parte superior derecha tiene cuatro pestañas, entre ellas: Workspace, donde aparece la lista de los objetos creados en la memoria; History, que contiene el histórico de las líneas de código ejecutadas; Connections, donde se puede realizar una conexión a fuentes de datos existentes, y finalmente Tutorial, donde se puede obtener información adicional sobre los paquetes desarrollados en *R*. La parte inferior derecha dispone de cuatro pestañas: Files, que da acceso al árbol de directorios y ficheros del disco duro; Plots, donde aparecen los gráficos creados en la consola; Packages, que facilita la administración de los paquetes de *R* instalados en la máquina, y Help, donde se abren las páginas de ayuda.

Desde la barra del menú principal, se puede acceder a todos los menús de *Rstudio*. Los menús: Archivo, Edición, Ver y Ayuda son habituales en los programas bajo el entorno de Windows. El resto de menús son específicos de *Rstudio*, estos permiten gestionar la interfaz, es decir, editar los ficheros, importar datos, instalar paquetes, gestionar las gráficas, etc,. Pero en ningún momento permite ningún cálculo estadístico o representación gráfica, todo esto lo debemos hacer por medio de comandos, tal como en la *R-console*.

R-Commander

R-Commander es una interfaz gráfica de usuario (GUI), creada por John Fox, que permite acceder a muchas capacidades del entorno estadístico *R* sin que el usuario tenga que conocer el lenguaje de comandos propio de este entorno. Para su utilización, se debe abrir *R* o *Rstudio* y ejecutarlo desde la consola (*R-Commander* no es una aplicación que funcione sola). La ventana que aparece es la siguiente (figura 3.7):

Fig. 1.7
Vista de la ventana de R-commander



Cada vez que, a través de los menús, se accede a las capacidades de *R* (gráficos, procedimientos estadísticos, modelos, etc.), en la ventana de instrucciones (R Script) se muestra la instrucción o conjunto de instrucciones que ejecutan la tarea que se ha solicitado, y en la consola (ya sea de *R* o de *Rstudio*) se muestra el resultado de dicha instrucción. De este modo, aunque el usuario no conozca el lenguaje de comandos de *R*, simplemente observando lo que va apareciendo en la ventana de instrucciones se irá familiarizando con dicho lenguaje.

Adicionalmente, el usuario puede introducir comandos directamente en esta ventana y tras clicar el botón Ejecutar, dichos comandos se ejecutarán y su resultado se visualizará igualmente. Las instrucciones pueden guardarse y volver a ser ejecutadas directamente con otros conjuntos de datos diferentes.

El acceso a las funciones implementadas en *R-Commander* es muy simple y se realiza utilizando el ratón para seleccionar, dentro de la barra de menú principal situada en la primera línea de la ventana, la opción a la cual queramos acceder. Allí se puede encontrar:

- **Fichero**: Para abrir ficheros con instrucciones a ejecutar, o para guardar datos, resultados, sintaxis, etc.
- **Editar**: Contiene las típicas opciones para cortar, pegar, borrar, etc.
- **Datos**: Utilidades para la gestión de datos (creación de datos, importación desde otros programas, recodificación de variables, etc.).
- **Estadísticos**: Para ejecutar procedimientos propiamente estadísticos.
- **Gráficas**: Contiene todos los gráficos disponibles.
- **Modelos**: Permite la definición y el uso de modelos específicos para el análisis de datos.
- **Distribuciones**: Ofrece probabilidades, cuantiles y gráficos de las distribuciones de probabilidad más habituales (Normal, t de Student, F de Fisher, binomial, etc.).
- **Herramientas**: Permite cargar librerías y definir el entorno.
- **Ayuda**: Ayuda sobre *R-Commander* (en inglés).

  Adicionalmente, existe una barra de herramientas bajo la barra de menú con los siguientes botones:

- **Conjunto de datos**: Muestra el nombre de la serie de datos activa. Inicialmente, no hay ninguna serie de datos activa. Al pulsar este botón, se puede elegir entre las series de datos que están actualmente en la memoria (si hay más de una).
- **Editar conjunto de datos**: Permite abrir el editor de datos de R para modificar la serie de datos activa.
- **Visualizar**: Permite abrir el editor de datos de R para examinar la serie de datos activa.
- **Modelo**: Indica el nombre del modelo estadístico activo, un modelo lineal (como el modelo de regresión lineal), un modelo lineal generalizado, etc. Inicialmente, no hay ningún modelo activo.

1.3. Primeros pasos con R

Para emprezar, se describen los primeros pasos para introducir datos, realizar operaciones, calcular funciones y representar gráficamente los datos y el análisis estadístico utilizando el lenguaje *R*, ya sea en la consola de *R*, en *Rstudio* o en el *R-Commander* . Finalmente, se mostrará cómo se guarda una sesión.

1.3.1. R como calculadora básica

En *R*, se pueden realizar operaciones de cálculo numérico básicas tales como: suma (+), resta (-), multiplicación (*||*), división (/), división entera (%/ %), residuo (% %), potencia (^), etc. Además, están disponibles operaciones lógicas como: igual (==), mayor que (>), menor que (<), mayor o igual que (>=), menor o igual que (<=), diferente (!=), and (&), or (||), etc. Por ejemplo:

```
2+2
```

```
## [1] 4
```

```
2+3*5/6+4^2
```

```
## [1] 20.5
```

```
31%%7
```

```
## [1] 3
```

```
202%%10
```

```
## [1] 20
```

Los operadores <- o = se utilizan para hacer asignaciones. Es preferible el uso del primero, ya que el signo igual tiene en algunas ocasiones, connotaciones lógicas. La variable se crea en el mismo instante de la asignación. Es más, no se puede declarar con anterioridad y dejarla vacía. A continuación, se pueden ver algunos ejemplos de asignación de variables.

```
x <- 4
y = 6
z = x+y  NO muestra el resultado
z = x+y; z  SI muestra el resultado
```

```
## [1] 10
```

```
(z = x+y)  SI muestra el resultado
```

```
## [1] 10
```

Los nombres de las variables también pueden contener períodos, demarcado con el punto (.), por ejemplo:

```
x.inicial <- 4
x.final = 10
x.dif = x.final-x.inicial; x.dif
```

```
## [1] 6
```

Al comparar dos variables, el resultado es una variable lógica que indica si la declaración es verdadera o falsa. Por ejemplo:

```
x==y
```

```
## [1] FALSE
```

```
x!=y
```

```
## [1] TRUE
```

```
x>y
```

```
## [1] FALSE
```

R también tiene algunas constantes integradas: π pi o las letras del alfabeto en inglés en mayúsculas y minúsculas (LETTERS, letters), entre otras. Por ejemplo, se puede calcular el perímetro de la circunferencia de la Tierra en el Ecuador, sabiendo que su radio es 6378km

```
pi
```

```
## [1] 3.141593
```

```
2*pi*6378
```

```
## [1] 40074.16
```

Adicionalmente, existen muchas funciones matemáticas integradas en *R*, entre ellas podemos destacar: la raíz cuadrada (sqrt()), las funciones exponencial y logarítmicas (exp(), log(), log10()), las funciones trigonométricas (sin(), cos(), tan()), el valor absoluto (abs()), las funciones de redondeo (ceiling(), floor(), trunc(), round()), etc. A continuación, se pueden ver algunos ejemplos:

```
sin(45*pi/180)
```

```
## [1] 0.7071068
```

```
sqrt(81)
```

```
## [1] 9
```

```
exp(2)
```

```
## [1] 7.389056
```

```
log(20)
```

```
## [1] 2.995732
```

Tips & Tricks!

- Para ejecutar las instrucciones que están en una línea, se pulsa la tecla Enter.
- Todo lo que va precedido por almohadillas (#) *R* lo considera un comentario y no lo interpreta.
- Varias instrucciones se pueden ejecutar en una misma línea si se separan por un punto y coma (;).
- Para visualizar los datos asignados a una variable, se introduce el nombre de la variable.
- Se pueden recuperar líneas de instrucciones introducidas anteriormente pulsando la tecla con la flecha ascendente del teclado, a fin de volver a ejecutarlas o modificarlas.
- Para abortar la ejecución de una instrucción y devolver el control al usuario, basta pulsar la tecla (Esc) del teclado. Así recuperaremos el símbolo (>) para volver a escribir las instrucciones.

1.3.2. Vectores y matrices

El uso de vectores y matrices es fundamental para poder organizar los datos de forma adecuada para su análisis estadístico posterior. Por tanto, es necesario saber cómo definirlos, utilizarlos y manipularlos en *R*.

Definición de vectores

Para construir un vector, primero se define un nombre (por ejemplo, x), acto seguido se ingresa el operador de asignación y después se introduce la letra c (de concatenar). Finalmente, se escriben las componentes del vector entre paréntesis y separadas por comas.

```
x <- c(1,2,3,4,5); x
```

```
## [1] 1 2 3 4 5
```

También se pueden introducir los datos por el teclado con la instrucción scan(). Los valores se teclean dejando espacios en blanco, cada vez que se pulsa la tecla Enter se cambia de línea y se puede continuar introduciendo valores. Para terminar, se pulsa Enter en una línea vacía:

```
y = scan()
```

Una vez definido el vector, mediante la función length() se puede conocer su longitud (número de elementos que lo componen):

```
length(y)
```

```
## [1] 0
```

Si el vector es una secuencia de valores enteros (por ejemplo, de 1 a 10), se puede definir de la siguiente manera:

```
x1 <- 1:10; x1
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Un vector también se puede definir como una secuencia de valores equidistantes mediante la función seq(). Como información adicional, se han de definir el valor inicial del vector (from =), su valor final (to =) y la distancia entre valores (by =) o la longitud del vector (length =):

```
x2 <- seq(from=2, to=18, by=2); x2; length(x2)
```

```
## [1]  2  4  6  8 10 12 14 16 18
## [1] 9
```

```
x3 <- seq(from=2, to=18, length=30); x3; length(x3)
```

```
##  [1]  2.000000  2.551724  3.103448  3.655172  4.206897  4.758621  5.310345
##  [8]  5.862069  6.413793  6.965517  7.517241  8.068966  8.620690  9.172414
## [15]  9.724138 10.275862 10.827586 11.379310 11.931034 12.482759 13.034483
## [22] 13.586207 14.137931 14.689655 15.241379 15.793103 16.344828 16.896552
## [29] 17.448276 18.000000
```

```
## [1] 30
```

O simplemente:

```
x2 <- seq(2,18,2); x2; length(x2)
```

```
## [1]  2  4  6  8 10 12 14 16 18
```

```
## [1] 9
```

También se pueden definir como repeticiones de un valor o de un vector definido con anterioridad:

```
x4 <- rep(1,5); x4; length(x4)
```

```
## [1] 1 1 1 1 1
## [1] 5
```

```
rep(x, length=8)
```

```
## [1] 1 2 3 4 5 1 2 3
```

Incluso, se puede definir mediante una fusión de los comandos anteriores:

```
x5 <- c(1:4, 8:10, seq(-7,5,by=2), rep(x,length=8)); x5; length(x5)
```

```
## [1] 1 2 3 4 8 9 10 -7 -5 -3 -1 1 3 5 1 2 3 4 5 1 2 3
```

```
## [1] 22
```

Manipulación de vectores

Si se quiere acceder a un elemento específico de un vector, se introducen el nombre del vector y la posición del elemento entre corchetes. Para acceder a más de un elemento, primero hay que crear un vector con sus posiciones.

```
x5[10]  Décimo elemento del vector x5
```

```
## [1] -3
```

```
x5[c(10,15,1)]  Décimo, decimoquinto y primer elemento de x5
```

```
## [1] -3  1  1
```

Se pueden eliminar uno o varios elementos de un vector si se introduce la posición del elemento, precedido del signo menos (-) y entre corchetes:

```
x6 <- x5[-10]; x6; length(x6)
```

```
## [1] 1  2  3  4  8  9 10 -7 -5 -1  1  3  5  1  2  3  4  5  1  2  3
```

```
## [1] 21
```

```
x7 <- x5[c(-8,-7,-2,-17)]; x7;
```

```
##  [1]  1  3  4  8  9 -5 -3 -1  1  3  5  1  2  4  5  1  2  3
```

```
x7 <- x5[-c(8,7,2,17)]; x7;
```

```
##  [1]  1  3  4  8  9 -5 -3 -1  1  3  5  1  2  4  5  1  2  3
```

Se puede insertar un nuevo elemento en un vector mediante la creación de un nuevo vector utilizando los elementos del vector anterior:

```
x8 <- c(x6[1:4],20,x6[5:length(x6)]); x8; length(x8)
```

```
## [1] 1 2 3 4  20  8 9 10 -7 -5 -1  1  3  5  1  2  3  4  5  1  2  3
```

```
## [1] 22
```

Las operaciones y funciones vistas anteriormente para variables escalares pueden aplicarse a vectores, con la salvedad de que las operaciones se harán para cada componente del vector:

```
sin(c(0,30,45,60,90)*pi/180)  Seno de varios ángulos dados en grados
```

```
## [1] 0.0000000 0.5000000 0.7071068 0.8660254 1.0000000
```

```
exp(x6)  Exponencial de un vector definido previamente
```

```
##  [1] 2.718282e+00 7.389056e+00 2.008554e+01 5.459815e+01 2.980958e+03
##  [6] 8.103084e+03 2.202647e+04 9.118820e-04 6.737947e-03 3.678794e-01
## [11] 2.718282e+00 2.008554e+01 1.484132e+02 2.718282e+00 7.389056e+00
## [16] 2.008554e+01 5.459815e+01 1.484132e+02 2.718282e+00 7.389056e+00
## [21] 2.008554e+01
```

```
x5>1
```

```
##  [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
      FALSE FALSE
## [13] TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

Por otra parte, también se puede acceder a los elementos de un vector que cumplen una determinada condición. Por ejemplo, si se desea conocer el valor de los elementos de x5 que son menores que 0, se utiliza la función which() primero para saber la posición de los elementos que cumplen la condición:

```
ind <- which(x5<0); ind
```

```
## [1]  8  9 10 11
```

```
x5[ind]
```

```
## [1] -7 -5 -3 -1
```

O simplemente:

```
x5[x5<0]
```

```
## [1] -7 -5 -3 -1
```

Además de la función length(), algunas de las funciones básicas para utilizar vectores son las siguientes:

```
sum(x8)  Suma de los elementos
```

```
## [1] 74
```

```
range(x8)  Mínimo y máximo
```

```
## [1] -7 20
```

```
summary(x8)  Resumen estadístico
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -7.000   1.000   3.000   3.364   4.750  20.000
```

```
sort(x8)  Organiza los elementos de menor a mayor
```

```
## [1] -7 -5 -1 1 1 1 1  2  2  2  3  3  3  3  4  4  5  5  8  9 10 20
```

```
order(x8)  Muestra las posiciones al organizarlos de menor a mayor
```

```
## [1] 9 10 11 1 12 15 20 2 16 21 3 13 17 22 4 18 14 19 6  7  8  5
```

```
rev(x8)  Invierte los elementos del vector
```

```
## [1] 3 2 1 5 4 3 2 1 5 3 1 -1 -5 -7 10 9 8 20 4 3 2 1
```

Es muy frecuente en estadística tener un conjunto de variables que describen a una serie de individuos, y que de un individuo concreto (o varios) no se disponga del valor de una (o varias) de esas variables. *R* tiene en cuenta esta posibilidad; en estos casos, aparece el valor faltante como NA (*Non Available*, *No Accessible*) y algunas de las funciones básicas para tratar este tipo de datos son las siguientes:

```
v1 <- c(7,0,NA,8,5,6,NA,4);v1
```

```
## [1]  7  0 NA  8  5  6 NA  4
```

```
is.na(v1)  Para saber en dónde están esos NA
```

```
## [1] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
```

```
sum(v1)  Como faltan valores, la suma no se ejecuta
```

```
## [1] NA
```

```
sum(v1,na.rm=TRUE)  Realiza la operación sin considerar NA
```

```
## [1] 30
```

En los vectores, también se pueden almacenar cadenas de caracteres. La sintaxis es similar; la única diferencia es que cada cadena de caracteres ha de ir entre comillas dobles:

```
v <- c("Jesus","Jaime","Javier") ; v
```

```
## [1] "Jesus"  "Jaime"  "Javier"
```

Tips & Tricks!

- A las funciones en R se les pueden agregar atributos.
- El atributo (na.rm=TRUE) da la orden de que la función se ejecute sin tener en cuenta los datos *No Accesibles* (NA).
- na significa No Accesible, rm significa remove (quitar) y TRUE verdadero. Este último debe ir siempre en mayúsculas porque, de lo contrario, no lo reconoce.

Definición de matrices

Una matriz en *R* es un conjunto de objetos ordenados por filas y columnas. Un array en *R* es lo mismo, salvo que puede tener más de dos dimensiones. En general, una matriz se puede crear de dos formas: utilizando la función matrix o la función array. Manipular los datos que se encuentran dentro de una matriz es igual que con vectores, la diferencia es que ahora se ha de tener en cuenta la posición de cada elemento en función de las filas y las columnas.

```
m1 <- matrix(1:20,nrow=5); m1
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    6   11   16
## [2,]    2    7   12   17
## [3,]    3    8   13   18
## [4,]    4    9   14   19
## [5,]    5   10   15   20
```

```
m2 <- array(x5,dim=c(7,3)); m2
```

```
##      [,1] [,2] [,3]
## [1,]    1   -7    1
## [2,]    2   -5    2
## [3,]    3   -3    3
## [4,]    4   -1    4
## [5,]    8    1    5
## [6,]    9    3    1
## [7,]   10    5    2
```

```
m1[2,3]  Visualiza el valor del elemento de la fila 2 y la columna 3
de m1
```

```
## [1] 12
```

```
m1[2,c(1,3)]  Visualiza el valor de los elemento de la fila 2,
las columnas 2 y 3 de m1
```

```
## [1]  2 12
```

```
m1[c(1:5),2]  Visualiza el valor de todos los elementos de la columna
2 de m1
```

```
## [1]  6  7  8  9 10
```

También se pueden visualizar todos los elementos de una fila o de una columna de la siguiente forma:

```
m2[2,]  Visualiza el valor de todos los elemento de la fila 2 de m2
```

```
## [1]  2 -5  2
```

```
m2[,3]  Visualiza el valor de todos los elemento de la columna 3 de m2
```

```
## [1] 1 2 3 4 5 1 2
```

1.3.3. Estructuras de datos

La forma más común de almacenar datos es utilizar tablas (data.frames en *R*). Es como una matriz, formada por filas y columnas, con la diferencia de que cada columna puede ser una variable de tipo diferente. En una tabla, pueden coexistir columnas con información numérica, entera, decimal; otras con información cualitativa de caracteres, otras lógicas, etc. Lo más frecuente es que estas tablas tengan dos dimensiones (filas y columnas), pero en algún caso pueden tener más de dos dimensiones. Para construir una estructura de tipo data.frame, se utiliza la función data.frames(v1,v2,...,v(n-1),vn) en que cada vector (vi) contiene todos los datos de cada variable.

```
x <- c(2,2,1,2,1,1,1,2,2,1,1,2)  n <- length(x);n
```

```
## [1] 12
```

```
sex <- rep("Boy",n);sex
```

```
## [1] "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy"
      "Boy" "Boy"
```

```
sex[x==2]="Girl"; sex
```

```
##  [1] "Girl" "Girl" "Boy"  "Girl" "Boy"  "Boy"  "Boy"  "Girl"
        "Girl" "Boy"
## [11] "Boy"  "Girl"
```

```
age <- c(3,6,4,2,8,9,5,4,4,7,1,10) ; age
```

```
##  [1]  3  6  4  2  8  9  5  4  4  7  1 10
```

```
table <- data.frame(age,sex); table
```

```
##    age  sex          ##    age  sex
## 1    3 Girl          ## 7    5  Boy
## 2    6 Girl          ## 8    4 Girl
## 3    4  Boy          ## 9    4 Girl
## 4    2 Girl          ## 10   7  Boy
## 5    8  Boy          ## 11   1  Boy
## 6    9  Boy          ## 12  10 Girl
```

El nombre que se asocia a cada columna o variable dentro de la estructura es el nombre que tienen los vectores. Para referirnos a cada variable de la estructura de datos por separado, se utiliza el signo $ entre el nombre del data.frame y el nombre de la variable:

```
tableage
```

```
##  [1]  3  6  4  2  8  9  5  4  4  7  1 10
```

```
tablesex[4]
```

```
## [1] "Girl"
```

Si no se quiere tener que utilizar en todo momento el signo del dólar, se puede utilizar el comando attach(). Ahora podremos acceder a cualquier variable de la tabla directamente, únicamente mediante el nombre de su variable.

```
attach(table)
```

```
## The following objects are masked _by_ .GlobalEnv:
##     age, sex
```

```
sex
```

```
##  [1] "Girl" "Girl" "Boy" "Girl" "Boy" "Boy" "Boy" "Girl" "Girl" "Boy"
## [11] "Boy"  "Girl"
```

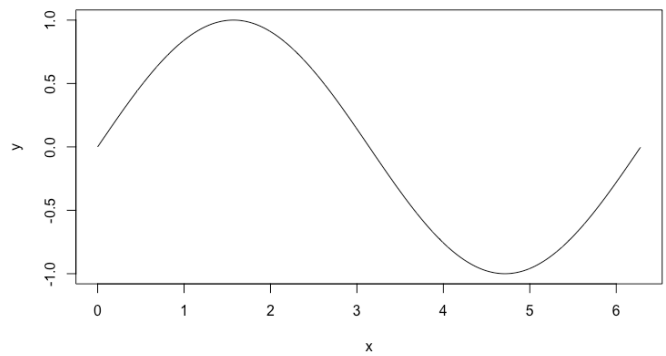
1.3.4. Funciones gráficas básicas

Otra gran ventaja de *R* son sus capacidades gráficas. Los gráficos se pueden exportar en diferentes formatos (pdf, eps, jpg, etc.). Para ver una selección de gráficos realizados con *R*, puede ejecutarse un programa de demostración mediante la siguiente instrucción:

```
demo("graphics")
```

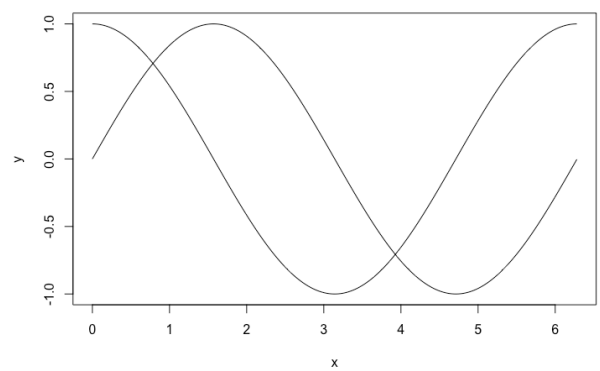
La función básica y quizás la más utilizada para generar gráficas de series o datos es plot(). Previamente, se ha de definir el vector de los datos que serán representados en el eje de las ordenadas (eje y) y, si es necesario, el vector de los datos del eje de las abscisas (eje x). Como atributo, se especifica el tipo de gráfica (puntos, líneas, ambos, etc.). Un ejemplo sencillo es el siguiente:

```
x <- seq(0,2*pi,0.01) y <- sin(x) plot(x,y,type="l")
```



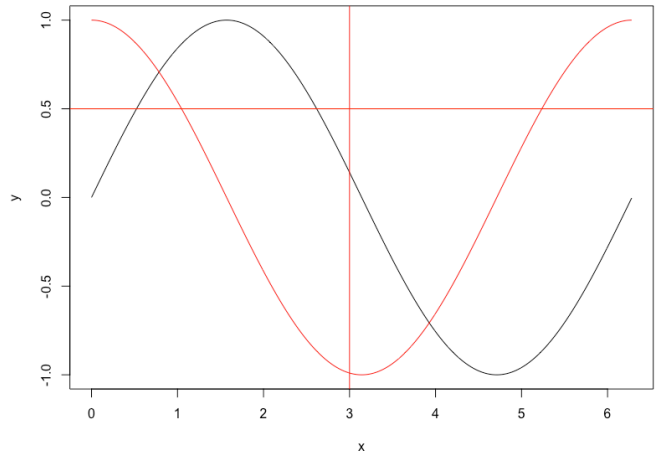
Se puede incluir la función coseno a la gráfica que está activa, utilizando la función lines():

```
x <- seq(0,2*pi,0.01) z <- cos(x) lines(x,z,type="l")
```



También, se pueden generar múltiples gráficas en una sola ventana de la siguiente forma:

```
>x=seq(0,2*pi,0.01)
>y=sin(x); z=cos(x)
>plot(x,y,type='l')
>par(col='red')
>lines(x,z)
>abline(h=0.5); abline(v=3)
```



1.3.5. Guardar y recuperar la sesión

Cuando terminamos una sesión de *R*, tenemos la opción de guardar el espacio de trabajo (*Environment*) o el historial (*History*). El espacio de trabajo incluye todos los objetos definidos por el usuario, que se van almacenando en la memoria intermedia mientras se está trabajando, pero se eliminan al cerrar una sesión. Por otra parte, el historial es el conjunto de todos los comandos que se han utilizado en la sesión. Si se quiere guardar un objeto en concreto (por ejemplo, el vector x) en el fichero MyData.Rdata, se utiliza la siguiente instrucción:

```
save(x,file="MyData.Rdata")
```

Para guardar todos los objetos existentes en el espacio de trabajo:

```
save.image("MyData.Rdata")
```

Y para recuperar todos los objetos previamente guardados en un fichero:

```
load("MyData.Rdata")
```

Si se desea guardar o cargar un historial de comandos:

```
savehistory("MyData.Rdata") loadhistory("MyData.Rdata")
```

Mediante la barra de menú también se pueden guardar y cargar tanto los objetos del espacio de trabajo como el historial.

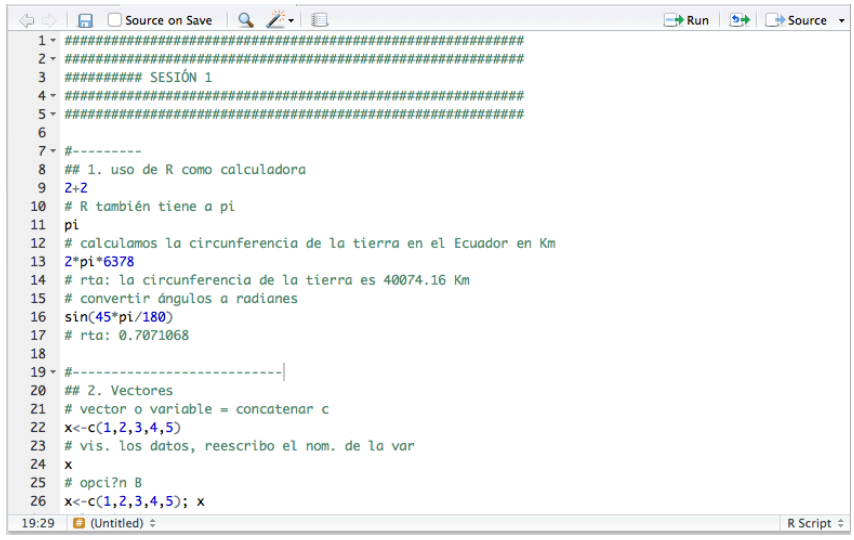
Tips & Tricks!

- Para limpiar la consola, usamos [Ctrl+L]
- Para visualizar los objetos almacenados en el espacio de trabajo, ls()
- Para eliminar un objeto, rm(name)
- Para eliminar todos los objetos, rm(list=ls())
- Para visualizar el directorio de trabajo (*working directory*), usamos getwd()
- Para ajustar el directorio de trabajo al especificado, setwd("midirectorio")

1.3.6. Scripts o guiones, la forma de organizar la sesión

Hasta ahora se ha trabajado directamente en la consola de *R* o *Rstudio* y se ha definido cómo se guardan todas las instrucciones que se han ejecutado (correctas y erróneas). Sin embargo, esta forma de guardar una sesión no es la más aconsejable. Se recomienda que el trabajo que se realice en cualquier entorno de programación, sin ser *R* la excepción, se guarde en forma de *scripts* o guiones. Un *script* no es más que un documento de texto plano que contiene el conjunto de instrucciones o códigos que se desean ejecutar. En él se pueden registrar comentarios de cada instrucción o de un conjunto de ellas. De esta manera, nuestro código queda guardado de una forma organizada y clara para poder recuperarse en una futura sesión.

Fig. 1.8
Vista de un script.



Para crear un nuevo *script*, seleccionamos en la barra de menú "Ficheros", "Nuevo fichero" y finalmente "R script". También podemos realizar esta acción con el teclado, pulsando simultáneamente *Ctrl+Shift+N*. Este documento se puede editar, modificar, guardar y ejecutar (todas las instrucciones, parte de ellas o solo las líneas deseadas). Un ejemplo de la sesión actual se puede ver en la figura 1.8

Tips & Tricks!

- Para ejecutar una línea de instrucciones desde la ventana R-script de *R-Studio*:
  - Pulsar [Ctrl+Intro] en el teclado estando el cursor en cualquier posición de esa línea.
  - Clica el botón "Ejecutar" con el ratón.
  - Para ejecutar todo el script, pulsa en el teclado [Ctrl+A] y luego [Ctrl+Intro].
- Para ejecutar una línea de instrucciones desde la ventana R-script de *R-Commander*
  - Pulsa [Ctrl+R] en el teclado estando el cursor en cualquier posición de esa línea.
  - Clica el botón "Ejecutar" con el ratón.
  - Para ejecutar todo el script, pulsa en el teclado [Ctrl+A] y luego [Ctrl+R].
- Todo el texto precedido por el carácter almohadilla # es ignorado por *R*; por tanto, se utiliza la para introducir comentarios.

1.3.7. Material extra

En los siguientes sitios web, se pueden encontrar los ficheros necesarios para instalar **R**, **R-Commander** y **Rstudio**, además de varios tutoriales para ampliar la información presentada en esta guía de prácticas.

- https://www.r-project.org/
- https://www.rstudio.com/products/rstudio/features/
- http://www.rcommander.com/
- https://support.rstudio.com/hc/en-us

Igualmente, existe una versión de **Rstudio** para trabajar directamente desde internet:

- https://rstudio.cloud/

1.4. Ejercicios propuestos

1. Crea un vector que contenga 12 valores: los cuatro primeros que sean igual a 3, los siguientes cuatro que sean igual a 6 y los últimos cuatro que sean igual a 18.
2. Introduce el vector $x = (3, 4, 30, 6, 85, 9)$:
   - Reemplaza el segundo dato por 8.
   - Introduce el número 11 entre el quinto y el sexto.
3. Crea el vector X que contiene la siguiente información: $\left[0, \frac{\pi}{16}, 2\frac{\pi}{16}, 3\frac{\pi}{16}, ..., 16\frac{\pi}{16}\right]$
   - Calcula la suma de todos sus datos: $\sum X_i$.
   - Crea un vector Y a partir del vector X eliminando los datos almacenados en las posiciones 4,9,14.
   - Calcula $\sum \text{sin}(X_i) - \sum \text{cos}(Y_i)$.
   - Compara los dos resultados anteriores y determina cuál de los dos es mayor.
4. Crea el vector X que contiene 100 datos entre -1 y 1 igualmente estpaciados.
   - Calcula la suma de todos sus datos: $\sum X_i$.
   - Calcula la suma de todos sus datos: $\sum e^{X_i}$.
5. A lo largo de un año, los importes de las facturas mensuales del móvil han sido: 23, 33, 25, 45, 10, 28, 39, 27, 15, 38, 34, 29. Escribe un *script* en *R* en que se cree una tabla de datos con meses y cargos. El *script* debe responder automáticamente las siguientes preguntas, incluso si cambia cualquier valor.
   - ¿Cuánto habéis gastado en total en el año?
   - ¿En qué mes habéis gastado menos dinero? ¿Cuánto ha sido?
   - ¿En qué mes habéis gastado más dinero? ¿Cuánto ha sido?
   - ¿En qué meses habéis gastado más que el promedio?


2
1
3

Estadística descriptiva

2.1. Introducción y objetivos

Supongamos que tenemos los siguientes datos de 1.384 pacientes con gammapatía monoclonal de significado incierto (MGUS, por sus siglas en inglés): ID del paciente (*id*), edad (*age*), sexo (*sex*), hemoglobina (*hgb*), creatinina (*creat*), tamaño del suero monoclonal (*mspike*), tiempo de progresión a la neoplasia maligna PCM (*ptime*), ocurrencia del PCM (*pstat*, 0=no, 1=sí), tiempo hasta la muerte (*futime*) y ocurrencia de la muerte (*death*, 0=no, 1=sí). A continuación, se muestran los valores de las primeras 18 observaciones.

```
##    id age sex dxyr  hgb creat mspike ptime pstat futime death
## 1   1  88   F 1981 13.1   1.3    0.5    30     0     30     1
## 2   2  78   F 1968 11.5   1.2    2.0    25     0     25     1
## 3   3  94   M 1980 10.5   1.5    2.6    46     0     46     1
## 4   4  68   M 1977 15.2   1.2    1.2    92     0     92     1
## 5   5  90   F 1973 10.7   0.8    1.0     8     0      8     1
## 6   6  90   M 1990 12.9   1.0    0.5     4     0      4     1
## 7   7  89   F 1974 10.5   0.9    1.3   151     0    151     1
## 8   8  87   F 1974 12.3   1.2    1.6     2     0      2     1
## 9   9  86   F 1994 14.5   0.9    2.4    57     0     57     0
## 10 10  79   F 1981  9.4   1.1    2.3   136     0    136     1
## 11 11  86   M 1972 11.8   1.0    2.3     2     0      2     1
## 12 12  89   F 1983 11.3   1.3    1.2   108     0    108     1
## 13 13  87   M 1968 11.2   1.1    1.3    10     0     10     1
## 14 14  80   F 1985 13.1   1.0    1.3    14     0     14     1
## 15 15  85   M 1979 13.0   1.1    1.0    18     0     18     1
## 16 16  90   F 1985 14.1   1.2    0.5    43     0     43     1
## 17 17  94   F 1975 11.0   1.1    0.7    34     0     34     1
## 18 18  86   M 1980 16.0   1.5    1.9    67     0     67     1
```

A simple vista, ¿qué información relevante podemos ver en el fichero? ¿Qué podemos concluir? Es evidente que el primer paso cuando se trabaja con datos es la exploración

inicial. Los datos se han de organizar, representar de alguna forma más amena y resumir. Por ejemplo, representar gráficamente (como se muestra en la figura) la proporción entre hombres y mujeres, analizar el rango de edad más común con la enfermedad, el valor medio de hemoglobina, etc.

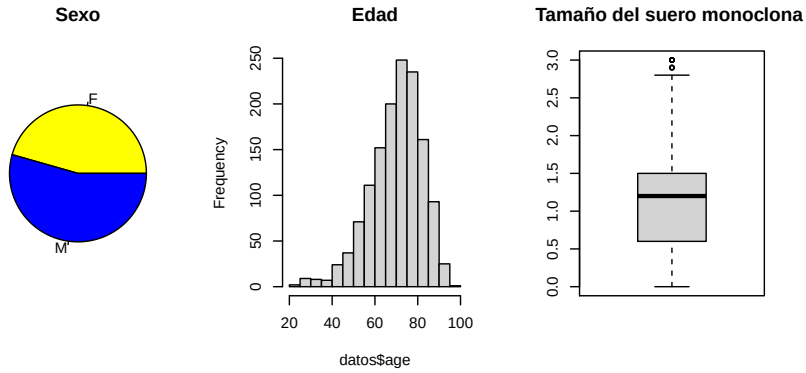


Fig. 2.1 Información resumida y organizada de los resultados de 1341 pacientes con MGUS

En esta sesión, se hace una introducción a las técnicas básicas para organizar, representar y resumir un conjunto de datos. En la estadística matemática, se conoce como **análisis exploratorio de datos** o **estadística descriptiva**. Además, se presentan las diferentes formas de aplicar la estadística descriptiva utilizando **R**, **Rstudio** y **R-Commander** en un conjunto de datos para su mejor interpretación. Al finalizar esta sesión, el alumno ha de ser capaz de:

- Identificar las principales maneras de describir, organizar, representar y resumir un conjunto de datos.
- Construir tablas de frecuencias, representarlas gráficamente, calcular algunos estadísticos importantes (media aritmética, varianza y moda) e interpretar todos estos resultados en **R**, **Rstudio** y **R-Commander**.

2.2. Estadística descriptiva

La estadística descriptiva es la disciplina de la estadística que se encarga de organizar y resumir información cuantitativa para describir las características principales de un conjunto de datos. Frecuentemente, el conjunto de datos incluye diferentes **variables** (por ejemplo: velocidad, resistencia, elasticidad, etc.). Por tanto, lo más usual es considerar las variables de una en una, sin tener en cuenta la posible correlación que existe entre ellas. Según sus características, se pueden encontrar **variables cualitativas** o **categóricas** (No necesitan números para expresarse, por ejemplo: sexo, color, etc.) y **variables cuantitativas** o **numéricas** (Sí necesitan números para expresarse, por ejemplo: edad, longitud, etc.). Por cada variable, hay una serie de observaciones; las anotaciones sobre qué modalidad (cualitativas) o qué valor (cuantitativas) tiene cada observación se denominan **datos**. Estos datos se pueden organizar, resumir y representar mediante:

- **Tablas:** Matrices donde se guardan los datos que toma una determinada variable para

cada objeto. Por ejemplo, tablas de frecuencia.

- **Gráficos:** Representaciones visuales de las tablas que otorgan una visión más general y completa de los datos. Por ejemplo, gráficos de barras, histogramas, gráficos sectoriales y polígonos de frecuencia.
- **Medidas de tendencia central:** Valores que pretenden proporcionar información sobre el centro de la distribución de datos. Algunos ejemplos son la media, la mediana y la moda.
- **Medidas de variabilidad:** Valores que pretenden proporcionar información sobre la homogeneidad de los valores entre sí. Algunos ejemplos son la desviación estándar, la varianza y los cuartiles.

2.2.1. Tabla de frecuencia

El modo más simple de presentar ordenadamente datos categóricos es mediante una tabla de frecuencias. Esta tabla indica el número de repeticiones de cada una de las clases de la variable cualitativa. Se pueden distinguir los siguientes tipos de frecuencias:

- **Frecuencia absoluta** (n_i)**:** Es el número de repeticiones que presenta una observación.
- **Frecuencia relativa** (f_i)**:** Es la frecuencia absoluta, dividida por el número total de datos.
- **Frecuencia absoluta acumulada** (N_i)**:** Es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado.
- **Frecuencia relativa acumulada** (F_i)**:** Es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos.

Ejemplo 1: El conjunto de datos para el control de calidad del agua de diferentes reactores es el siguiente, en que cada número representa el reactor que se eligió como el mejor:

1,5,3,1,2,3,4,5,1,4,2,4,4,5,1,4,2,4,2,2

Reactor	Frec. absoluta	Frec. relativa	Frec. abs. acumulada	Frec. rel. acumulada
1	4	0.20	4	0.20
2	5	0.25	9	0.45
3	2	0.10	11	0.55
4	6	0.30	17	0.85
5	3	0.15	20	1.00

En **R**, la tabla de frecuencias se puede calcular de la siguiente manera:

```
datos_1 = c(1,5,3,1,2,3,4,5,1,4,2,4,4,5,1,4,2,4,2,2)
ni = table(datos_1) Frecuencia absoluta
fi = table(datos_1)/length(datos_1)  Frecuencia relativa
Ni = cumsum(ni)  Frecuencia absoluta acumulada
Fi = cumsum(fi)  Frecuencia relativa acumulada
Tabla_Frec = cbind(ni,fi,Ni,Fi)  Tabla con todas las frecuencias
Tabla_Frec  Se visualiza la tabla
```

```
##   ni   fi Ni   Fi
## 1  4 0.20  4 0.20
## 2  5 0.25  9 0.45
## 3  2 0.10 11 0.55
## 4  6 0.30 17 0.85
## 5  3 0.15 20 1.00
```

Ejemplo 2: Las resistencias a la compresión de la aleación en libras por pulgada cuadrada (psi) de 80 especímenes de una nueva aleación de aluminio-litio sometida a evaluación como material posible para elementos estructurales de aeronaves son:

105,221,183,186,121,181,180,143,167,141,97,154,153,174,120,168,176,110,158,133,
245,228,174,199,181,158,156,123,229,146,163,131,154,115,160,208,158,169,148,158,
207,180,190,193,194,133,150,135,118,149,134,178,76,167,184,135,218,157,101,171,
165,172,199,151,142,163,145,171,160,175,149,87,160,237,196,201,200,176,150,170

Cuando los valores de la variable son muchos, conviene agrupar los datos en intervalos o clases para así realizar un mejor análisis e interpretación de ellos. Para construir una tabla de frecuencias con datos agrupados, conociendo los intervalos, se deben determinar la frecuencias correspondientes a cada intervalo.

	ni	fi	Ni	Fi
70 <= x <90	2	0.0250	2	0.0250
90 <= x <110	3	0.0375	5	0.0625
110 <= x <130	6	0.0750	11	0.1375
130 <= x <150	14	0.1750	25	0.3125
150 <= x <170	22	0.2750	47	0.5875
170 <= x <190	17	0.2125	64	0.8000
190 <= x <210	10	0.1250	74	0.9250
210 <= x <230	4	0.0500	78	0.9750
230 <= x <250	2	0.0250	80	1.0000

En **R**, la tabla de frecuencias con datos agrupados se puede calcular de la siguiente manera:

```
datos_2=c(105,221,183,186,121,181,180,143,167,141,97,154,153,174,120,
          168,176,110,158,133,245,228,174,199,181,158,156,123,229,146,
          163,131,154,115,160,208,158,169,148,158,207,180,190,193,194,
          133,150,135,118,149,134,178,76,167,184,135,218,157,101,171,
          165,172,199,151,142,163,145,171,160,175,149,87,160,237,196,
          201,200,176,150,170)
breaks = seq(70,250,by=20);
breaks  Se crea el vector que contiene los intervalos
```

```
##  [1]  70  90 110 130 150 170 190 210 230 250
```

```
datos_2a = cut(datos_2, breaks, right=FALSE)
Asigna c/valor a un intervalo
head(datos_2a, n=40)  Visualiza los primeros 40 elementos
```

```
##  [1] [90,110)  [210,230) [170,190) [170,190) [110,130) [170,190) [170,190)
##  [8] [130,150) [150,170) [130,150) [90,110)  [150,170) [150,170) [170,190)
## [15] [110,130) [150,170) [170,190) [110,130) [150,170) [130,150) [230,250)
## [22] [210,230) [170,190) [190,210) [170,190) [150,170) [150,170) [110,130)
## [29] [210,230) [130,150) [150,170) [130,150) [150,170) [110,130) [150,170)
## [36] [190,210) [150,170) [150,170) [130,150) [150,170)
## 9 Levels: [70,90) [90,110) [110,130) [130,150) [150,170) ... [230,250)
```

```
ni = table(datos_2a)  Frecuencia absoluta
fi = table(Datos_2a)/length(Datos_2a)  Frecuencia relativa
Ni = cumsum(ni)  Frecuencia absoluta acumulada
Fi = cumsum(fi)  Frecuencia relativa acumulada
Tabla_Frec = cbind(ni,fi,Ni,Fi)  Se crea una tabla con todas
                                 las frecuencias
Tabla_Frec  Se visualiza la tabla
```

```
##  ni       fi     Ni          Fi
## [70,90)    2     0.0250      2 0.0250
## [90,110)   3     0.0375      5 0.0625
## [110,130)  6     0.0750      11 0.1375
## [130,150) 14     0.1750      25 0.3125
## [150,170) 22     0.2750      47 0.5875
## [170,190) 17     0.2125      64 0.8000
## [190,210) 10     0.1250      74 0.9250
## [210,230)  4     0.0500      78 0.9750
## [230,250)  2     0.0250      80 1.0000
```

Tips & Tricks!

- table() crea resultados tabulares de variables categóricas, o sea, determina la frecuencia absoluta de los datos.
- cumsum() calcula un vector cuyos elementos son la suma acumulada del vector de entrada.
- cbind() y rbind() combinan varios objetos de R en un solo objeto: por columnas y por filas, respectivamente.
- cut() divide el rango del vector "datos" en los intervalos "breaks" y codifica los valores de los datos de acuerdo con el intervalo a que pertenecen.

2.2.2. Gráficas estadísticas

Las distribuciones de frecuencias se pueden presentar en tablas como las anteriores, o bien en gráficas. La representación gráfica se utiliza para facilitar la comprensión de los resultados, pero no añade ninguna información extra con respecto a la que contendría una tabla de frecuencias. Sin embargo, como reza el dicho popular "Vale más una imagen que mil palabras". Existen diversos tipos de gráficas, cada una de ellas adecuada a un tipo de variables. A continuación, se describen las más utilizadas y cómo se realizan en **R** utilizando los dos ejemplos anteriores.

Diagrama de tallos y hojas

Es una herramienta que presenta una tabla de datos en un formato gráfico para ayudar a visualizar la forma de la distribución. Es una tabla en que cada dato es dividido, según su valor, en un tallo y una hoja. El último dígito del dato representa la hoja y los demás dígitos representan el tallo. Este tipo de gráficos otorgan información sobre la localización, la dispersión y los valores extremos de nuestros datos. El diagrama de tallos y hojas se calcula en R mediante la función stem(). La longitud del gráfico se puede modificar utilizando el atributo scale= , donde 1 es el valor por defecto, 2 produce un gráfico aproximadamente el doble de largo, etc. El gráfico del ejemplo 2 se genera de la siguiente forma:

```
>stem(datos_2,scale=2)
The decimal point is 1 digit(s) to the right of the |
  7 | 6              16 | 0003357789
  8 | 7              17 | 0112445668
  9 | 7              18 | 0011346
 10 | 15             19 | 034699
 11 | 058            20 | 0178
 12 | 013            21 | 8
 13 | 133455         22 | 189
 14 | 12356899       23 | 7
 15 | 001344678888   24 | 5
```

Gráfico de puntos

Cuando se tiene una tabla de frecuencias pequeña de variables categóricas y los valores no distan mucho entre sí, se trata de representar los datos obtenidos de una forma atractiva. En este gráfico, el eje horizontal representa los posibles valores de los datos y el eje vertical corresponde a la localización de cada dato dentro de la lista. Cada dato se representa con un punto y se coloca encima del valor que le corresponda y a una altura proporcional al orden que tiene en el conjunto. En el ejemplo 1, el primer dato (1) se representa por un punto en la posición (1,1); el segundo dato (5), por un punto en la posición (5,2); el tercero, que es 3, en la posición (3,3); el cuarto, que es 1, en (1,4); y así sucesivamente. El gráfico de puntos se genera con el comando dotchart(), tal como se muestra a continuación.

```
dotchart(datos_1)
```

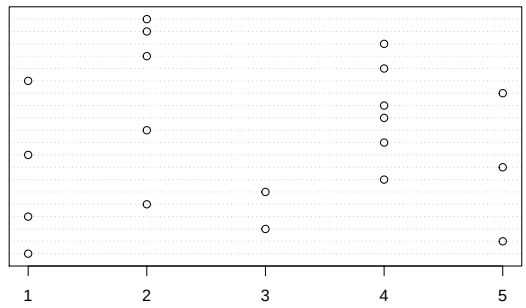
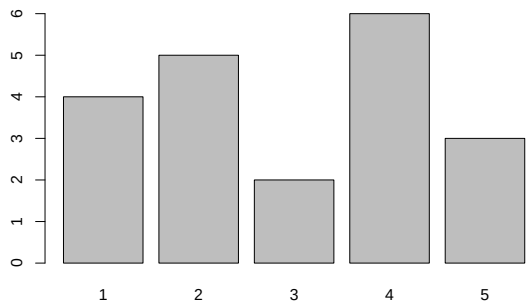


Gráfico de barras

Este gráfico representa visualmente la frecuencia de variables categóricas mediante barras rectangulares de igual anchura. A cada categoría o clase de variable se le asocia una barra cuya altura representa la frecuencia absoluta o la frecuencia relativa de esa clase. Para generar el gráfico de barras, se utiliza el comando barplot(); sin embargo, es necesario definir primero la tabla de frecuencias. Para el ejemplo 1, el gráfico de barras para la frecuencia absoluta es el siguiente:

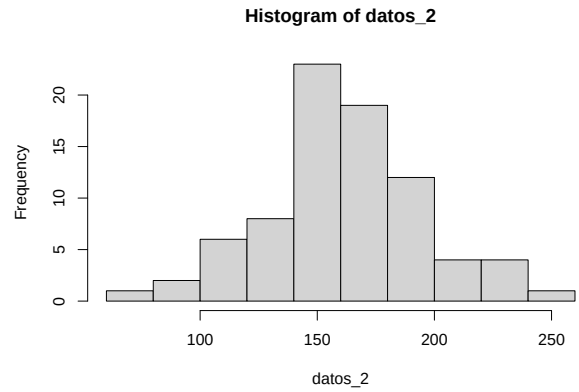
```
barplot(table(datos_1))
```



Histograma

Es la gráfica adecuada para representar variables cuantitativas con un gran número de valores distintos. Los datos se agrupan en intervalos y se representan gráficamente por rectángulos yuxtapuestos cuyas bases descansan sobre el eje horizontal y cuyas alturas son tales que el área de cada rectángulo es proporcional a la frecuencia de cada intervalo. Si todos los intervalos tienen igual longitud, entonces la altura de cada rectángulo es proporcional a la frecuencia del intervalo. Para evitar confusiones, la diferencia principal con el gráfico de barras es la inexistencia de espacios entre rectángulos. La función hist() permite hacer el histograma de unos datos y, además, modificar la longitud de los intervalos, si se desea. A diferencia del gráfico de barras, la función calcula automáticamente la frecuencia del intervalo. El histograma del ejemplo 2 se genera de la siguiente forma:

```
h=hist(datos_2)
```



Si el único argumento de la función es el vector de datos, el histograma se realiza con el número de intervalos (y, por tanto, su longitud) calculados de forma automática. Si el histograma se guarda en un objeto h = hist(), este objeto contiene determinada información, como los límites de los intervalos, la frecuencia de cada intervalo, su densidad, el punto medio, etc.

```
h$breaks  Límites de los intervalos
```

```
##  [1]  60  80 100 120 140 160 180 200 220 240 260
```

```
h$counts  Frecuencia de cada intervalo
```

```
##  [1]  1  2  6  8 23 19 12  4  4  1
```

```
h$density  Densidad de cada intervalo
```

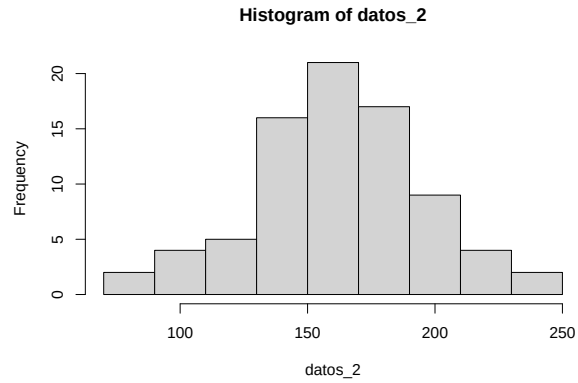
```
##  [1] 0.000625 0.001250 0.003750 0.005000 0.014375 0.011875
        0.007500 0.002500
##  [9] 0.002500 0.000625
```

```
h$mids  Punto central de cada intervalo
```

```
##  [1]  70  90 110 130 150 170 190 210 230 250
```

También se pueden seleccionar los límites de los intervalos o el número de intervalos en que se quieren agrupar.

```
new_breaks = seq(70,250,by=20)
h1 = hist(datos_2,breaks=new_breaks)
```




```
h2=hist(datos_2,breaks = 3)
```

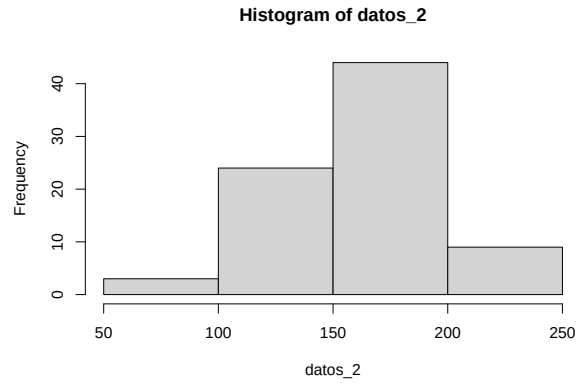
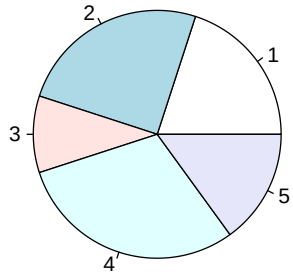


Gráfico de sectores

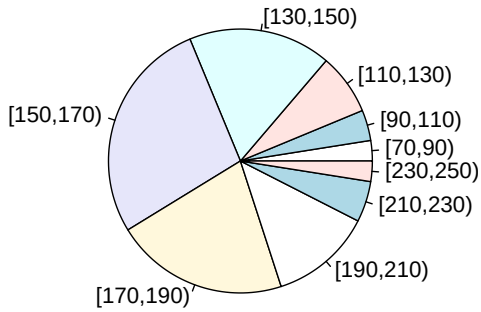
Este gráfico se representa como un círculo dividido en porciones, que son proporcionales a la frecuencia relativa de cada categoría. La función pie() permite realizar el gráfico de sectores. Al igual que en el gráfico de barras, es necesario definir previamente la tabla de frecuencias. Para el ejemplo 1, el diagrama se realiza de la siguiente forma:

```
pie(table(datos_1))
```



Si se quiere hacer el gráfico de sectores para el ejemplo 2, los datos se tienen que agrupar para poder visualizar información relevante.

```
pie(table(datos_2a))
```



Tips & Tricks!

- Las funciones stem(), dotchart(), barplot(), hist() y pie() permiten resumir visualmente los datos.
- Estos gráficos se pueden mejorar definiendo algunos atributos, como por ejemplo: col, main, names.arg, etc. Utiliza la ayuda para conocer un poco más sobre ellos.

2.2.3. Medidas de posición y tendencia central

En ocasiones, es conveniente resumir la información de un conjunto de datos numéricos en un solo valor para obtener indicadores del comportamiento de la variable y poder realizar comparaciones. Las medidas de tendencia central, también conocidas como *medidas de posición* o *localización*, describen un valor en torno del cual se encuentran las observaciones.

Media

También conocida como el valor medio, se define como la suma de todos los valores de cada observación (x_i), dividido por el número total de observaciones del conjunto de datos (N).

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si se dispone de un conjunto de datos agrupados en que se conoce el valor medio de cada intervalo $(\bar{x_i})$ y el número de datos de cada uno de ellos (n_i) , la media viene dada por:

$$\bar{X} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \cdots + x_N n_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i n_i$$

donde $n_1 + n_2 + n_3 + n_4 + \cdots + n_n = N$. Para los ejemplos anteriores, las medias se puede calcular conforme a la definición de la siguiente manera:

```
sum(datos_1)/length(datos_1)
```

```
## [1] 2.95
```

```
sum(datos_2)/length(datos_2)
```

```
## [1] 162.6625
```

Sin embargo, la función mean() calcula la media directamente.

```
mean(datos_1)
```

```
## [1] 2.95
```

```
mean(datos_2)
```

```
## [1] 162.6625
```

Mediana

La mediana es el dato que ocupa la posición central en la muestra ordenada de menor a mayor; es un punto que divide la muestra ordenada en dos grupos iguales (deja el 50 % de los valores por debajo y el otro 50 % por encima). Para calcularla, se ordenan los datos de menor a mayor, y el dato central es el que ocupa la posición $\frac{N+1}{2}$ donde N es el número total de datos. Si N es impar, la mediana es el mismo dato central; si N es par, existen dos datos centrales, por lo que la mediana es el promedio de ambos. Igualmente, existe una función que la calcula directamente: median().

```
median(datos_1)
```

```
## [1] 3
```

```
median(datos_2)
```

```
## [1] 161.5
```

Moda

La moda es el valor con mayor frecuencia absoluta en los datos obtenidos. Indica cuál es el valor más frecuente, pero no cuántas veces se repite. Si existen más de dos valores que se repiten con mayor frecuencia, se dice que los datos son *multimodales*. Se puede calcular la moda sigiendo las siguientes instrucciones:

```
table(datos_1)
```

```
## datos\_1
## 1 2 3 4 5
## 4 5 2 6 3
```

```
Se organiza la tabla de frecuencias de mayor valor
(el más frecuente) a menor
freq_ord=sort(table(datos_1), decreasing = TRUE); freq_ord
```

```
## datos\_1
## 4 2 1 5 3
## 6 5 4 3 2
```

```
Se toma el/los valor/es que más se repite/n
(el primero de la tabla ordenada)
moda = names(freq_ord[1]); moda
```

```
## [1] "4"
```

Cuantiles

Los cuantiles son valores de la lista de datos que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores. Los más usados son los percentiles, los deciles y los cuartiles. Los percentiles son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Por ejemplo, el percentil de orden 15 deja por debajo al 15 % de las observaciones y por encima quedan el 85 %. Los deciles son los nueve valores que dividen el conjunto de datos ordenados en diez partes iguales; son un caso particular de los percentiles. Los cuartiles son los tres valores que dividen el conjunto de datos ordenados en cuatro partes iguales; son también un caso particular de los percentiles. En *R*, cualquiera de estos se calcula con la función quantile(), donde adicionalmente se ha de especificar el cuantil o los cuantiles deseados (como un valor entre 0 y 1) de la siguiente forma:

```
quantile(datos_2,0.95)  Percentil de orden 95
```

```
##    95%
## 221.35
```

```
quantile(datos_2,seq(0.1,0.9,by=0.1))  Todos los deciles
```

```
##   10%   20%   30%   40%   50%   60%   70%   80%   90%
## 119.8 135.0 149.0 156.6 161.5 170.4 176.6 186.8 201.6
```

```
quantile(datos_2,seq(0.25,0.75,by=0.25))  Todos los cuartiles
```

```
##   25%   50%   75%
## 144.5 161.5 181.0
```

Finalmente, el rango intercuartílico es la extensión cubierta por la mitad central de los datos ordenados, excluyendo la cuarta parte inicial (los que son inferiores al primer cuartil) y la cuarta parte final (los que son superiores al tercer cuartil). La función IQR() calcula directamente el rango intercuartílico.

```
quantile(datos_2,0.75) - quantile(datos_2,0.25)
```

```
##  75%
## 36.5
```

```
IQR(datos_2)
```

```
## [1] 36.5
```

La media, la mediana, el mínimo, el máximo y los cuartiles se pueden calcular directamente mediante la función summary().

```
summary(datos_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    76.0   144.5   161.5   162.7   181.0   245.0
```

2.2.4. Medidas de variabilidad y dispersión

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no nos dicen cuán disperso es el conjunto de datos. Las medidas de dispersión o variabilidad describen lo cerca que se encuentran los datos entre ellos o de alguna medida de tendencia central.

Rango

Es el intervalo entre el valor máximo y el valor mínimo del conjunto de datos. Es altamente sensible a los valores extremos, es decir, es un parámetro estadístico débil. Con la función range(), también se obtienen el valor mínimo y el máximo del conjunto de datos; por tanto, para calcular el rango, basta con calcular su diferencia.

```
max(datos_1)-min (datos_1)
```

```
## [1] 4
```

```
max(datos_2)-min(datos_2)
```

```
## [1] 169
```

```
diff(range(datos_1))
```

```
## [1] 4
```

```
diff(range(datos_2))
```

```
## [1] 169
```

Varianza y desviación típica

Estas medidas miden cuán lejos difieren los datos de la media. Específicamente, expresan "el promedio de la distancia de cada punto respecto de la media". La varianza se calcula según:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

donde x_i es el valor de cada observación, $\overline{X}$ es la media y N es el número total de datos. Nótese que las unidades de la varianza están expresadas al cuadrado; por tanto, si se tienen datos de longitud (en mm), la varianza resulta con unidades de superficie (en mm^2), lo cual no tiene mucho sentido. Así pues, se dispone de la desviación estándar o típica, que no es más que la raíz cuadrada de la varianza; de esta forma, las unidades de la medida de dispersión son las mismas de los datos.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}$$

Para el ejemplo 1, la varianza y la desviación típica se pueden calcular usando la definición de la siguiente forma:

```
sum((datos_1-mean(datos_1))2)/length(datos_1)  Varianza
```

```
## [1] 1.9475
```

```
sqrt(sum((datos_1-mean(datos_1))2)/length(datos_1))  Desviación típica
```

```
## [1] 1.395529
```

En *R*, la varianza y la desviación estándar se pueden calcular mediante las funciones var () y sd(), respectivamente; sin embargo, estas funciones utilizan $N-1$ (o $\sqrt{N-1}$) en el denominador, en lugar de N (o $\sqrt{N}$), para poderlas usar como estimadores no sesgados en inferencia estadística. Estas medidas se conocen como; varianza y la desviación típica corregidas. Por tanto, para conocer la varianza y la desviación típica sin corregir, se tienen que multiplicar por los factores $\frac{N-1}{N}$ y $\sqrt{\frac{N-1}{N}}$, respectivamente.

```
N = length(datos_1) var(datos_1)  Varianza corregida
```

```
## [1] 2.05
```

```
((N-1)/N)*var(datos_1)  Varianza NO corregida
```

```
## [1] 1.9475
```

```
sd(datos_1)  Desviación típica corregida
```

```
## [1] 1.431782
```

```
sqrt((N-1)/N)*sd(datos_1)  Desviación típica NO corregida
```

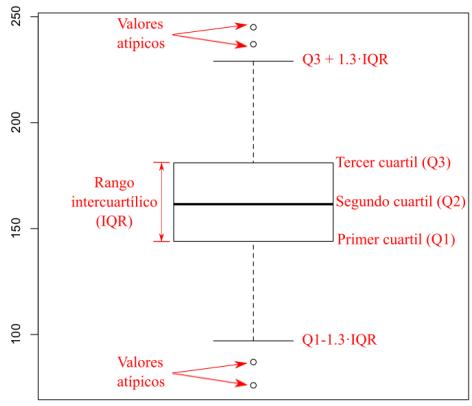
```
## [1] 1.395529
```

2.2.5. Gráfico de caja

Los diagramas de caja son una presentación visual que describe varias características importantes al mismo tiempo, tales como la tendencia central, la dispersión y la simetría. Para su realización, se representan los tres cuartiles y los valores mínimo y máximo de los datos sobre un rectángulo, alineado horizontal o verticalmente. Los valores con dispersión hasta 1.3 veces el rango intercuartílico se representan como unas líneas rectas o bigotes. Los valores fuera de ese intervalo se representan mediante puntos y se consideran valores extremos atípicos.

boxplot() es la función que se utiliza para la creación del gráfico. Al igual que como con el histograma, si se guarda el gráfico de caja en un objeto h = boxplot(), este objeto contiene información tales como: los límites para considerar los valores atípicos, los valores atípicos, los cuartiles, etc.

```
bp = boxplot(datos_2); bp
```




```
## $stats
##       [,1]
## [1,]  97.0
## [2,] 144.0
## [3,] 161.5
## [4,] 181.0
## [5,] 229.0
##
## $n
## [1] 80
##
## $conf
##          [,1]
## [1,] 154.964
## [2,] 168.036
##
## $out
## [1] 245  76  87 237
##
## $group
## [1] 1 1 1 1
##
## $names
## [1] "1"
```

Tips & Tricks!

- Las funciones mean(), median(), quantile(), IQR(), var(), sd() y boxplot() nos dan información sobre la tendencia central y la variabilidad de los datos.
- Recuerde que puede consultar más información sobre cada función mediante la instrucción ?NombreDeLaFuncion, por ejemplo ?boxplot.

2.3. Ejercicios propuestos

Los siguientes datos se extrajeron de la revista estadounidense *Motor Trend* en 1974 y resumen el consumo y diez aspectos de diseño y rendimiento de 32 automóviles (modelos 1973-1974). Este conjunto de datos, que se denomina mtcars, contiene 11 variables con 32 observaciones y está almacenado en **R** . Para poder trabajar con ellos, solo hace falta adjudicarle un nombre al objeto, como por ejemplo:

```
a = mtcars
```

Las variables son las siguientes:

mpg:	millas por galón de combustible
cyl:	número de cilindros
disp:	desplazamiento
hp:	caballos de potencia
drat:	relación del eje trasero
wt:	peso (1000 lbs)
qsec:	tiempo a 1/4 milla
vs:	V/S
am:	transmisión (0 = automático, 1 = manual)
gear:	número de marchas adelante
carb:	número de carburadores

Una vez cargado el conjunto de datos, procede a la resolución del cuestionario.

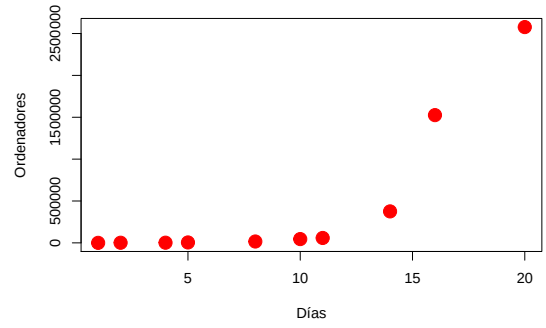
1. Determina la media, la mediana, la moda y la desviación estándar de cada una de las variables. Se puede calcular para todas las variables? para cuáles no? Justifica la respuesta
2. Determina qué variable presenta valores atípicos. Cómo los has encontrado?
3. Realiza un gráfico de sectores para cada una de las variables. El gráfico de qué variables no cumple con el objetivo de “impactar” o “ser más claro” que la tabla de datos?
4. Realiza el histograma para cada una de las variables usando 5 intervalos. De nuevo, ¿esta gráfica es útil para todas las variables? Justifica la respuesta.
5. Realiza una gráfica que incluya el diagrama de cajas de todas las variables de modo de que se puedan comparar.

Regresión lineal

3.1. Introducción y objetivos

En la sesión 2 “Estadística descriptiva”, hemos aprendido a resumir y a presentar un conjunto de datos en tablas y gráficas. Estas ayudan a interpretarlos para la toma de decisiones. Sin embargo, en algunas ocasiones, cuando se tienen los datos organizados y representados en una gráfica, puede surgir la necesidad de estimar un valor que no se tiene, por diferentes motivos, ya sea por pérdida de la información, por un error o fallo del sensor, o por fallo técnico, etc. Por ejemplo, el número de días que han pasado desde que se ha detectado un virus informático y el número de ordenadores infectados en una determinada region de Europa están registrados en la siguiente tabla y representados en su diagrama de dispersión:

Días	Ordenadores
1	255
2	1500
4	2105
5	5050
8	16300
10	45320
11	58570
14	375800
16	1525640
20	2577000



En esta sesión, se podrá dar respuesta a una serie de preguntas que pueden surgir, como: ¿Cuántos ordenadores se infectaron a los 12 días? ¿Qué cantidad de ordenadores estarán infectados a los 22 días? El caso de los ordenadores infectados a los 12 días, es un evento del pasado que no ha sido medido. Por contra, estimar la cantidad de ordenadores que estarán infectados a los 22 días corresponde a un evento futuro inmediato (no muy lejano). Las dos cuestiones se pueden resolver aplicando la técnica de la regresión lineal.

Esta técnica es un modelo matemático básico de regresión y de análisis predictivo de uso común. La idea general de la regresión es resumir y estudiar la relación entre dos variables continuas, hallar la relación matemática lineal que existe entre ellas. Sin embargo, un buen observador puede inferir que, por ejemplo, la relación entre los días y el número de ordenadores NO es lineal. Esta "complicación" puede ser fácilmente resuelta para después aplicar la técnica de regresión lineal.

Ahora bien, generalmente los conjuntos de datos con que habitualmente se trabaja en estadística suelen ser extensos o simplemente se obtienen de algún equipo de medición. Entonces, lo más cómodo es que estén guardados en ficheros típicos de hojas de cálculo o como ficheros de texto. Para esta tercera sesión, además de aplicar los conceptos de regresión lineal, también se utiliza un conjunto de datos guardado en un fichero externo. Por tanto, en la primera parte de la sesión se detalla el procedimiento para importarlos a los entornos con los cuales se puede trabajar: **R-Console**, **R-Commander** y **Rstudio**. Adicionalmente, se estudian las dos medidas que describen la bondad de la relación encontrada en la regresión lineal. Finalmente, se presentan las instrucciones básicas para desarrollar la técnica utilizando **R-Console**, **R-Commander** y el entorno **Rstudio**. Al finalizar esta sesión, el alumno ha de ser capaz de:

- Importar datos utilizando la consola **R**, así como el paquete **R-Commander** y el entorno **Rstudio**.
- Comprender el concepto del criterio de mínimos cuadrados.
- Interpretar la intercepción y la pendiente de una ecuación de regresión estimada.
- Calcular, entender e interpretar el coeficiente de determinación y el coeficiente de correlación.
- Saber obtener la estimación de la línea recta de regresión, los coeficientes de determinación y de correlación utilizando **R**.
- Estimar los valores faltantes o desconocidos a partir de la relación encontrada.

3.2. Importar datos a R-Console Rstudio y R-Commander

Generar una estructura de datos en **R-Console** puede ser dispendioso según el tamaño de los datos. Por ello, en algunos casos es más fácil o cómodo crearlos y/o editarlos previamente utilizando una hoja de cálculo. Por otra parte, muchos datos los podemos tener ya en otros programas o simplemente pueden proceder directamente de algún software de adquisión de datos. A continuación, se explica brevemente la manera de importar datos utilizando **R-Console**, así como el paquete **R-Commander** y el entorno **Rstudio**.

3.2.1. Importar datos con R-Console

Los datos contenidos en archivos externos se pueden importar directamente desde la consola mediante la función read.table(file,header,sep,dec,...), donde se ha de especificar en

primera instancia el nombre del fichero. Dependiendo de la forma en que los datos están organizados dentro del fichero, se han de definir los siguientes atributos heater, sep, dec:

- **file** es el nombre del archivo donde están guardados los datos.
- **header** es un tipo de dato lógico que indica si el archivo tiene en la primera fila los nombres de las variables.
- **sep** es el carácter que separa las magnitudes entre sí.
- **dec** es el carácter que separa la parte entera de la parte decimal de un número cualquiera (normalmente, es un punto o una coma).

Por ejemplo, al abrir el fichero coches.txt en cualquier editor de texto, se puede apreciar que el nombre de las variables está incluido en la primera fila del fichero, los datos están separados por tabulaciones y el carácter decimal está definido por el punto, tal como se muestra en la figura.

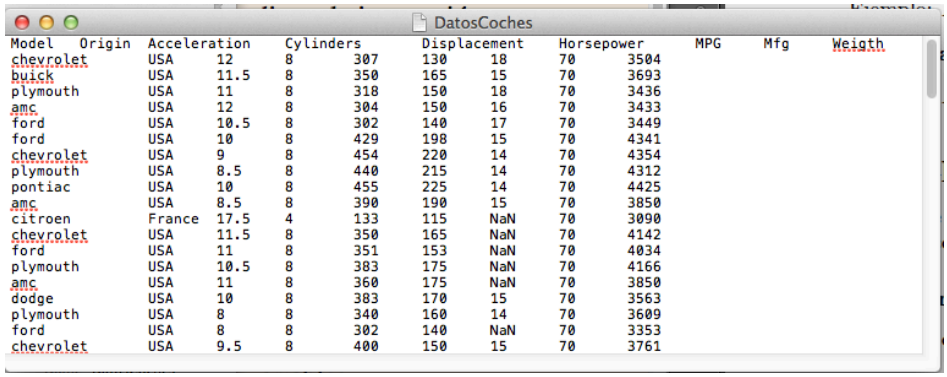


Fig. 3.1
Vista del fichero coches.txt

Para importar sus datos y guardarlos en una estructura de datos (data.frame) llamada Datos, se ejecuta:

```
Datos = read.table("datos_Sesion3/coches.txt",
                   header=TRUE, sep="\t", dec=".")
head(Datos)
```

```
##           Model  Origin Acceleration Cylinders Displacement
## 1 chevrolet     USA             12.0         8          307
## 2 buick         USA             11.5         8          350
## 3 plymouth      USA             11.0         8          318
## 4 amc           USA             12.0         8          304
## 5 ford          USA             10.5         8          302
## 6 ford          USA             10.0         8          429

##   Horsepower MPG Mfg Weight
```

```
## 1          130   18   70   3504
## 2          165   15   70   3693
## 3          150   18   70   3436
## 4          150   16   70   3433
## 5          140   17   70   3449
## 6          198   15   70   4341
```

3.2.2. Importar datos conRstudio

Para importar datos desde **Rstudio**, debemos ir a Tools >Import Dataset >From Local File. Ahí seleccionamos el fichero que queremos importar. En nuestro caso, vamos a importar el fichero coches2.txt.

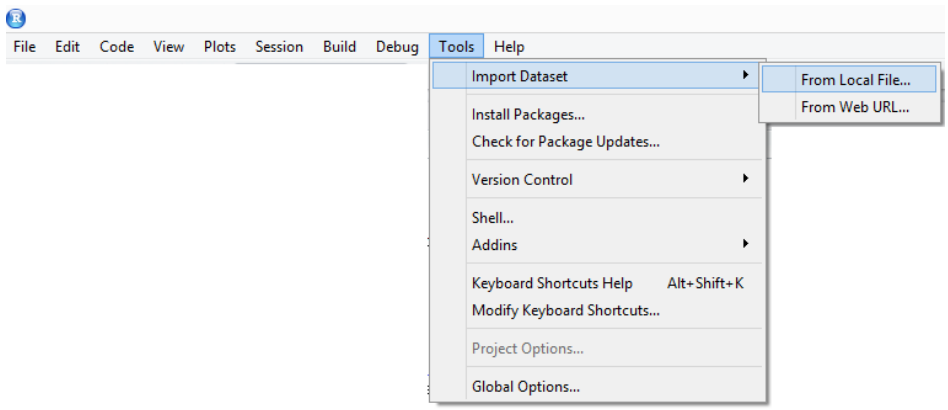


Fig. 3.2 Importar un fichero de datos desde Rstudio

Después de seleccionar el fichero, en la ventana emergente "Import Dataset" se han de indicar diversas características que conciernen al fichero (al igual que en **R-Console**): si el nombre de las variables está incluido en la primera fila del fichero o no, cómo están separados los datos, el carácter utilizado para separar los números decimales, etc.

Una vez realizada con éxito la importación de los datos, en la esquina superior derecha, en la pestaña de "Environment" se puede observar que, entre las variables actuales en el espacio de trabajo, se encuentra el fichero "Datos". Nótese que se especifica que es una estructura de datos (data.frame) de 100 observaciones de 9 variables.

3.2.3. Importar datos con R-commander

Si **R-Commander** no está en ejecución, basta con cargarlo desde la **R-Console** o **Rstudio**.

```
library(Rcmdr)
## Loading required package: splines
```

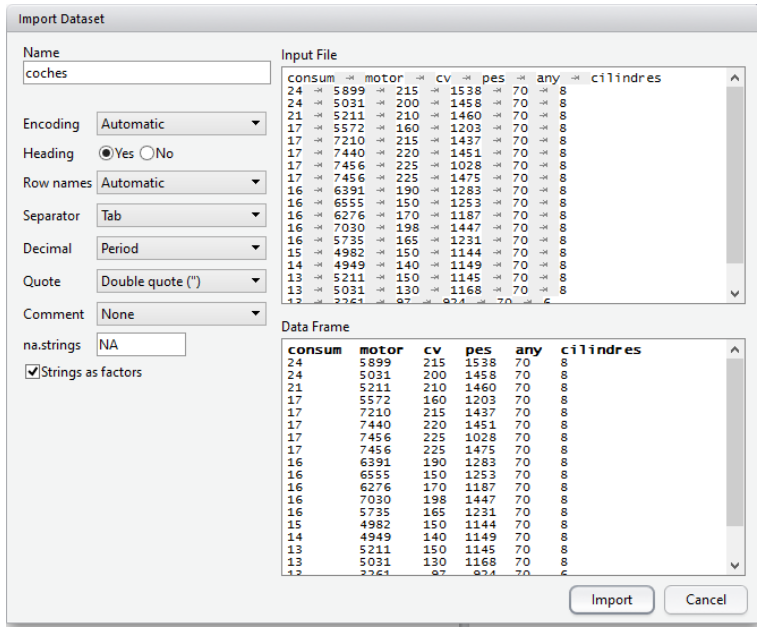


Fig. 3.3
Selección de características del fichero de datos desde Rstudio

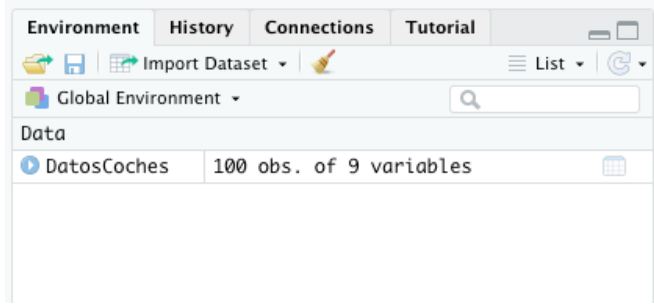


Fig. 3.4
Vista de las variables en el entorno de trabajo en Rstudio

```
## Loading required package: RcmdrMisc
## Loading required package: car
```

Para importar datos desde **R-Commander**, debemos ir a Datos >Importar datos y seleccionar si el origen de los datos es un archivo de texto, portapapeles o URL.

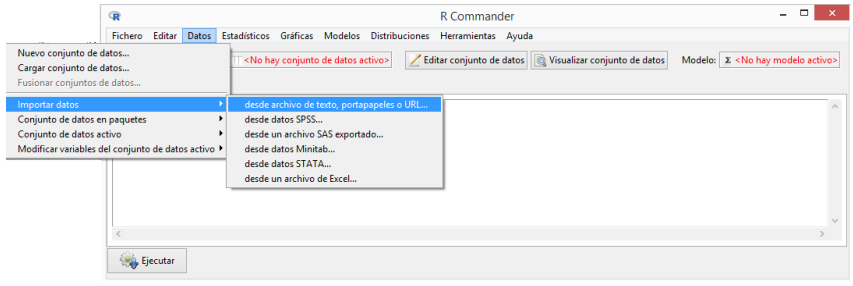


Fig. 3.5
Importar un fichero de datos desde R-Commander

De igual forma que en **R-console** y **Rstudio**, se han de indicar las características del fichero de datos.

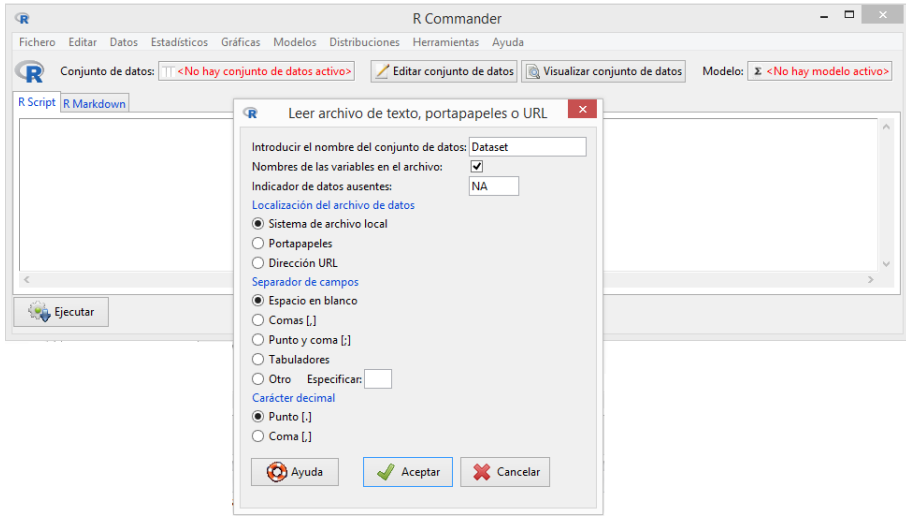


Fig. 3.6 Selección de características del fichero de datos desde R-Commander.

Nótese que, en la ventana principal, debajo del menú, la casilla Conjunto de datos ahora aparece activada con el nombre que le hemos dado al importar el fichero, en nuestro caso Dataset. Por otra parte, en la pestaña del "R Script" aparece la línea de instrucciones que se pueden ejecutar desde la **R-console** o **Rstudio**.

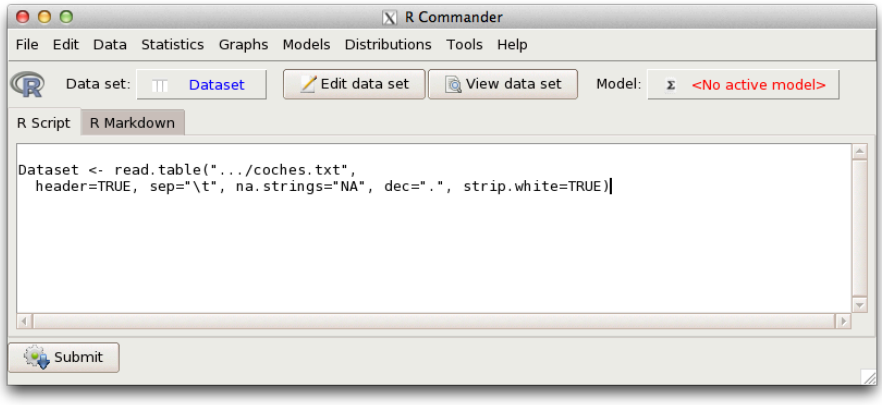


Fig. 3.7 Vista del R Script en R-Commander después de haber importado un fichero de datos.

Tips & Tricks!

- Para ejecutar una línea de instrucciones desde la ventana R Script de **Rstudio**:
  - Pulsa [Ctrl+Intro] en el teclado estando el cursor en cualquier posición de esa línea.
  - Clica el botón "Ejecutar" con el ratón.
  - Para ejecutar todo el *script*, pulsa en el teclado [Ctrl+A] y luego [Ctrl+Intro].
- Para ejecutar una línea de instrucciones desde la ventana R Script de **R-Commander**:
  - Pulsa [Ctrl+R] en el teclado estando el cursor en cualquier posición de esa línea.
  - Clica el botón "Ejecutar" con el ratón.
  - Para ejecutar todo el *script*, pulsa en el teclado [Ctrl+A] y luego [Ctrl+R].

3.3. Regresión lineal

La regresión lineal es una herramienta estadística que aporta la habilidad de estimar la relación matemática entre una variable dependiente (o respuesta, normalmente y) y una variable independiente (o predictor, normalmente x). Su objetivo principal es utilizar la información obtenida sobre un fenómeno para predecir su comportamiento en el futuro. Esta información suele estar organizada por parejas de valores observados y se representa gráficamente en una nube de puntos o diagrama de dispersión.

Ejemplo 1: La relación entre el tamaño del lote en la fábrica de cierto producto y las horas de trabajo necesarias viene dada por la siguiente tabla.

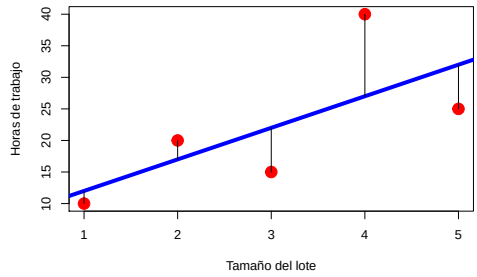
Lotes	Horas
1	10
2	20
3	15
4	40
5	25

3.3.1. Modelo de regresión lineal simple

La regresión lineal simple consiste en encontrar la línea recta ($\hat{y} = mx + b$) que se ajuste mejor a través de los puntos. La línea que se ajusta mejor se denomina **línea de regresión** o **recta de regresión**. En la figura 3.8, se puede apreciar la gráfica de dispersión de los datos del ejemplo 1 (puntos en rojo). La línea diagonal azul es la línea de regresión y

representa la predicción en y para cada valor posible de x. Las líneas verticales desde los puntos hasta la línea de regresión representan los errores de predicción $(y_i - \hat{y}_i)$. Como se puede ver, el punto en $x = 1$ está muy cerca de la línea de regresión, de modo que, su error de predicción es pequeño. Por el contrario, el punto en $x = 4$ está mucho más lejos de la línea de regresión y, por tanto, su error de predicción es mayor.

Fig. 3.8 Diagrama de dispersión y línea de regresión



La línea recta que se ajusta mejor a los datos es aquella para la cual los n errores de predicción (uno por cada punto de datos) son tan pequeños como sea posible en sentido general. Una forma de lograr este objetivo es invocar el "criterio de los mínimos cuadrados", consistente en "minimizar la suma de los errores de predicción al cuadrado". Es decir, se han de buscar los valores de m (pendiente) y b (intercepción) tales que la suma del cuadrado de los errores de predicción $Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ sea la más pequeña posible.

Por tanto, se ha de minimizar la ecuación:

$$Q = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

Derivando e igualando a cero, se obtiene:

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{y} \qquad b = \bar{y} - m\bar{x}$$

Para el ejemplo 1, se tiene:

x	y	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	10	-2	4	-12	24
2	20	-2	1	-2	2
3	15	0	0	-7	0
4	40	1	1	18	18
5	25	2	4	3	6
$\bar{x} = 3$	$\bar{y} = 22$		$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 10$		$\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = 50$

Por tanto, $m = \frac{50}{10} = 5$ y $b = 22 - 5 \times 3 = 7$. De esta manera, la recta de regresión viene dada por: $\hat{y} = 5x + 7$.

3.3.2. Modelo de regresión exponencial

En otros casos, la línea que une los valores obtenidos no se aproxima a una recta sino a una función exponencial. En otras palabras, se asemeja a una función tipo $y = \alpha e^{\beta x}$. Por tanto, la regresión consiste en encontrar los valores de α y β que se ajusten mejor a los datos. En este tipo de regresiones, también podemos encontrar el valor del coeficiente de determinación R^2, el cual sigue el mismo criterio que para la regresión lineal (cuanto más cerca esté de 1, más precisa será la aproximación).

Aunque no lo parezca, esta aproximación no es más difícil que la lineal. Debido a las propiedades de los logaritmos neperianos, una relación exponencial se puede convertir en una relación lineal de una forma muy sencilla:

$$\ln(y) = \ln(\alpha e^{\beta x}) = \ln(\alpha) + \ln(e^{\beta x}) = \ln(\alpha) + \beta \ln(e^{x}) = \ln(\alpha) + \beta x$$

La solución al problema inicial sería la regresión lineal entre x y $\ln(y)$.
Retomando el ejemplo de la introducción de esta sesión sobre el registro del número de días que han pasado desde que se ha detectado un virus informático y el número de ordenadores infectados, se puede observar en el diagrama de dispersión que la relación entre las dos variables no es lineal, sino que tiene un comportamiento exponencial.

Dias	Ordenadores
1	255
2	1500
4	2105
5	5050
8	16300
10	45320
11	58570
14	375800
16	1525640
20	2577000

Al realizar el diagrama de dispersión entre la variable Días y el logaritmo a la variable Ordenadores log(Ordenadores), se observa que la relación tiene una tendencia lineal.

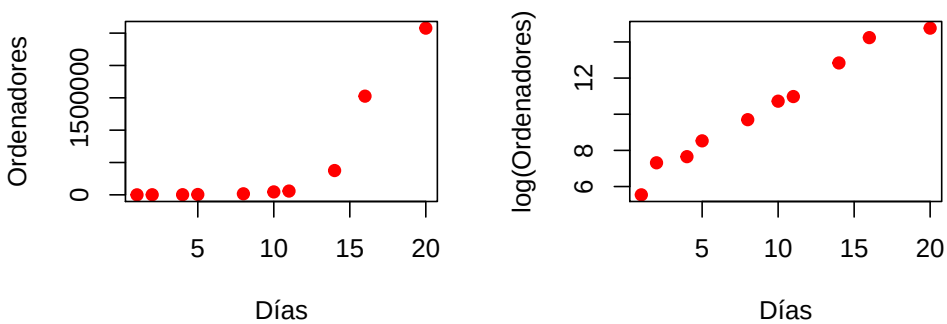


Fig. 3.9
Diagrama de dispersión: (izq.) Días versus Ordenadores, (der.) Días versus log(Ordenadores)

Por tanto, siguiendo el criterio de mínimos cuadrados, se obtiene:

$$m = \beta = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{171.0931}{354.9} = 0.4821$$

$$b = \ln(\alpha) = \bar{y}^* - m\bar{x} = 10.2269 - 0.4821 \times 9.1 = 5.8399$$

$$\alpha = e^b = e^{5.8399} = 343.7072$$

donde $\bar{y}^*$ es el logaritmo de la variable Ordenadores. De esta manera, la línea de regresión viene dada por: $\ln(y) = 0.4821 + 5.8399x$, o por $y = 343.7072e^{0.4821x}$

3.3.3. Evaluar la exactitud del modelo de regresión

Existen varias formas de evaluar en qué medida se ajusta nuestro modelo a los datos; la bondad de ajuste o la calidad de la regresión las determina normalmente el **coeficiente de determinación (R^2)** o **el coeficiente de Pearson (R)**. Estos números característicos de cada regresión indican cuán bien se ajusta la línea a los datos. Por ejemplo, $R^2 = 0.85$ quiere decir que el 85 % de la variación total en y se puede explicar por la relación lineal entre x e y. En consecuencia, cuanto más se acerque a 1, mejor se ajustará a los valores. En este caso, la línea pasa exactamente por cada punto y es capaz de detallar toda la variación. Cuanto más lejos esté de los puntos, peor será la aproximación. El coeficiente de determinación es la relación entre la variabilidad explicada por la regresión y la variabilidad total. Se calcula mediante la siguiente fórmula:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

donde $\hat{y}_i$ es la estimación del valor de y_i. En el ejemplo dado:

x	y	$\hat{y} = 5x + 7$	$(\hat{y} - \bar{y})^2$	$(y - \bar{y})^2$
1	10	12	100	144
2	20	17	25	4
3	15	22	0	49
4	40	27	25	324
5	25	32	100	9
$\bar{x} = 3$	$\bar{y} = 22$		$\sum_{i=1}^{5}(\hat{y}_i - \bar{y})^2 = 250$	$\sum_{i=1}^{5}(y_i - \bar{y})^2 = 530$

Por tanto, $R^2 = \frac{250}{530} = 0.4717$ y $R = \pm\sqrt{R^2} = \sqrt{0.4717} = 0.6868$, donde el signo viene determinado por la pendiente de la recta de regresión.

3.4. Regresión lineal con R-Console o Rstudio

A continuación, se describe detalladamente el procedimiento para realizar una regresión lineal utilizando **R**. Los datos que se utilizarán están guardados en el fichero coches.txt, que se puede descargar desde Atenea.

3.4.1. Cargar datos

Tal como se ha explicado anteriormente, se importan los datos desde **Rstudio**, **R-Commander** o **R-Console** mediante la función read.table(). Se observa que hay datos que no están disponibles NaN y, si se quiere hacer alguna operación (por ejemplo, una media), el resultado será también NaN. Este problema se puede evitar agregando un atributo a la función:

```
Datos = read.table("datos_Sesion3/coches.txt",
                   header=TRUE,sep="\t",na.strings="NA",dec=".")
mean(Datos$MPG)
```

```
## [1] NaN
```

```
mean(Datos$MPG, na.rm = TRUE)
```

```
## [1] 23.71809
```

Si resulta engorroso tener que invocar la variable por medio del nombre de la estructura Datos, más el símbolo dólar $, más el nombre propio de la variable MPG, se puede utilizar la función attach() para vincular todas las variables de la estructura de datos a la ruta de búsqueda de *R*, es decir, las variables se pueden invocar solo por sus nombres.

```
attach(Datos)
```

```
## The following objects are masked from Datos (pos = 3):
##
##     Acceleration, Cylinders, Displacement, Horsepower,
##     Mfg, Model, MPG, Origin, Weight

## The following objects are masked from Datos (pos = 4):
##
##     Acceleration, Cylinders, Displacement, Horsepower,
##     Mfg, Model, MPG, Origin, Weight
```

```
mean(MPG)
```

```
## [1] NaN
```

```
mean(MPG, na.rm = TRUE)
```

```
## [1] 23.71809
```

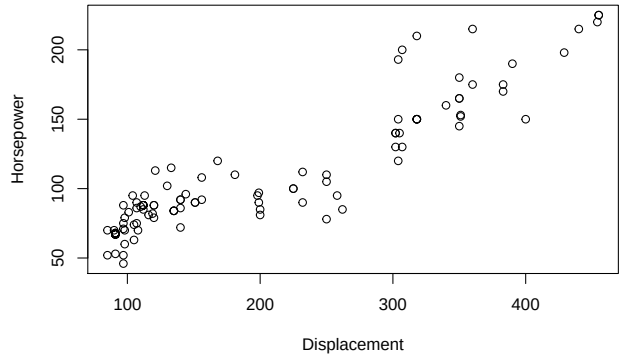
Tips & Tricks!

- La función attach() permite invocar las variables de una estructura de datos solo por su nombre.
- A las funciones en *R* se le pueden agregar atributos.
- El atributo na.rm=TRUE da la orden de que la función se ejecute sin tener en cuenta los datos No Accesibles.
- na significa No Accesible, rm significa *Remove* (quitar) y TRUE, verdadero. Esta última debe ir siempre en mayúsculas porque, de lo contrario, no la reconoce.

3.4.2. Diagrama de dispersión

Un diagrama de dispersión permite ver claramente si existe alguna relación entre las variables que estamos estudiando. Esta dispersión se puede obtener mediante la función plot(). Para ver un ejemplo, haremos un diagrama de dispersión de los caballos de potencia (*Horsepower* (y)) y el desplazamiento (*Displacement* (x)).

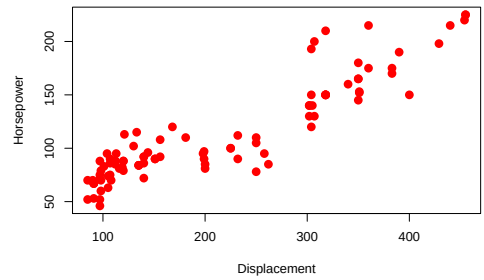
```
plot(Horsepower Displacement)
```



Tips & Tricks!

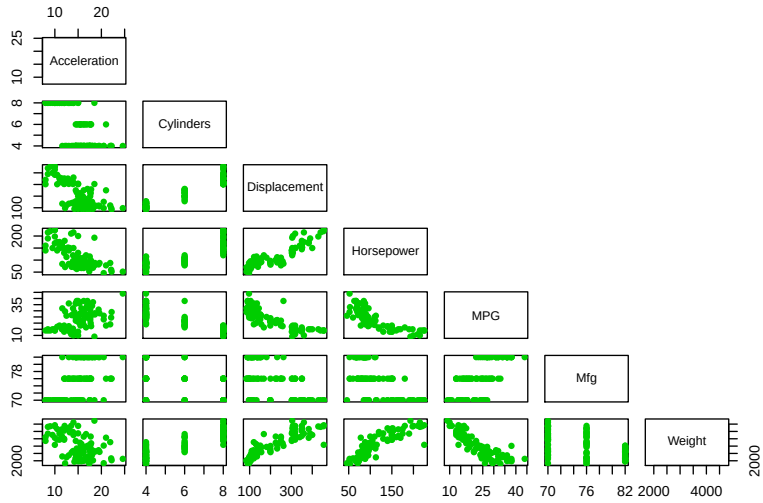
- Con la función plot(), se generan gráficas. Existen diferentes atributos que permiten mejorar su visualización, por ejemplo:
  - Asignando un valor a pch se escoge el símbolo para representar los puntos (entero entre 0 y 25, o símbolos comunes como: *,.,o,O,0,+,-,|, %,#).
  - lwd define el grosor de las líneas en la gráfica.
  - col asigna un color a los puntos (en este caso, para el color rojo, se puede poner red o 2)
  - cex determina el factor por el cual se multiplica el tamaño del punto original.
- plot(x,y) y plot(y∼x) generan la misma gráfica, x en el eje de las abscisas e y en el eje de las ordenadas del plano XY.

```
plot(Displacement,Horsepower,pch=16,col=red",cex=1.5)
```



Si se desea ver, de forma general, los diagramas de dispersión entre todas las combinaciones de las variables de una estructura de datos, se utiliza la función pairs().

```
pairs(Datos[3:9], upper.panel = NULL, pch = 16, col="green3")
```



Tips & Tricks!

Con la función pairs(), se crea una matriz gráfica de la correlación entre todas las variables numéricas del conjunto de datos

- El atributo upper.panel = NULL muestra la matriz inferior de las gráficas, evitando su duplicidad.
- También es posible crear la gráfica en color con la opción bg = c() listando los colores elegidos separados por comas y entre " ".

3.4.3. Modelo lineal de los mínimos cuadrados

El modelo de los mínimos cuadrados implementado en *R* consiste en aproximar una serie de valores con un polinomio del mínimo grado posible. Si los valores se asemejan a una recta, la función de la línea de regresión será de la forma $y = mx + b$, de primer grado. Este polinomio se consigue mediante la función lm(y $\sim$ x). Si el modelo se quiere utilizar después, por ejemplo para estimar algún valor, representar gráficamente los puntos y su recta, etc., este se debe asignar a un objeto con un nombre escogido por el usuario. Por ejemplo, queremos buscar la relación existente entre las variables Displacement y Horsepower.

```
model1=lm(Horsepower Displacement); model1
```

```
##
## Call:
## lm(formula = Horas ~ Lotes)
##
## Coefficients:
## (Intercept)        Lotes
##           7            5
```

Interpretando los resultados obtenidos, se deduce que la recta de regresión es $y = 34.8870 + 0.3706x$, donde x =Displacement e y =Horsepower. El objeto model1 guarda todos los parámetros del modelo. Por ejemplo, otra forma de conocer los valores de la pendiente y la intercepción de la recta de regresión es mediante las siguientes instrucciones:

```
model1$coef[1]
```

```
## (Intercept)
##    34.88703
```

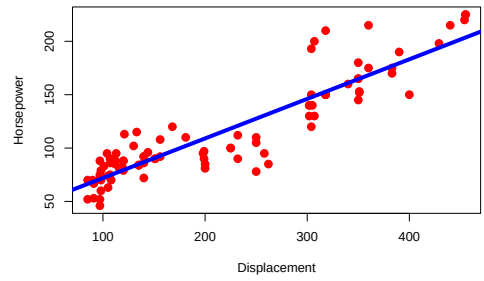
```
model1$coef[2]
```

```
## Displacement
##    0.3706237
```

3.4.4. Añadir la recta de regresión

Con la función abline(), se añade la recta de regresión creada con la función lm() al diagrama de dispersión. La primera entrada que hay que poner es el nombre de la variable que contiene el modelo; en esta ocasión, "model1", y los parámetros adicionales de visualización.

```
plot(Displacement,Horsepower,pch=16,col=red",cex=1.5)
abline(model1,col="blue",lwd=5)
```



3.4.5. Coeficientes de determinación (R^2) y de correlación (R)

Una vez calculado el modelo lineal, el coeficiente de determinación se visualiza utilizando la función summary(). Con este comando, se obtienen los parámetros más importantes del modelo. Sin embargo, para este coeficiente nos interesan solamente multiple R-squared.

```
summary(model1)
```

```
##
## Call:
## lm(formula = Horsepower ~ Displacement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.543 -11.425  -0.704   9.604  57.255
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.88703    3.96235   8.805 5.09e-14 ***
## Displacement  0.37062    0.01677  22.094  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.63 on 97 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.8325
## F-statistic: 488.2 on 1 and 97 DF,  p-value: < 2.2e-16
```

Por tanto, el coeficiente R cuadrado múltiple = 0.8342
El coeficiente de correlación de Pearson, que es la raíz cuadrada del coeficiente de determinación, también se puede calcular mediante la función cor().

```
sqrt(0.8342)
```

```
## [1] 0.9133455
```

```
cor(Horsepower,Displacement,use="na.or.complete")
```

```
## [1] 0.9133645
```

3.4.6. Estimación de valores indeterminados

Una de las aplicaciones principales de la regresión es la posibilidad de estimar el valor de la variable dependiente para un valor determinado de la variable independiente. Este recurso es muy útil ya que permite conocer aproximadamente el comportamiento de las dos variables en situaciones sin datos. Para realizar la estimación basta con reemplazar en la fórmula de la recta los valores de la pendiente, la intercepción y la variable dependiente. Por ejemplo, si se quiere estimar cuál será el valor faltante de los caballos de potencia (observación 77), sabiendo que ese coche tiene un valor de desplazamiento de 151, basta con ejecutar:

```
34.8870+0.3706*151
```

```
## [1] 90.8476
```

Sin embargo, si se quiere minimizar el error por redondeo y, además, estimar varios valores, lo mejor que se puede hacer es crear una función utilizando los valores directamente del modelo de la siguiente forma:

```
f1 <- function(x) {model1$coef[1] + model1$coef[2]*(x)}
```

Así, cada vez que se quiera hacer una estimación solo habrá que ejecutar la función creada.

```
Displacement_pred = 151
Horsepower_pred=  f1 (Displacement_pred);Horsepower_pred
```

```
## (Intercept)
##    90.85121
```

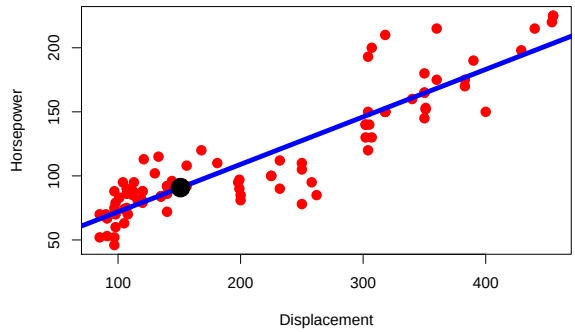
O utilizando la función predict()

```
Displacement_pred = 151
Horsepower_pred=predict(model1,data.frame(Displacement=Displacement_pred));
Horsepower_pred
```

```
##        1
## 90.85121
```

Finalmente, esta estimación se puede representar en la gráfica mediante un punto negro (por ejemplo), mediante la siguiente instrucción:

```
plot(Displacement,Horsepower,pch=16,col=red",cex=1.5)
abline(model1,col="blue",lwd=5)
points(Displacement_pred,Horsepower_pred, col ="black", pch=20, cex
=4)
```



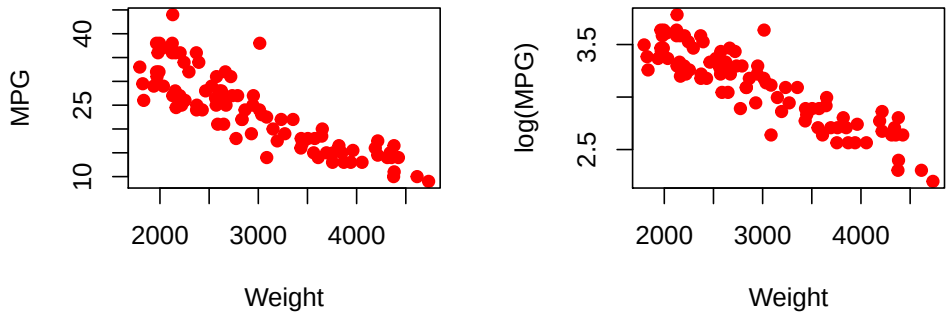
Tips & Tricks!

- lm(y∼x) crea un modelo lineal entre las variables x e y.
- abline() agrega una línea recta a la gráfica actual; los parámetros más usuales son:
  - a=A, b=B definen la intercepción (A) y la pendiente de la recta (B); también se puede poner el nombre de la variable que contiene un modelo lineal.
  - h=H para definir una línea horizontal en $y = H$
  - v=V para definir una línea vertical en $x = V$
- summary() muestra los resultados del modelo lineal.
- predict() genera la predicción para valores nuevos. Estos valores deben organizarse en un data.frame y la variable independiente debe tener el mismo nombre que se ha utilizado en la creación del modelo.
- points() agrega puntos a una gráfica existente.

3.4.7. Regresión exponencial

Como ya se ha visto, si dos variables tienen una relación exponencial, esta se puede linealizar por medio de logaritmos. Utilizando el fichero coches.txt, la relación entre las variables peso (Weight(y)) y consumo (MPG (x)) se puede observar mediante el diagrama de dispersión. Se puede intuir que la relación tiende a ser más parecida a una exponencial que a una lineal.

```
par(mfrow=c(1,2))
plot(Weight,MPG,xlab="Weight", ylab="MPG",col=red",pch=20,cex=1.5)
plot(Weight,log(MPG),xlab="Weight", ylab="log(MPG)",col=red",
pch=20,cex=1.5)
```




```
model2=lm(log(MPG) Weight) summary(model2)
```

```
##
## Call:
## lm(formula = log(MPG) ~ Weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41639 -0.13593  0.00205  0.11741  0.55351
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.300e+00  6.447e-02   66.69   <2e-16 ***
## Weight      -4.033e-04  2.101e-05  -19.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1624 on 92 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8002, Adjusted R-squared:  0.798
## F-statistic: 368.4 on 1 and 92 DF,  p-value: < 2.2e-16
```

De estos datos, obtenemos que $\ln(\text{MPG}) = 4.3 - 0.0004033 \times \text{Weight}$, o que $\text{MPG} = e^{4.3} \cdot e^{-0.0004033 \times \text{Weight}}$. Además, R cuadrado múltiple = 0.8002 y R cuadrado ajustado = 0.798.

Finalmente, queremos estimar los valores de la variable MPG que no han sido registrados (NaN), es decir, las observaciones: 11, 12, 13, 14, 15, y 18. Para ello, utilizamos la

ecuación de la regresión y los valores del peso de las observaciones: 11, 12, 13, 14, 15, y 18.

```
obs = which(MPG=="NaN"); obs
```

```
## [1] 11 12 13 14 15 18
```

```
Weight_pred = Weight[obs]
Log_MPG_pred = predict(model2,data.frame(Weight=Weight_pred));
Log_MPG_pred
```

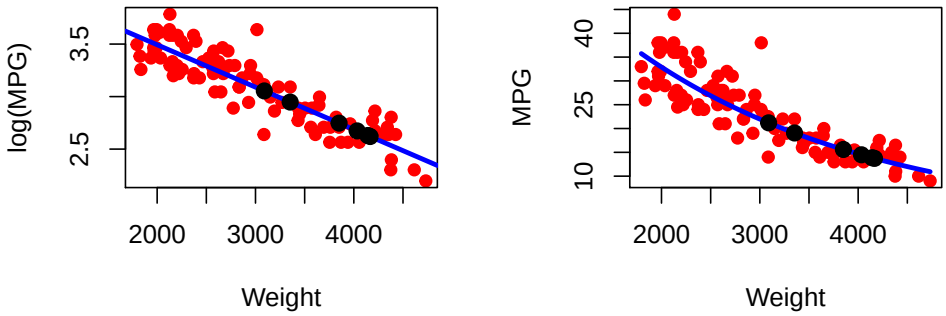
```
##        1        2        3        4        5        6
## 3.053833 2.629596 2.673149 2.619917 2.747350 2.947774
```

```
MPG_pred = exp(Log_MPG_pred); MPG_pred
```

```
##        1        2        3        4        5        6
## 21.19644 13.86816 14.48551 13.73459 15.60123 19.06347
```

Representando gráficamente los datos existentes, el resultado de la regresión y la estimación de los valores no registrados, se tiene:

```
par(mfrow=c(1,2))
plot(Weight,log(MPG),xlab="Weight", ylab="log(MPG)",
     col=red",pch=20,cex=1.5)
abline(model2,col="blue",lwd=3)
points(Weight_pred,Log_MPG_pred, col = "black", pch= 20, cex = 2)
plot(Weight,MPG,xlab="Weight", ylab="MPG",col=red",pch=20,cex=1.5)
curve(exp(4.3)*exp(-0.0004033*x),add=T,col="blue",lwd=3)
points(Weight_pred,MPG_pred, col = "black", pch= 20, cex = 2)
```

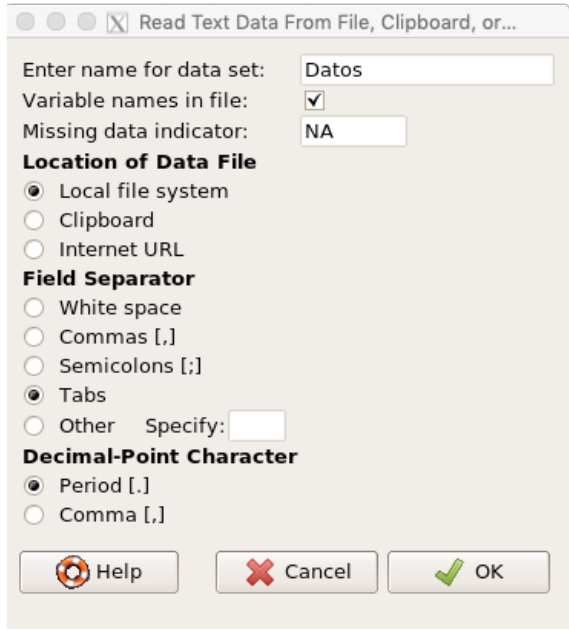


3.5. Regresión lineal con R-Commander

De igual forma que en el apartado anterior, se utiliza el fichero coches.txt disponible en Atenea.

3.5.1. Cargar datos

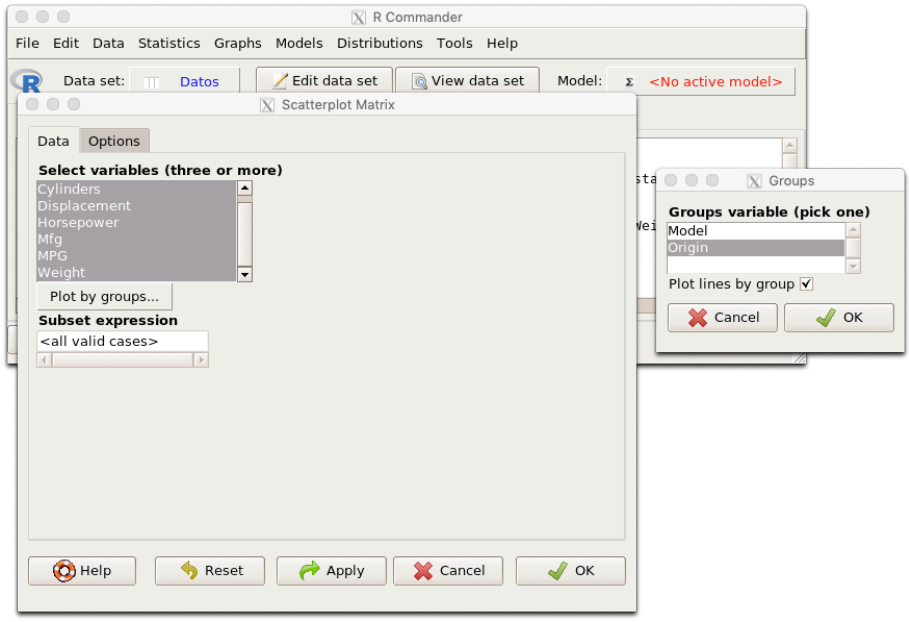
Se cargan los datos del fichero con las siguientes características:



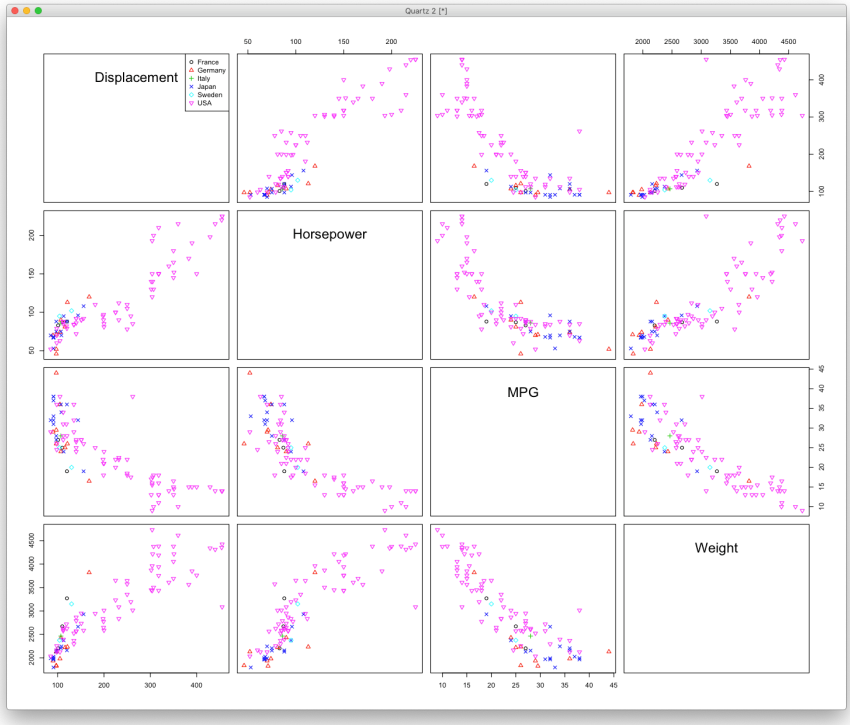
Se recomienda visualizar los datos para verificar que los datos se han importado correctamente.

3.5.2. Diagramas de dispersión

Para visualizar las relaciones existentes entre diferentes variables en forma de una matriz de gráficos, se selecciona Gráficos >Matriz de diagramas de dispersión en la barra de menú. Aparece la siguiente ventana, donde se pueden elegir las variables a graficar y, si se quiere, discriminarlas por grupos de acuerdo con las variables cualitativas existentes. También se pueden cambiar varias opciones de visualización por medio de la pestaña Opciones. Por ejemplo, haremos una matriz con las variables: Displacement, Horsepower, MPG, Weight, pero agrupadas confrome al origen.

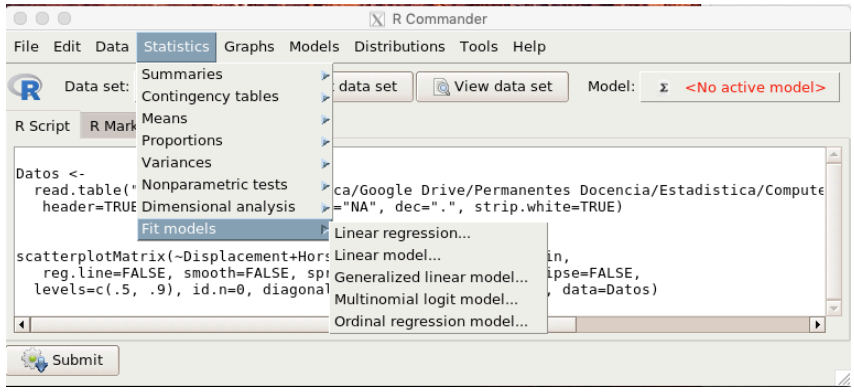


La figura resultante es:

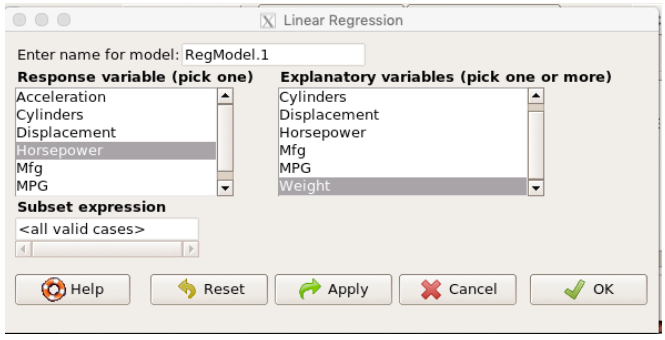


3.5.3. Modelo lineal de los mínimos cuadrados

Ahora buscaremos la recta de regresión entre los caballos de potencia (Horsepower (y)) y el peso (Weight (x)) para poder estimar el valor faltante de la variable Horsepower (observación 77) utilizando el valor del peso de dicha observación. Se selecciona Estadística >Ajuste de modelos >Regresión lineal en la barra de menú.



Se define el nombre del modelo, se selecciona Horsepower como variable independiente y Weight como variable dependiente.



Al aceptar, se puede observar que en la ventana principal, debajo del menú, la casilla Model ahora aparece activada con el nombre que le hemos dado al crear el modelo lineal, en nuestro caso RegModel.1. Por otra parte, en la pestaña de "R Script" aparece la línea de instrucciones que se pueden ejecutar desde la **R-Console** o **Rstudio**.

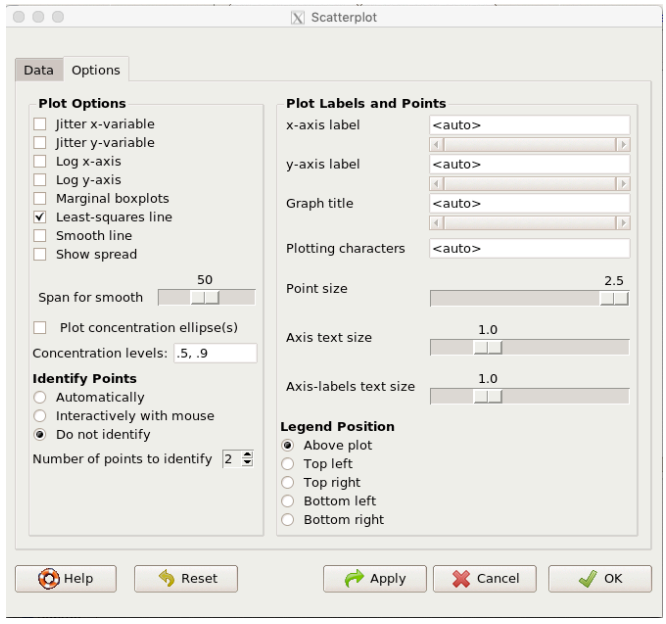
```
RegModel.1 <- lm(Horsepower Weight, data=Datos) summary(RegModel.1)
```

```
##
## Call:
## lm(formula = Horsepower ~ Weight, data = Datos)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.610 -12.510   0.872   9.627 109.312
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.592850   8.656252  -4.112 8.22e-05 ***
## Weight        0.049022   0.002776  17.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.29 on 97 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.7603
## F-statistic: 311.8 on 1 and 97 DF,  p-value: < 2.2e-16
```

3.5.4. Añadir la recta de regresión

Para añadir la recta de regresión, solo hay que hacer la gráfica de dispersión entre las variables y seleccionar, dentro de la pestaña Opciones, la recta de regresión. También se pueden modificar otras opciones de visualización.

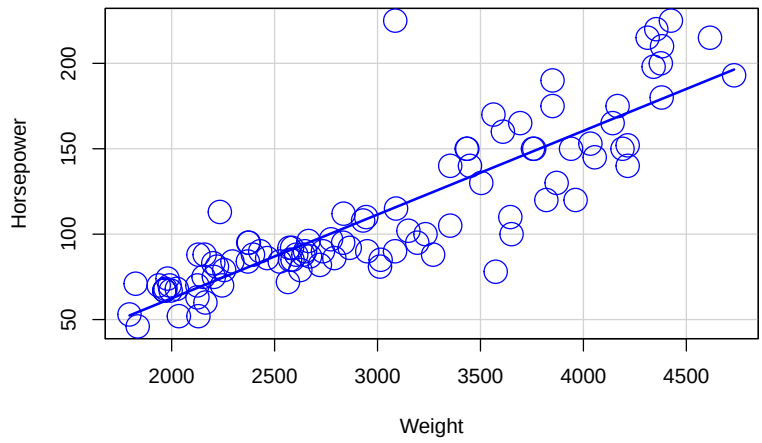


El resultado en la ventana de salida y en la figura es el siguiente:

```
library(car)
```

```
## Loading required package: carData
```

```
scatterplot(Horsepower Weight, reg.line=lm, smooth=FALSE,
spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
levels=c(.5, .9), cex=2.5, data=Datos)
```



3.5.5. Estimación de valores indeterminados

Finalmente, la estimación de los valores indeterminados se hace de la misma forma que en **R-Console** o **Rstudio**.

```
obs = which(Horsepower=="NaN"); obs
```

```
## [1] 77
```

```
Weight_pred = Weight[obs]
Horsepower_pred = predict(RegModel.1,data.frame(Weight=Weight_pred));
Horsepower_pred
```

```
##        1
## 113.1877
```

3.6. Ejercicios propuestos

1. Tenemos las siguientes alturas (en cm) de un conjunto de personas:

   78, 181, 168, 183, 164, 181, 174, 176, 174, 176, 181, 168, 164, 174, 171

   A este mismo conjunto de personas, se las pesa y se obtienen los siguientes pesos:

   82, 89, 68, 91, 65, 80, 79, 81, 80, 79, 82, 69, 67, 80, 78

   Haz un estudio de regresión lineal, calculando la recta de regresión y el coeficiente de determinación, y traza el diagrama de dispersión. ¿Qué puedes deducir a partir del valor de **R**?

2. Utilizando los datos del fichero reg.txt, analiza si es adecuado realizar modelos de regresión lineal entre las variables x1-y1, x2-y2, x3-y3, x4-y4. Para cada par de variables, puedes seguir la siguiente metodología:

   – Diagrama de dispersión. ¿Te parece adecuado un modelo lineal para describir cada conjunto de datos?
   – Ajuste de un modelo lineal. Comenta los resultados.
   – En caso de que un modelo lineal no sea adecuado, ¿qué se podría hacer para ajustar un modelo que pueda predecir la variable y en función de la variable x?

3. Se quiere estudiar la resistencia de unas piezas de cemento en función de su edad en días. Utilizando los datos que se encuentran en el fichero cemento.txt, propón un modelo que relacione la resistencia con el tiempo de secado. ¿Qué resistencia tendrán a los 5 días? ¿Y a los 50? Utiliza el coeficiente de determinación para justificar dichos valores.

4. Queremos estudiar la evolución del nivel máximo anual del mar (en cm) en Venecia. Los datos de que disponemos corresponden a los años 1931-1981 y están contenidos en el fichero venecia.txt (datos reales, publicados en Smith R.L, "Extreme value theory based on the r largest annual events", *Journal of Hydrology*, 86 (1986)). Realiza un estudio de regresión y comenta la evolución del máximo anual del nivel del mar en Venecia.

5. Se quiere estudiar la evolución de la producción mundial de petróleo de 1880 a 1973. Los datos se encuentran en el fichero petroleo.txt.

6. La hidrólisis de un cierto éster tiene lugar en un medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial de 30 mM del éster, se han medido concentraciones a diferentes tiempos y se han obtenido los resultados registrados en el fichero ester.txt. ¿Cuál estimas que ha sido la concentración a los 70 segundos de comenzar el proceso? Explica los resultados.

7. Estudiando los incendios forestales, deducimos que puede existir una relación entre la cantidad de lluvia caída durante los meses de verano (en mm) y el número de incendios declarados. Recopilamos la información de los últimos diez años y obtenemos los datos siguientes:

Lluvia	Incendios
97	521
27	863
93	712
175	163
38	138
192	811
28	534
182	442
61	963
77	313

A la vista de estos datos, ¿podríamos hacer una previsión de cuál será el número aproximado de fuegos que se declararán con una lluvia de 120 mm? ¿y de 10 mm? Explica las conclusiones.


2
1
3

Variables aleatorias discretas y distribuciones de probabilidad

4.1. Introducción y objetivos

Supongamos que, en una inspección de fabricación, se tiene un lote de 100 piezas y una de ellas está contaminada. Si se inspecciona una pieza de ese lote, ¿cuál es la probabilidad de que esta pieza esté contaminada? Si se inspeccionan dos, ¿cuál es la probabilidad de que ninguna esté contaminada? ¿Cuál es la probabilidad de que se tengan que inspeccionar cinco piezas hasta encontrar la contaminada? El hecho de inspeccionar una pieza, dos o varias hasta encontrar la contaminada se denomina *experimento aleatorio*. El resultado del experimento se asigna a una variable. Si los posibles resultados no pueden tomar valores dentro de un mínimo conjunto numerable, la variable se denomina *variable aleatoria discreta*. Si un experimento aleatorio se repite varias veces, se espera que durante todo el tiempo los resultados estén condicionados por sus probabilidades. Si estas probabilidades siguen un comportamiento específico, el experimento se puede clasificar dentro de ciertos modelos de distribución.

Esta sesión se centra en conocer cómo se representan las probabilidades de un experimento aleatorio de variable aleatoria discreta, simular repeticiones de un experimento y comparar los resultados con las probabilidades asignadas y, finalmente, describir los modelos más frecuentes utilizando los mismos términos empleados para describir los datos recogidos (esperanza y varianza). Al finalizar esta sesión, el alumno ha de ser capaz de:

- Representar gráficamente una distribución de variable aleatoria discreta usando **R**.
- Simular la repetición de diferentes experimentos aleatorios discretos y comparar el resultado de estos experimentos con las probabilidades estudiadas previamente.
- Calcular e interpretar el valor esperado y la varianza de una variable aleatoria discreta.
- Reconocer y aplicar correctamente las distribuciones de probabilidad discretas más comunes en ingeniería.

4.2. Variables aleatorias discretas (VAD)

Normalmente, expresamos el resultado de un experimento aleatorio con un simple número, pero no siempre es posible, como en los casos de lanzar una moneda, atrapar un balón, etc. En estos casos, no es razonable sugerir hacer un análisis cuantitativo. Así pues, es necesario asignar un número real a cada uno de los sucesos elementales del espacio muestral. Una variable aleatoria X se puede definir como la función que transforma los resultados del espacio muestral Ω en puntos sobre la recta real $\mathbb{R}$. En otras palabras, es una función cuyo dominio es Ω y rango $\mathbb{R}$.

$$\mathbf{X} : \Omega \to \mathbb{R}$$

Una variable aleatoria es una variable aleatoria discreta (VAD o DRV, por sus iniciales en inglés) si el conjunto de sus posibles resultados es contable. Esto es, si el espacio muestral contiene un número finito de posibilidades o una secuencia inacabada con tantos elementos como números enteros hay. Dicho con más rigor, se determina una VAD como la variable que hay entre dos valores observables, en que hay por lo menos un valor no observable.

Pero una variable aleatoria cuyo conjunto de posibles valores es un intervalo completo de números es no discreta, es decir, si el espacio muestral contiene un número infinito de posibilidades igual al número de puntos en un segmento de línea, se denomina *variable aleatoria continua*.

En la mayoría de problemas prácticos, las variables aleatorias continuas representan los datos medidos, tales como los pesos, las temperaturas, la distancia o los períodos de vida, mientras que las variables aleatorias discretas representan datos contables, como el número de elementos defectuosos en un muestreo, los objetos o el número de accidentes mortales al año en un cierto lugar de la autopista.

Ejemplo: Si lanzamos 3 monedas, o una misma moneda 3 veces (experimento aleatorio), podemos recoger el resultado como un trío, por ejemplo, si C es "cara" y + es "cruz", el conjunto

$$\Omega = \{(CCC), (CC+), (C+C), (+CC), (++C), (+C+), (C++), (+++)\}$$

contiene todos los posibles resultados. Este espacio muestral no está representado por números. Sin embargo, se pueden definir diferentes variables aleatorias dependiendo de lo que se quiera analizar. Por ejemplo, las variables aleatorias X_1, X_2 y X_3 se pueden definir tal como se ve en la tabla siguiente, donde el valor de X_1 indica el número de "caras", X_2 toma como valor 1 si se ha obtenido exactamente una "cara" y, X_3 toma como valor 1 si el último lanzamiento es "cara" y 0 el resto de casos.

Ω	CCC	$CC+$	$C+C$	$+CC$	$++C$	$+C+$	$C++$	$+++$
X_1	3	2	2	2	1	1	1	0
X_2	0	0	0	0	1	1	1	0
X_3	1	0	1	1	1	0	0	0

4.2.1. Función de densidad

La función de densidad (también conocida como distribución de probabilidad, función de probabilidad o función de masa de probabilidad), asociada a una variable aleatoria discreta X definida sobre el espacio muestral Ω, es la aplicación f que asigna a cada elemento x_i de $X(\Omega)$ la probabilidad de que la variable X tome este valor x_i:

$$f : X(\Omega) \to \mathbb{R}$$

$$x_i \to f(x_i)$$

Esta función tiene las siguientes propiedades:

- Su valor siempre es un número real positivo entre cero y uno: $0 \leq f(x) \leq 1$.
- La probabilidad de que X tome un valor exacto esta definida por su función de densidad:
$$P(X = x) = f(x).$$
- La suma total de sus valores es 1: $\sum_{x_i} f(x_i) = 1$.
- La probabilidad de que X tome un valor en el intervalo $[a, b]$ es la suma de las probabilidades en ese intervalo: $P(a \leq X \leq b) = \sum_{x_i=a}^{b} f(x_i)$.

Ejemplo 1: Consideremos la variable aleatoria discreta X_1 que representa el "Resultado obtenido" al lanzar un dado normal de seis caras. Así, X_1 puede tomar los siguientes valores: 1,2,3,4,5,6. De acuerdo con la teoría de la probabilidad, la probabilidad de que X_1 sea 1 viene dada por:

$$P(X_1 = 1) = f_1(1) = \frac{N^{\text{o}} \text{ eventos favorables}}{N^{\text{o}} \text{ eventos totales}} = \frac{1}{6}$$

Del mismo modo, $f_1(2) = f_1(3) = f_1(4) = f_1(5) = f_1(6) = \frac{1}{6}$. Por tanto, la función de densidad se puede representar por medio de la siguiente tabla:

X_1	1	2	3	4	5	6
$f_1(x)$	1/6	1/6	1/6	1/6	1/6	1/6

Ejemplo 2: Ahora, consideramos un dado trucado en que la probabilidad de que salga un número es proporcional al valor del mismo. Esto significa que la probabilidad de obtener un 1 es 1α, de obtener un 2 es 2α, para 3 es 3α, y lo mismo con los demás. Consideramos la variable aleatoria discreta X_2 que representa el "Resultado obtenido" al lanzar un dado trucado. Partiendo de que $\Sigma p_k = 1$, podemos calcular $\alpha = 1/21$; por tanto, $f_2(1) =$

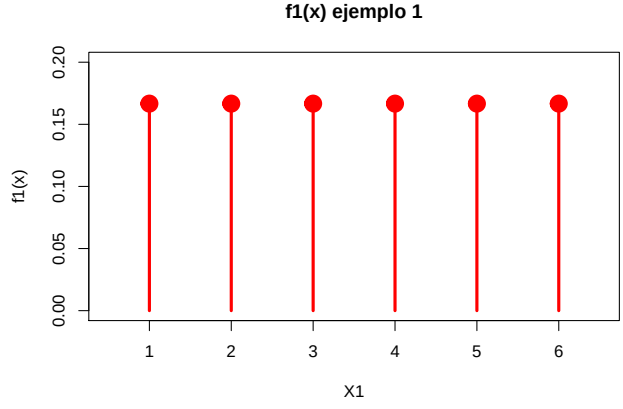
$1/21, f_2(2) = 2/21, f_2(3) = 3/21, f_2(4) = 4/21, f_2(5) = 5/21, f_2(6) = 6/21$. De esta manera, su función de densidad es:

X_2	1	2	3	4	5	6
$f_2(x)$	1/21	2/21	3/21	4/21	5/21	6/21

La función de densidad se representa gráficamente mediante líneas verticales. Estas líneas se extienden desde la línea de base a lo largo del eje x, donde se representan los posibles valores de la variable aleatoria $(X = k)$. La probabilidad de ocurrencia de cada valor $f(x) = P(X = k)$ se indica mediante puntos gruesos al final de la línea. La función de densidad del ejemplo 1, el dado normal, se representa gráficamente de la siguiente manera:

```
rm(list=ls())  Elimina todos los objetos del espacio de trabajo
graphics.off()  Elimina las gráficas creadas con anterioridad
```

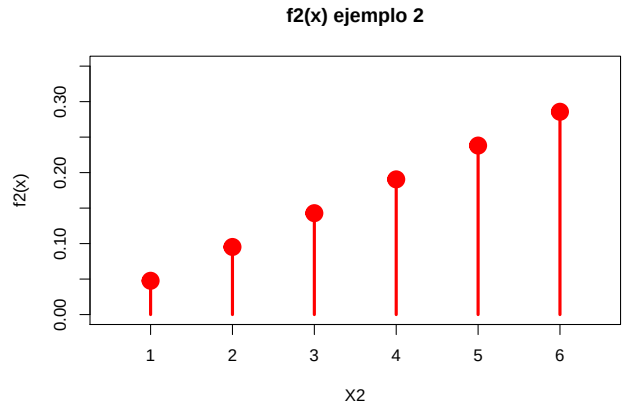
```
x1 = 1:6  Posibles resultados
f1 = rep(1/6,6)  Probabilidad de ocurrencia de cada resultado
Creación de las líneas verticales
plot(x1, f1, type="h", col=red", lwd=3, main="f1(x) ejemplo 1",
xlab="X1", ylab="f1(x)", xlim=c(0.5,6.5), ylim=c(0,0.20))
Se crean los puntos y se guarda la gráfica completa en un objeto
para su uso posterior
points(x1, f1, col=red", lwd=10); gra.fx.ej1 = recordPlot()
```



Si consideramos el ejemplo 2, el dado trucado, su función de densidad es:

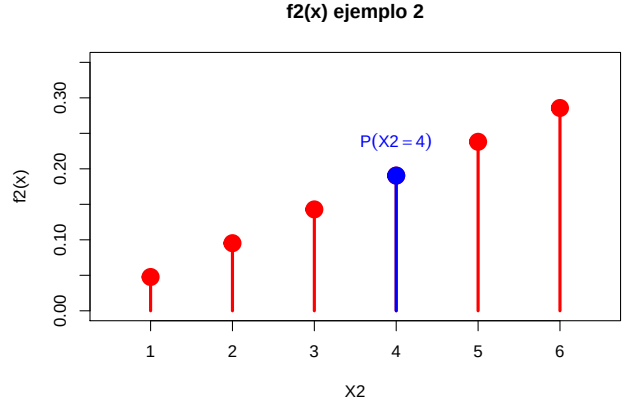
```
x2 = 1:6  Posibles resultados
f2 = x2/21  Probabilidad de ocurrencia de cada resultado  Creación de
las líneas verticales
plot(x2, f2, type="h", col=red", lwd=3, main="f2(x) ejemplo 2",
```

```
xlab="X2", ylab="f2(x)", xlim=c(0.5,6.5), ylim=c(0,0.35))
Se crean los puntos y se guarda la gráfica completa en un objeto
para su uso posterior
points(x2, f2, col=red", lwd=10); gra.fx.ej2 = recordPlot()
```



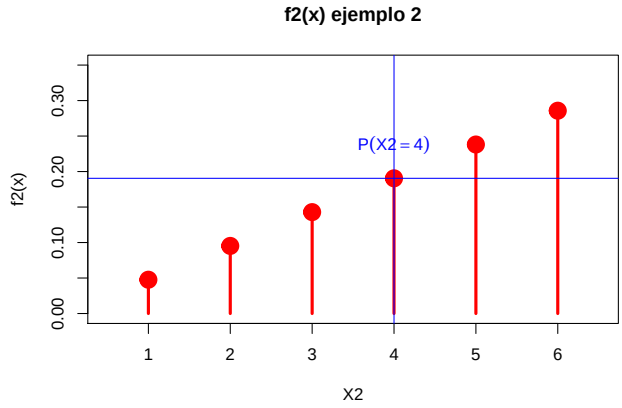
Finalmente, si se quiere resaltar en la gráfica una probabilidad específica, por ejemplo $P(X = 4)$, se puede optar por cambiar el color de la línea y el punto en $X = 4$:

```
gra.fx.ej2  Se recupera la gráfica creada anteriormente
lines(4, f2[4], type="h", col="blue", lwd=3)
points(4, f2[4], col="blue", lwd=10)
text(4, f2[4]+0.02, expression(P(X2 == 4)), pos=3, col="blue")
```



o agregando una línea vertical en $x = 4$ y una horizontal en $f_2(4)$con la función abline()

```
gra.fx.ej2  Se recupera la gráfica creada anteriormente
abline(v=4, col="blue")
abline(h=f2[4], col="blue")
text(4, f2[4]+0.02, expression(P(X2 == 4)), pos=3, col="blue")
```



4.2.2. Función de distribución

La función de distribución o función de probabilidad acumulada asociada a una variable aleatoria discreta X, definida sobre un espacio muestral, es la aplicación F que asigna a cada elemento x_i de $X\,(\Omega)$ la probabilidad de que la variable X tome cualquier valor menor o igual que x_i:

$$F : X(\Omega) \to \mathbb{R}$$

$$x_i \to F(x_i) = P(X \leq x_i) = \sum_{x_i \leq x} P(X = x_i)$$

Por tanto:

$$P(a < X \leq b) = F(b) - F(a), \qquad \forall a \leq b$$

De esta forma, la función de densidad y la función de distribución del ejemplo 1 son:

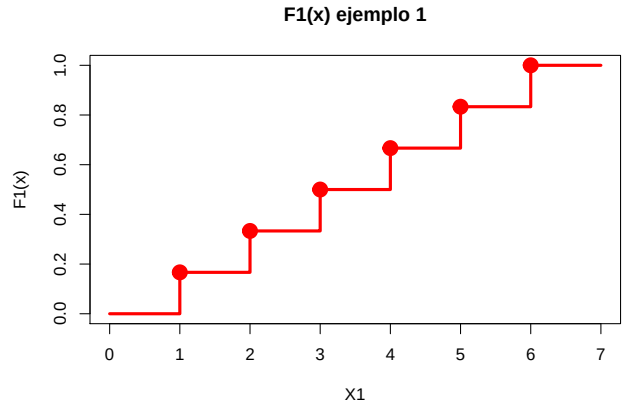
X_1	1	2	3	4	5	6
$f_1(x)$	1/6	1/6	1/6	1/6	1/6	1/6
$F_1(x)$	1/6	2/6	3/6	4/6	5/6	1

Y, para el ejemplo 2, son:

X_2	1	2	3	4	5	6
$f_2(x)$	1/21	2/21	3/21	4/21	5/21	6/21
$F_2(x)$	1/21	3/21	6/21	10/21	15/21	1

La función de distribución se representa gráficamente mediante líneas horizontales escalonadas, que indican la probabilidad $F(x) = P(X \leq x)$, donde x puede ser cualquier valor real. Los puntos gruesos al inicio de cada línea horizontal indican el valor que toma la función para los valores de la variable aleatoria X. La función de distribución del ejemplo 1, el dado normal, se representa gráficamente de la siguiente manera:

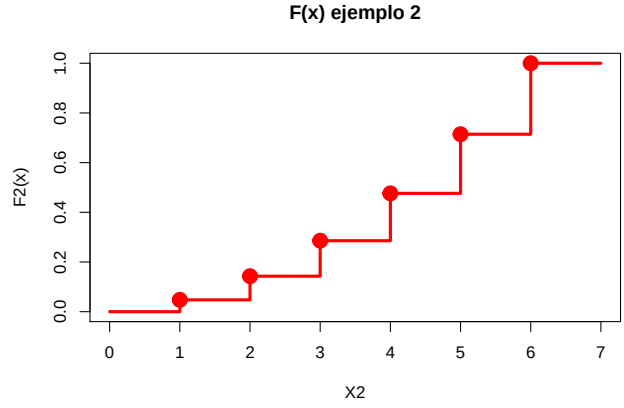
```
F1 = cumsum(f1)  Se genera un vector con la suma acumulada
plot(c(0,x1,7), c(0,F1,1), type="s", col=red", lwd=3,
main="F1(x) ejemplo 1", xlab="X1",  ylab="F1(x)")
points(x1, F1, col=red", lwd=8); gra.Fx.ej1 = recordPlot()
```



Nótese que se han agregado los elementos 0 y 7 en el eje x y sus imágenes $F_1(0) = 0$ y $F_1(7) = 1$ para completar la gráfica.

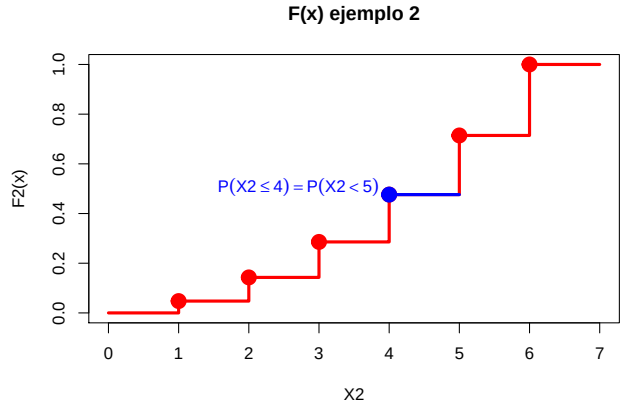
Si consideramos el ejemplo 2, el dado trucado, su función de densidad es:

```
F2 = cumsum(f2)  Se genera un vector con la suma acumulada
plot(c(0,x2,7), c(0,F2,1), type="s", col=red", lwd=3,
main="F(x) ejemplo 2", xlab="X2", ylab="F2(x)")
points(x2, F2, col=red", lwd=8); gra.Fx.ej2 = recordPlot()
```



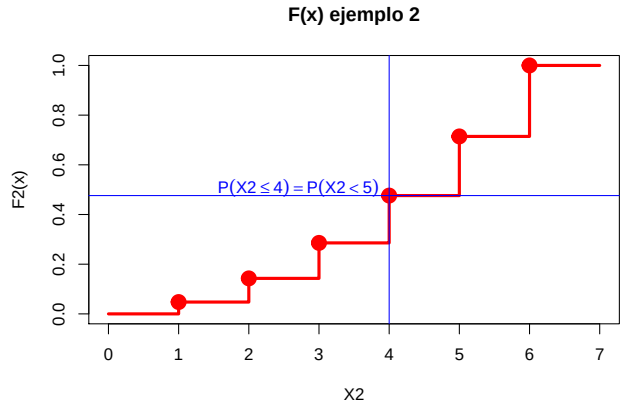
Finalmente, si se quiere resaltar en la gráfica una probabilidad específica, por ejemplo $P(X \leq 4)$, se puede optar por cambiar el color de la línea y el punto en $X = 4$

```
gra.Fx.ej2
lines(c(4,5), rep(F2[4],2), type="s", col="blue", lwd=3)
points(4, F2[4], col="blue", lwd=8)
text(4, F2[4]+0.02,
expression(P(X2 <= 4) == P(X2<5)), pos=2, col="blue")
```



o agregando una línea vertical en $x = 4$ y una horizontal en $F(4)$.

```
gra.Fx.ej2
abline(v=4, col="blue")
abline(h=F2[4], col="blue")
text(4, F2[4]+0.02,
expression(P(X2 <= 4) == P(X2<5)), pos=2, col="blue")
```



Tips & Tricks!

- rm(list = ls()) elimina todos los objetos (variables, datos, funciones) cargados en el espacio de trabajo.
- graphics.off() elimina o limpia la ventana de gráficas.
- plot() representa gráficamente los datos especificados eliminando las gráficas que previamente se hayan realizado. Las opciones type, col, lwd, xlim, ylim, main, xlab e ylab configuran el tipo de gráfica, el color de las líneas, el grosor de la línea, los límites, el título y las etiquetas de los ejes x e y, respectivamente. Véase la ayuda de **R** para conocer más detalles.
- El parámetro type= especifica el tipo de gráfica deseada: p para puntos, l para líneas, b para puntos y lineas, c para puntos vacíos unidos por líneas, o para puntos y líneas superpuestos, s y S para escalones y h para líneas verticales tipo histograma. Finalmente, n no produce puntos ni líneas.
- points() agrega puntos a la gráfica previamente realizada.
- lines() agrega líneas a la gráfica previamente realizada. lines(x,y,type=h) realiza la misma gráfica que plot(x,y,type=h) con la diferencia de que, con el primero, la gráfica que esté previamente realizada no se elimina. Esta propiedad es útil si se quieren superponer dos tipos de gráficas diferentes, por ejemplo con histograma.
- abline() agrega una línea recta a la gráfica activa; esta línea puede ser horizontal en un valor dado, h=valor, o vertical, v=valor
- text(x,y,expression()) ubica un texto dado en la posición (x, y) de una gráfica previamente realizada.
- recordPlot() permite guardar una gráfica en un objeto de **R** para su uso posterior.

4.2.3. Medidas características de las VAD

Estos importantes parámetros o características cuantifican la tendencia central y la variabilidad o dispersión de la variable aleatoria discreta. De hecho, conocer estas cantidades, dejando aparte la distribución completa, puede darnos una idea de la naturaleza del sistema.

Valor esperado

Dese el punto de vista de la *frecuencial de probabilidad*, representa la cantidad promedio que "esperamos' como resultado final de un experimento aleatorio repetido muchas veces. El valor esperado $E(X)$ es justamente la media ponderada de una variable aleatoria discreta X:

$$\mu = E[X] = \sum_{i=1}^{k} x_i f(x_i)$$

De esta forma, el valor esperado en el ejemplo 1 es:

$$\mu_{X_1} = E[X_1] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Y, en el ejemplo 2 es:

$$\mu_{X_2} = E[X_2] = 1 \cdot \frac{1}{21} + 2 \cdot \frac{2}{21} + 3 \cdot \frac{3}{21} + 4 \cdot \frac{4}{21} + 5 \cdot \frac{5}{21} + 6 \cdot \frac{6}{21} = 4.33$$

Varianza

Este parámetro mide la dispersión o *scatter* de los posibles valores de X. La varianza es el promedio (esperado) de la distancia al cuadrado (o desviación) de la media:

$$\sigma^2 = V(X) = Var(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 =$$
$$= \sum_{i=1}^{k} (x_i - \mu)^2 f(x_i) = \sum_{i=1}^{k} x_i^2 f(x_i) - \mu^2$$

Por tanto, la varianza del ejemplo 1 es:

$$\sigma^2_{X_1} = V(X_1) = (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + (3-3.5)^2 \cdot \frac{1}{6} + (4-3.5)^2 \cdot \frac{1}{6} +$$
$$+ (5-3.5)^2 \cdot \frac{1}{6} + (6-3.5)^2 \cdot \frac{1}{6} = 2.91667$$

Y la del ejemplo 2 es:

$$\sigma^2_{X_2} = V(X_2) = (1-4.33)^2 \cdot \frac{1}{21} + (2-4.33)^2 \cdot \frac{2}{21} + (3-4.33)^2 \cdot \frac{3}{21} + \ldots$$
$$+ (6-4.33)^2 \cdot \frac{6}{21} = 2.222$$

Todos estos resultados se pueden validar o verificar si podemos repetir el experimento muchas veces. Por ejemplo, si lanzamos el dado n veces, obtenemos n_1 veces el número 1, n_2 veces el número 2, n_3 veces el número 3, y así sucesivamente. La frecuencia relativa de obtener un valor i viene dada por $f_i = n_i/n$. Si n es muy grande, seguramente la frecuencia será $f_i = f\,(i)$, aproximadamente. Por otro lado, la media y la varianza de las repeticiones estarán cercanas a μ y σ^2, respectivamente.

4.2.4. Uso de sample() para generar simulaciones

Realizar un experimento aleatorio conociendo su espacio muestral se puede considerar como hacer una selección aleatoria (muestreo) del conjunto del espacio muestral. Teniendo en cuenta lo anterior, con **R** se pueden simular observaciones o experimentos de una VAD por medio de la función sample().

Si el vector de datos x contiene los elementos del conjunto al cual estamos haciendo el muestreo (espacio muestral), entonces la instrucción sample(x) reorganiza el contenido de x en una secuencia aleatoria mientras mantiene todos los valores numéricos intactos. En otras palabras, se puede considerar que se hacen tantas simulaciones del experimento como elementos en el espacio muestral.

El tamaño de la muestra (o número de simulaciones) puede especificarse agregando el atributo size =. Por defecto, el muestreo se hace sin reemplazo; sin embargo, ello se puede cambiar con el atributo replace =. Existen opciones más avanzadas a la hora de hacer un muestreo, como por ejemplo especificar la probabilidad de cada elemento para ser seleccionado; por defecto, cada valor del espacio muestral es igualmente probable. De esta manera, por medio de la instrucción sample(x, size=n, replace=T, prob=p), se seleccionan n elementos del vector x (con reemplazo) cuyas probabilidades se especifican en el vector p.

Considerando el ejemplo 1 del lanzamiento de un dado normal de 6 caras, un resultado puede simularse como:

```
x = 1:6 sample(x, size=1, prob=)
```

```
## [1] 1
```

Cada vez que se realiza el experimento, es decir, cada vez que se ejecuta la instrucción sample(), el resultado es diferente, debido a la aleatoriedad.

```
sample(x, size=1, prob=)
```

```
## [1] 3
```

Para simular que lanzamos el dado 100 veces:

```
fair_die = sample(x, 100, replace=T, prob=); fair_die
```

```
##   [1] 4 1 3 5 4 5 5 1 2 1 6 6 4 1 3 5 1 3 5 5 2 3 3 1 4 5 5 1 3 2
##       6 4 1 6 5 1 4
##  [38] 3 6 2 2 4 4 3 1 3 3 3 1 5 3 3 6 5 5 5 5 5 2 5 1 1 1 5 5 1 5
##       1 2 3 1 4 2 4
##  [75] 2 4 6 6 3 1 4 3 2 1 1 3 5 6 6 4 1 2 4 6 6 5 6 5 6 1
```

Como ya se ha dicho, los resultados obtenidos son diferentes cada vez que se ejecuta la instrucción. Esto es así porque se han generado números pseudoaleatorios a partir de un número dado denominado *semilla* (*seed*). Si esta semilla no se establece a priori, **R** utiliza el reloj del sistema, de ahí la aleatoriedad del proceso. Si se desea reproducir los mismos resultados en la simulación, se puede asignar el valor de la semilla mediante la instrucción set.seed(k) justo antes de la simulación, donde k es un número real. Por ejemplo:

```
set.seed(1) fair_die = sample(x, 100, replace=T, prob=); fair_die
```

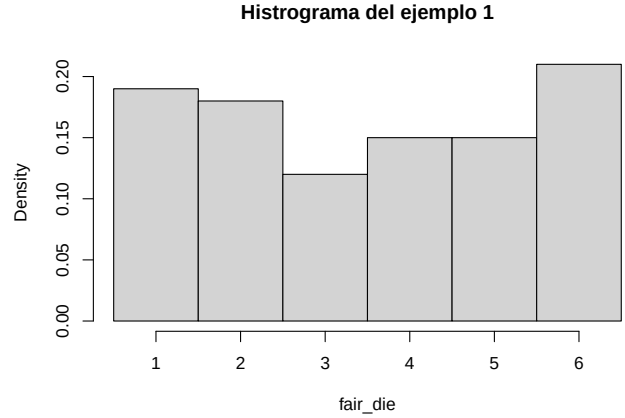
```
##   [1] 1 4 1 2 5 3 6 2 3 3 1 5 5 2 6 6 2 1 5 5 1 1 6 5 5 2 2 6 1
        4 1 4 3 6 2 2 6
##  [38] 4 4 4 2 4 1 6 1 4 1 6 2 3 2 6 6 2 5 2 6 6 6 1 3 3 6 4 6 3
        1 4 5 1 1 6 4 5
##  [75] 5 4 6 5 4 4 1 5 5 6 1 1 3 6 2 2 3 6 2 4 3 5 2 2 1 3
```

```
set.seed(1) fair_die = sample(x, 100, replace=T, prob=); fair_die
```

```
##   [1] 1 4 1 2 5 3 6 2 3 3 1 5 5 2 6 6 2 1 5 5 1 1 6 5 5 2 2 6 1
        4 1 4 3 6 2 2 6
##  [38] 4 4 4 2 4 1 6 1 4 1 6 2 3 2 6 6 2 5 2 6 6 6 1 3 3 6 4 6 3
        1 4 5 1 1 6 4 5
##  [75] 5 4 6 5 4 4 1 5 5 6 1 1 3 6 2 2 3 6 2 4 3 5 2 2 1 3
```

Una forma de resumir los resultados obtenidos de la simulación o el muestreo es por medio de la tabla de frecuencia relativa y su representación gráfica por medio del histograma (o diagrama de barras).

```
h_fair_die = hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histrograma del ejemplo 1")
```

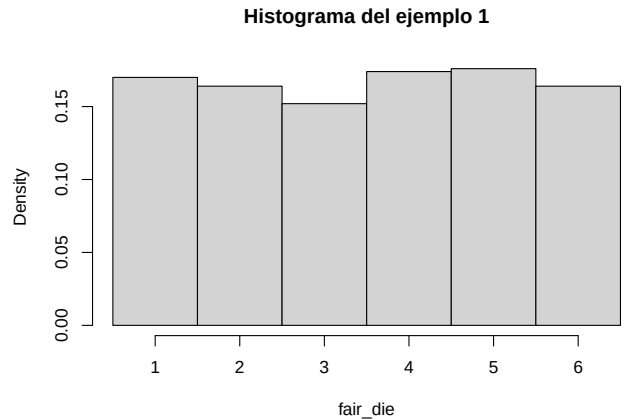


```
data.frame(Resultado=h_fair_die$mids, Frecuencia=h_fair_die$density)
```

```
##   Resultado Frecuencia
## 1         1       0.19
## 2         2       0.18
## 3         3       0.12
## 4         4       0.15
## 5         5       0.15
## 6         6       0.21
```

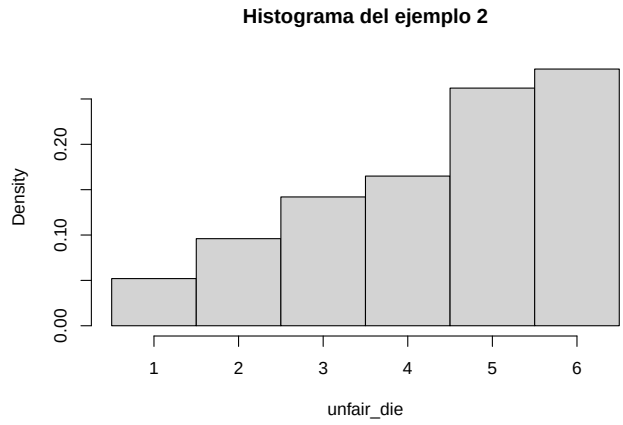
Al simular 1000 lanzamientos y representar gráficamente su tabla de frecuencias relativa, se obtiene:

```
set.seed(1) fair_die = sample(x, 1000, replace=T, prob=)
h_fair_die = hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histograma del ejemplo 1"); gra.hist.ej1 = recordPlot()
```



Si consideramos ahora el ejemplo 2 (el dado trucado), la simulación de 1000 lanzamientos de dicho dado y la representación gráfica de su frecuencia relativa se pueden realizar de la siguiente manera:

```
x=1:6; f=x/21;
set.seed(1) unfair_die = sample(x, 1000, replace=T, prob=f)
h_unfair_die = hist(unfair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histograma del ejemplo 2");
gra.hist.ej2 = recordPlot()
```

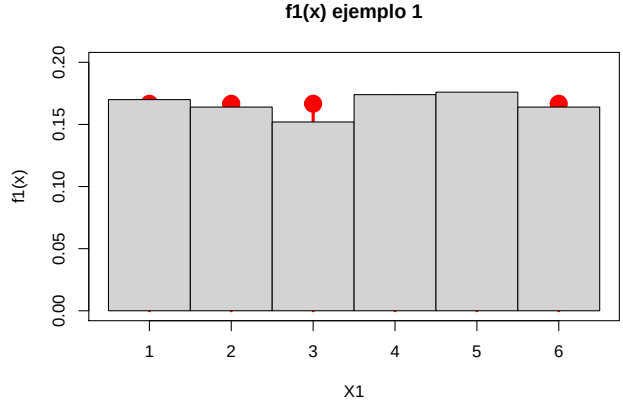


Nótese que los resultados no tienen la misma probabilidad de ocurrir, por lo que, se ha de definir previamente el vector de probabilidades.

4.2.5. Validación de los experimentos simulados y su distribución de probabilidad

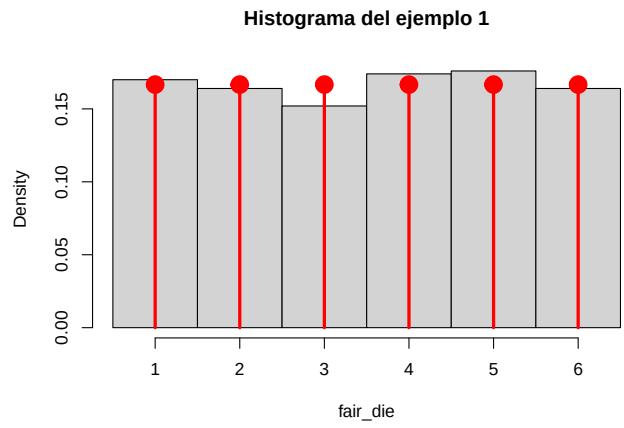
Para comparar la distribucción de una VAD con las simulaciones realizadas, lo más sencillo es realizar una gráfica en que se superpongan la función de densidad y el histograma de los resultados de las simulaciones (tabla de frecuencias). Por defecto, las funciones plot() e hist() borran la figura que se haya cargado previamente. Sin embargo, la función hist() tiene una opción que permite realizar la gráfica sin borrar la anterior:add=TRUE. Esta opción no está disponible en la función plot(). De esta manera, podríamos pensar que la comparación en el ejemplo 1 se puede realizar haciendo primero la función de densidad y posteriormente el histograma de la siguiente manera:

```
gra.fx.ej1hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE, add=T)
```



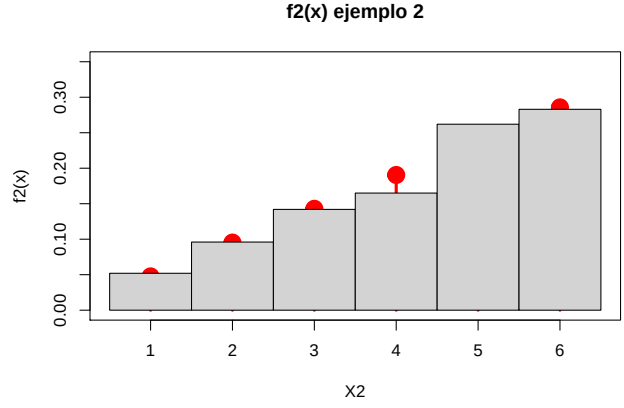
Otra forma de hacer gráficas en **R** es mediante la función lines(), que agrega líneas a una gráfica existente de una forma muy similar a como lo hace la función plot(), pero sin borrar las gráficas anteriores. Por tanto, se puede hacer primero el histograma y después la función de densidad.

```
gra.hist.ej1 x = 1:6 f = rep(1/6,6)  lines(x, f, type="h", col=red",
lwd=3)
points(x, f, col=red", lwd=10)
```



La comparación de la función de densidad del ejemplo 2 y de los resultados de simulación es:

```
gra.fx.ej2hist(unfair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE, add=T)
```



Por otro lado, también se puede comparar la media y la varianza de los experimentos simulados con el valor esperado y la varianza de la VAD; estos valores han de ser similares.

Recordemos que, para el ejemplo 1, el valor esperado es 3.5 y la varianza es 2.91667. La media y la varianza de los datos simulados del ejemplo son:

```
mean(fair_die)
```

```
## [1] 3.514
```

```
var(fair_die)
```

```
## [1] 2.936741
```

Para el ejemplo 2, el valor esperado es 4.33 y la varianza es 2.222. La media y la varianza de los datos simulados en dicho ejemplo son:

```
mean(unfair_die)
```

```
## [1] 4.338
```

```
var(unfair_die)
```

```
## [1] 2.276032
```

Tips & Tricks!

- sample(x,n,replace=T,prob=) selecciona una muestra con reemplazo de n elementos del vector x en que la probabilidad de selección es la misma para todos los elementos. En otras palabras, se están simulando n experimentos equiprobables del espacio muestral x.
- sample(x,n,replace=T,prob=p) simula n experimentos del espacio muestral x teniendo en cuenta la probabilidad de ocurrencia definida por el vector p.
- Utiliza set.seed() para asegurarte de que los resultados de cualquier función que incluye aleatoriedad sean reproducibles.

4.3. Distribuciones de probabilidad discretas más comunes

Hay muchas situaciones prácticas en la ciencia y la ingeniería en que las distribuciones de probabilidad y sus propiedades se utilizan para resolver problemas importantes. En algunas de estas situaciones, la naturaleza de la distribución e incluso una buena estimación de la estructura de la probabilidad pueden ser determinantes para los datos históricos o

de estudio a largo plazo, e incluso para grandes cantidades de datos ya previstos. No obstante, no todas las funciones de probabilidad se derivan de grandes cantidades de datos históricos. Hay numerosas situaciones cuya naturaleza sugiere un tipo concreto de distribución. Estas (también llamadas *distribuciones estándar*) se utilizan en todo el mundo en problemas de la vida real, pues el escenario científico que da lugar a cada uno de ellos es reconocible y ocurre en la práctica de forma general. En esta sesión, se analizarán y aplicarán las distribuciones de probabilidad discretas más típicas que se utilizan en ingeniería.

Distribución binomial

Muchos experimentos consisten en la repetición del ensayo, con dos posibles resultados, que pueden marcarse como exitoso o fallido (ensayo dicotómico). Si la probabilidad de éxito (p) es la misma en cada ensayo y estos son independientes, se denominan *ensayos de Bernoulli*. Si una variable discreta aleatoria (VAD) indica el número de éxitos en n ensayos de Bernoulli, con una probabilidad de éxito p, decimos que esta variable sigue una distribución binomial con los parámetros n y p. Su notación es $X \hookrightarrow B(n,p)$. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{para} \qquad x = 0, 1, 2, ..., n.$$

$$E(X) = np, \qquad V(X) = np(1-p).$$

Distribución geométrica

Si una VAD indica el número de ensayos de Bernoulli necesarios hasta el primer suceso, con una probabilidad de éxito p, decimos que esta variable sigue una distribución geométrica con un parámetro p y su notación es $X \hookrightarrow G(p)$. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = P(X = x) = p(1-p)^{x-1} \qquad \text{para} \qquad x = 1, 2, ...$$

$$E(X) = \frac{1}{p}, \qquad V(X) = \frac{(1-p)}{p^2}.$$

Distribución binomial negativa

Si una VAD indica el número de ensayos de Bernoulli necesarios hasta obtener r éxitos, con una probabilidad de éxito p, podemos decir que esta variable sigue la distribución binomial negativa con los parámetros r y p. Su notación es $X \hookrightarrow NB(r,p)$. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad \text{para} \qquad x = r, r+1, r+2, ...$$

$$E(X) = \frac{r}{p}, V(X) = r\frac{(1-p)}{p^2}.$$

Distribución hipergeométrica

Si una VAD indica el número de éxitos en n ensayos dependientes dicotómicos, con una probabilidad de éxito que cambia en cada ensayo (población que consiste en N éxitos y $N-k$ fracasos), podemos decir que esta variable sigue la distribución hipergeométrica con parámetros n, N y k. Su notación es $X \hookrightarrow H(n, N, k)$. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \quad \text{para} \quad x = max\{0, n+k-N\}, ..., min\{k, n\}$$

$$p = \frac{k}{N}, E(X) = np, V(X) = np(1-p)\frac{N-n}{N-1}.$$

Distribución de Poisson

Los experimentos que tienen como resultado el número de eventos que ocurren durante un intervalo de tiempo dado o en una región específica se denominan *ensayos de Poisson*. Si una VAD indica el número de *ensayos de Poisson*, con una frecuencia de ocurrencia media λ, podemos decir que esta variable sigue la distribución de Poisson con parámetro λ. Su notación es $X \hookrightarrow P(\lambda)$ y su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{para} \qquad x = 0, 1, 2, ...$$

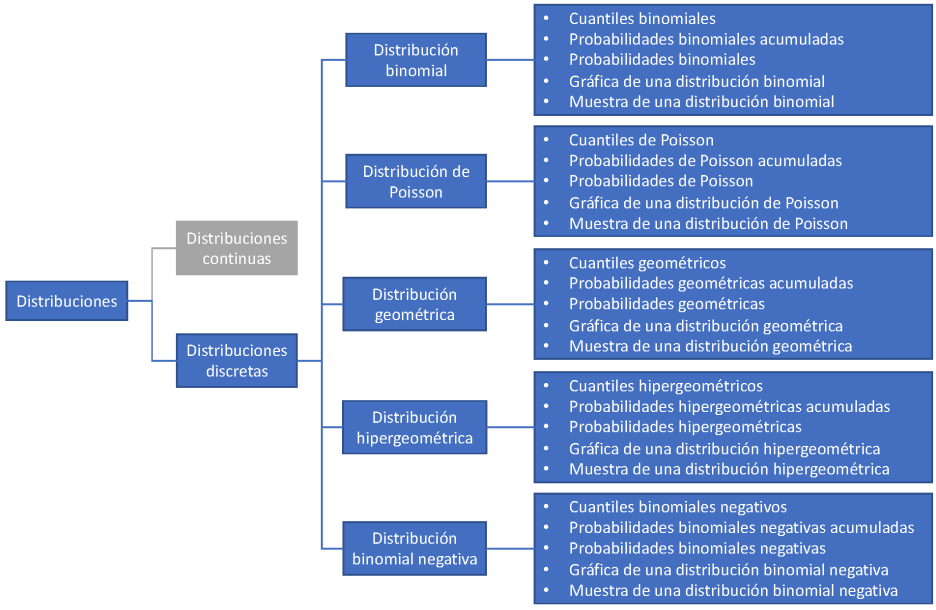
$$E(X) = \lambda, \qquad V(X) = \lambda.$$

Dada una función de distribución discreta comúnmente usada en ingeniería (binomial, geométrica, binomial negativa, hipergeométrica o de Poisson) con sus respectivos parámetros, utilizando **R** y/o **R-Commander** su pueden obtener muy fácilmente la probabilidad de un evento elemental, las probabilidades acumulativas, los cuantiles y los gráficos de distribuciones estadísticas estándar (que pueden usarse, por ejemplo, como sustituto de las tablas estadísticas); además, se pueden generar muestreos o simulaciones de estas distribuciones.

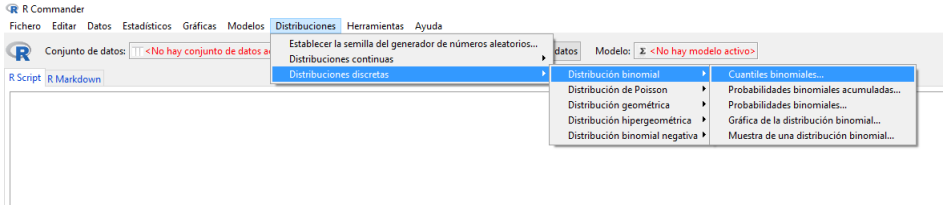
Si se dispone de **R-Commander**, en la barra superior podemos encontrar en *Distribuciones* > *Distribuciones discretas* todas las distribuciones estudiadas. El árbol del submenú completo para distribuciones de probabilidad discretas se muestra en la siguiente figura. La mayoría de opciones del menú nos llevan a diferentes cuadros de diálogo. Las opciones del menú están inactivas (en gris) si no se pueden aplicar al contexto actual.

Si no se dispone de **R-Commander**, o si se prefiere usar la **R-Console** o el **Rstudio**, o simplemente se está creando un *script* para su posterior uso, también se indicarán las instrucciones adecuadas para cada caso. Por otra parte, para no ser repetitivos, en esta guía solo se explicarán todas las opciones concernientes a la distribución binomial, las demás

distribuciones tienen las mismas opciones, con la diferencia de que los parámetros son diferentes para cada distribución. Por ejemplo, la distribución binomial tiene como parámetros n y p, mientras que la distribución geométrica solo tiene p, y la hipergeométrica tiene N, n y k.



Dada una VAD X que sigue una distribución binomial con los parámetros n y p, es decir, $X \hookrightarrow B(n, p)$, se pueden calcular las probabilidades binomiales (función de densidad) y las probabilidades acumuladas, realizar los gráficos de las funciones de densidad y de distribución, calcular los cuantiles binomiales y realizar un muestreo o simulación de experimentos binomiales, tal como se muestra en la figura.

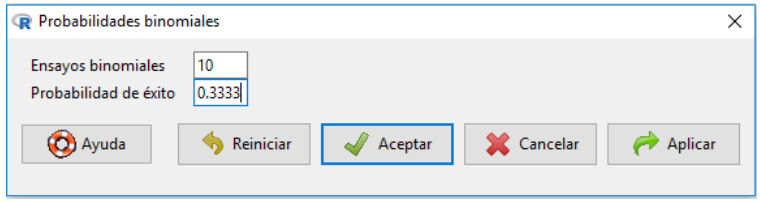


4.3.1. Probabilidades elementales

Calcula la función de densidad de la distribución dada, es decir, la probabilidad de que la variable X tome cada uno de los valores del espacio muestral, $P(X = x)$.

Ejemplo: Un tercio de la población de una comunidad tiene una enfermedad concreta. De estas, se escogen 10 personas de forma aleatoria. ¿Cuál es la probabilidad de que 4 personas en un muestreo aleatorio de 10 personas de esta comunidad padezcan la enfermedad?

Si definimos la VAD X que denota el número de personas que padecen la enfermedad de un muestreo aleatorio de 10 personas, donde la probabilidad de éxito es 1/3, entonces $X \hookrightarrow B(10, 1/3)$. Por tanto, $P(X = x)$ se obtiene al introducir los siguientes parámetros en el cuadro de diálogo que se genera al ejecutar *Distribuciones -> Distribuciones discretas -> Distribución binomial -> Probabilidades binomiales*....



En la ventana de instrucciones, se generan el siguiente código con su respectiva respuesta en la ventana de resultados:

```
local({  .Table <- data.frame(Probability=dbinom(0:10, size=10,
prob=0.3333))
rownames(.Table) <- 0:10 print(.Table) })
```

```
##     Probability
## 0  1.735020e-02
## 1  8.673800e-02
## 2  1.951312e-01
## 3  2.601359e-01
## 4  2.275848e-01
## 5  1.365304e-01
## 6  5.687914e-02
## 7  1.624874e-02
## 8  3.046183e-03
## 9  3.384140e-04
## 10 1.691816e-05
```

Por tanto, la probabilidad de que 4 personas padezcan la enfermedad es $P(X = 4) = 0.2275848$. Nótese que los valores de la función de densidad ($f(x)$) se calculan mediante la función dbinom(x, size=n, prob=p), donde x es el vector de valores de X del cual se quiere calcular la probabilidad, n es el número total de experimentos y p es la probabilidad

de éxito. De esta manera, si se prefiere usar directamente los comandos en la **R-Console** o **Rstudio**, se pueden calcular solo las probabilidades deseadas siguiendo la siguiente instrucción:

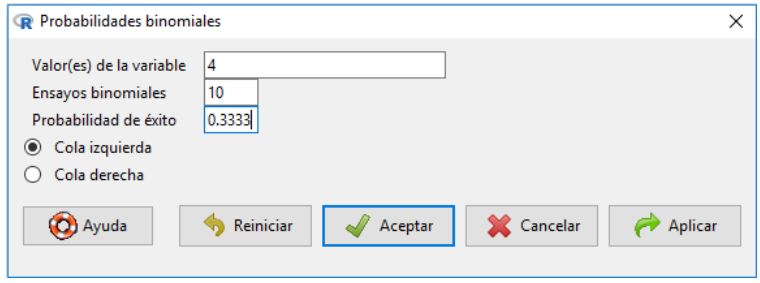
```
dbinom(4, size=10, prob=0.3333)
```

```
## [1] 0.2275848
```

4.3.2. Probabilidades acumuladas

Calcula la función de distribución acumulada de la distribución dada: la probabilidad de que la variable X sea como máximo x, $P(X \leq x)$. Además, $P(X > x)$ también se puede calcular seleccionando la opción "Cola derecha" en lugar de "Cola izquierda". Siguiendo el ejemplo anterior, ¿cuál es la probabilidad de que 4 personas o menos en un muestreo aleatorio de 10 personas de esta comunidad padezcan la enfermedad?

Para calcular $P(X \leq 4)$, se introducen los siguientes parámetros en su correspondiente cuadro de diálogo:



Y se genera el siguiente código y su resultado:

```
pbinom(c(4), size=10, prob=0.3333, lower.tail=TRUE)
```

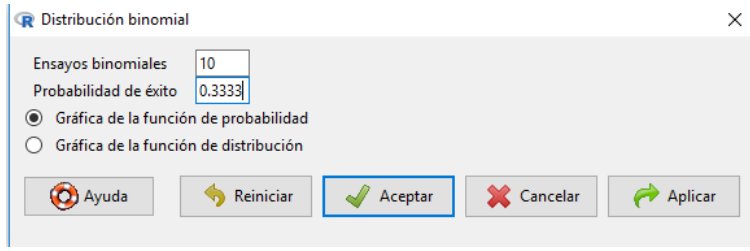
```
## [1] 0.7869402
```

Por tanto, $P(X \leq 4) = 0.7875542$. Si se selecciona "Cola derecha", la probabilidad calculada sería: $P(X > 4) = P(X \geq 5) = 1 - P(X \leq 4)$.

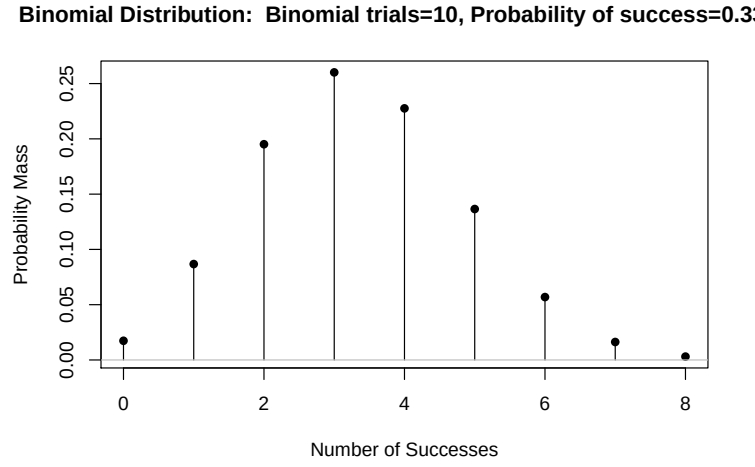
Nótese que las probabilidades acumuladas se calculan mediante la función pbinom(x, size=n, prob=p, lower.tail=), donde, además de x, size y prob, se ha de definir la cola lower.tail, TRUE para la izquierda y FALSE para la derecha.

4.3.3. Gráfica de una distribución

Esta opción permite generar la representación gráfica de la función de densidad o la función de distribución de la distribución dada. La función de densidad del ejemplo se genera de la siguiente manera:

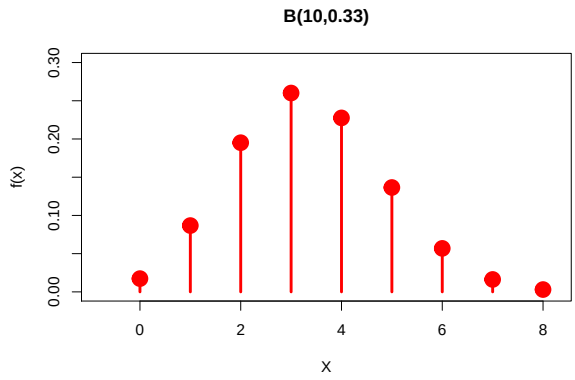



```
local({  .x <- 0:8 plotDistr(.x, dbinom(.x, size=10, prob=0.3333
), xlab="Number of Successes",  ylab="Probability Mass",
main="Binomial Distribution: Binomial trials=10,
      Probability of success=0.3333",
discrete=TRUE) })
```



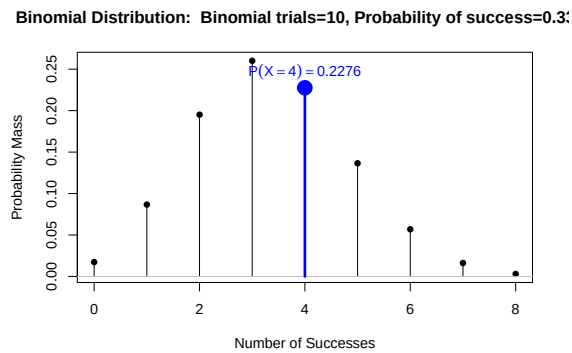
Para hacer la gráfica directamente desde **R-Console** o **Rstudio**, se pueden utilizar las instrucciones explicadas al principio de la guía de la siguiente forma:

```
x = 0:8  Posibles resultados f = dbinom(x, size=10, prob=0.3333)
Probabilidad de ocurrencia de cada resultado
plot(x, f, type="h", col=red", lwd=3, main="B(10,0.33)",
xlab="X", ylab="f(x)", xlim=c(-0.8,8.2), ylim=c(0,0.30))
Líneas verticales
points(x, f, col=red", lwd=10); gra.fx.binom = recordPlot() Puntos
```

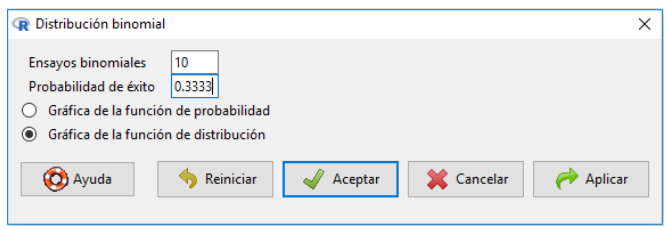


Independientemente de cómo se ha generado la gráfica de la función de densidad, es posible resaltar alguna probabilidad específica. Continuando con el ejemplo, la probabilidad de que 4 personas en un muestreo aleatorio de 10 personas padezcan la enfermedad se puede representar agregando el siguiente código:

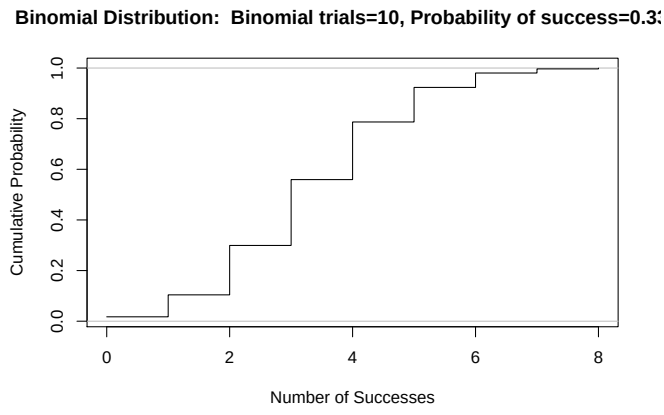
```
x = 4 f_4 = dbinom(4, size=10, prob=0.3333)  Cálculo de P(X=4)=f(4)
lines(x, f_4, type="h", col="blue", lwd=3)  Agrega la línea en X=4
points(x, f_4, col="blue", lwd=10)  Agrega el punto en (4,f(4))
text(x, f_4,expression(P(X==4)), pos=3, col="blue")
```



Por otra parte, la gráfica de la función de distribución se puede generar al introducir los siguientes parámetros en el cuadro de diálogo.

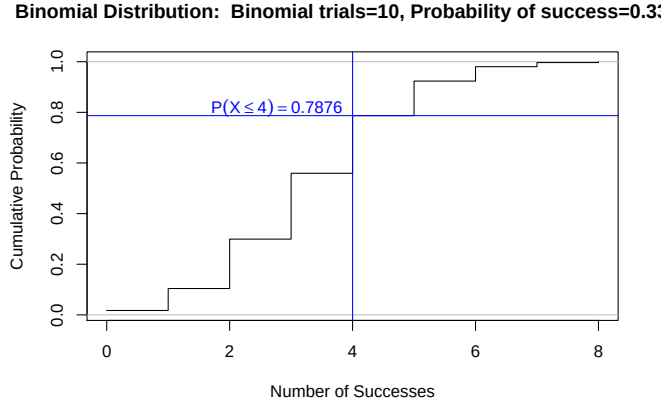


```
local({  .x <- 0:8 plotDistr(.x, pbinom(.x, size=10, prob=0.3333),
xlab="Number of Successes", ylab=Çumulative Probability",
main="Binomial Distribution: Binomial trials=10,
      Probability of success=0.3333",
discrete=TRUE, cdf=TRUE) })
```



De la misma forma, se puede representar la probabilidad de que 4 personas o menos en un muestreo aleatorio de 10 personas de esta comunidad padezcan la enfermedad:

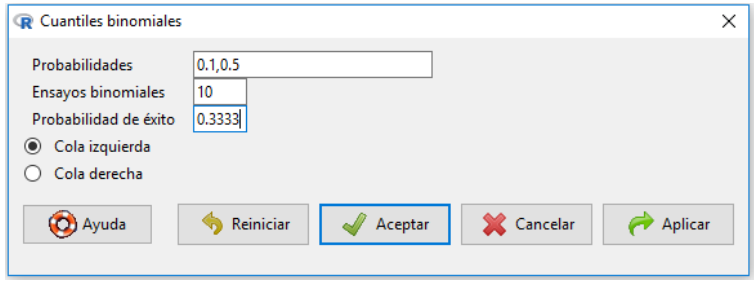
```
x = 4 f_4 = pbinom(4, size=10, prob=0.3333)
abline(v=x, col="blue") abline(h=f_4, col="blue")
text(x, f_4,expression(P(X<=4)== P(X<5)),
pos=2, col="blue")
```



4.3.4. Cuantiles

El p-ésimo cuantil se define como el valor más pequeño de x, de modo que $F(x) = P(X \leq x) \geq p$. Este se calcula seleccionando la opción "Cola izquierda". También se puede calcular el valor más pequeño de x tal que $P(X > x) \geq p$ seleccionando la opción "Cola derecha". Continuando con el ejemplo, se determinará el décimo cuantil, la mediana y el valor de x tal que $P(X > x) = 0.15$.

El décimo cuantil y la mediana indican el valor de x tal que $P(X \leq x) = 0.1$ y $P(X \leq x) = 0.5$, respectivamente. Por tanto, si seleccionamos la primera opción del submenú de "Distribución binomial", nos aparece la siguiente ventana:




```
qbinom(c(0.1,0.5), size=10, prob=0.3333, lower.tail=TRUE)
```

```
## [1] 1 3
```

Entonces, el décimo cuantil es 1, es decir $P(X \leq 1) \geq 0.1$, y la mediana es 3, o sea, $P(X \leq 3) \geq 0.5$. Para calcular el valor de x tal que $P(X > x) = 0.15$, el parámetro "probabilidades" se configura en 0.15 y se selecciona la opción "Cola derecha"; de esta manera, se genera el siguiente comando y salida:

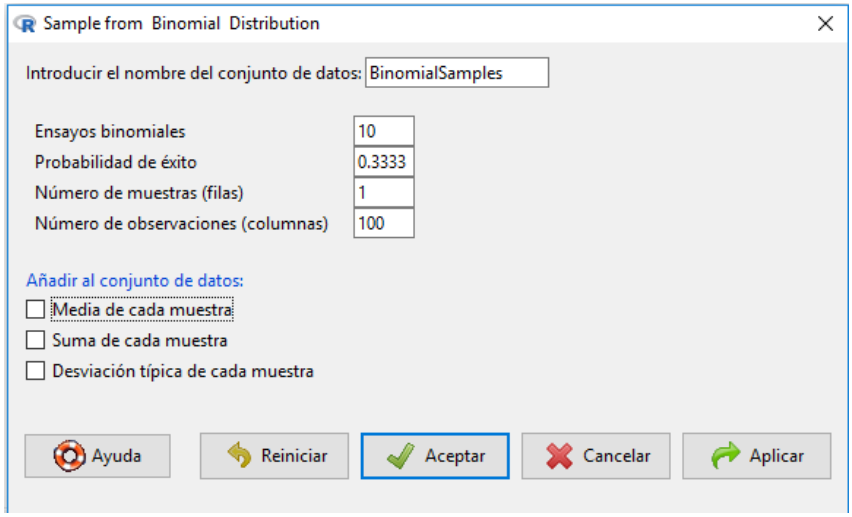
```
qbinom(c(0.15), size=10, prob=0.3333, lower.tail=FALSE)
```

```
## [1] 5
```

Lo cual indica que, $P(X > 5) \leq 0.15$.

4.3.5. Muestreo

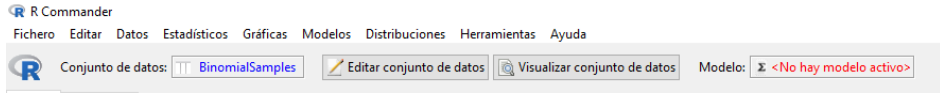
Simula escenarios aleatorios para la distribución dada. Para el ejemplo previo, si queremos simular una observación de 100 muestras, cada observación o experimento consiste en seleccionar 10 personas de la población y contar cuántas de ellas padecen la enfermedad. Los parámetros que debemos introducir en el cuadro de diálogo son los siguientes:



También se pueden añadir algunas medidas extras, como la media, la suma de cada muestra y la desviación típica.

```
BinomialSamples <- as.data.frame(matrix(rbinom(1*100, size=10,
                  prob=0.3333), ncol=100))
rownames(BinomialSamples) <- "sample"
colnames(BinomialSamples) <- paste(.ºbs", 1:100, sep=)
```

Una vez hemos simulado los datos, estos ya aparecen en el conjunto de datos del **R Commander**. En consecuencia, el nombre del conjunto de datos aparece en el botón que hay junto al extremo izquierdo en la ventana principal. Los datos simulados se guardan en el vector llamado BinomialSamples y se pueden ver haciendo clic en *Visualizar conjunto de datos*.



Tips & Tricks!

En caso de no disponer o hacer uso del **R-Commander**, se pueden introducir directamente en **R-Console** o **Rstudio** las siguientes funciones:

- dbinom() calcula la función de distribución de una distribución binomial, es decir, $f(x) = P(X = x)$ para todos los posibles valores de x.
- pbinom() calcula la probabilidad acumulada de una distribución binomial, es decir, $F(x) = P(X \leq x)$ o $1 - F(x) = P(X > x)$ para un valor de x dado.
- qbinom() calcula el p-ésimo cuantil de una distribución binomial, es decir, el valor de x tal que $F(x) = P\left(X \leq x\right) \geq p$ para un valor de p dado.
- rbinom() realiza un muestreo o simulación de un número determinado de experimentos de una distribución binomial.

Nótese que la primera letra indica: función de distribución (d), probabilidad (p), cuantil (q) o muestra aleatoria (r). El resto de letras indican la distribución: binomial (binom), geométrica (geom), binomial negativa (nbinom), hipergeométrica (hyper) o Poisson (pois). De esta manera, si lo que se quiere es calcular un cuantil de una distribución geométrica, la función correspondiente es qgeom(); para la distribución binomial negativa es qnbinom(); para la hipergeométrica es qhyper(), y para la de Poisson es qpois().

4.4. Ejercicios propuestos

1. Se lanzan dos dados de seis caras. Uno de ellos solo tiene números impares y el otro, pares. Define la variable aleatoria X como la suma de los números que han salido.
   - Determina la función de densidad y grafícala.
   - Simula 1000 lanzamientos y muestra su frecuencia relativa.
   - Compara (en el mismo gráfico) la función de densidad y la frecuencia relativa de las simulaciones.
2. Según *Chemical Engineering Progress* (noviembre de 1990), aproximadamente el 30 % de todos los problemas de tuberías en las plantas químicas se deben a errores del operador.
   - Con los siguientes 20 fallos en las tuberías, determina la probabilidad de que:
     - Al menos 10 se deban a un error del operador.
     - No más de 4 sean por un error del operador.
     - Exactamente 5 sean por un error del operador.
   - ¿Cuál es el número esperado de errores causados por un operador que pueden ocurrir en los siguientes 20 problemas en las tuberías? ¿Y la varianza?

- Simula una observación de 1000 muestras (cada muestra consiste en contar cuántos fallos son causados por un error del operador en los siguientes 20 problemas en las tuberías) y compara los resultados.

3. Durante la Segunda Guerra Mundial, se lanzaron 535 bombas sobre el sur de Londres. Esta área ha sido dividida en una cuadrícula de 576 pequeños cuadrados de 0,25 metros cuadrados cada uno. Asumiendo que el objetivo es aleatorio: cada bomba impacta en un solo lugar a la vez, cada lugar tiene la misma probabilidad de ser impactado y los impactos son sucesos independientes, encuentra:

   - La gráfica de las funciones de densidad y de distribución de la variable que indica el número de bombas que impactan en un cuadrado en particular.
   - ¿Cuál es la probabilidad de que exáctamente 2 bombas impacten en una zona en particular?
   - ¿Cuál es la probabilidad de que una zona concreta sea bombardeada (caiga al menos una bomba sobre ella)?
   - La gráfica de la función de densidad de la variable que indica el número de zonas que reciben exáctamente 2 impactos.
   - ¿Cuántas zonas se espera que sufran exáctamente dos impactos?
   - La gráfica de la función de densidad de la variable que indica el número de zonas que deben ser inspeccionadas para encontrar 10 que hayan sido bombardeadas.
   - ¿Cuántas zonas han de ser inspeccionadas para encontrar 10 que hayan sido bombardeadas?


2
1
3

Variables aleatorias continuas y distribuciones de probabilidad

5.1. Introducción y objetivos

La sesión anterior estaba centrada en el análisis de probabilidades de variables aleatorias discretas (VAD) y la aplicación de sus distribuciones de probabilidad más comunes en ingeniería. El objetivo de esta sesión es mostrar, analizar y aplicar el concepto de la variable aleatoria continua (VAC), simular repeticiones de un experimento y comparar sus resultados con su distribución. Además, se describen los modelos de distribución continuos más utilizadas. Al finalizar esta sesión, el alumno ha de ser capaz de:

- Representar gráficamente una distribución de variable aleatoria continua usando **R**.
- Simular la repetición de diferentes experimentos aleatorios continuos y comparar el resultado de estos experimentos con las probabilidades estudiadas previamente.
- Calcular e interpretar el valor esperado y la varianza de una variable aleatoria continua.
- Reconocer y aplicar correctamente las distribuciones de probabilidad continua más comunes en ingeniería.

5.2. Variables aleatorias continuas (VAC)

Como ya se ha descrito en la sesión previa, una variable aleatoria **X** se puede definir como la función que asigna un número real a cada salida de un espacio muestral Ω. En otras palabras, es una función de dominio Ω y rango $\mathbb{R}$. Una variable aleatoria es una variable aleatoria continua (VAC o CRV, por sus iniciales en inglés) si su conjunto de salidas no puede contarse. Esto es, un espacio muestral contiene un número infinito de posibilidades igual al número de puntos en un segmento de línea. Toma valores en una escala continua, normalmente valores que son precisamente los mismos valores que están contenidos en

un espacio muestral continuo. En los problemas prácticos, estas variables representan datos medidos, como todos los posibles pesos, alturas, temperaturas, distancias, etc.

5.2.1. Función de densidad

La función de densidad de probabilidad (o simplemente, función de densidad) de una variable aleatoria continua X, definida en el espacio muestral Ω, es la aplicación de f tal que:

$$f : X(\Omega) \to \mathbb{R}$$
$$x_i \to f(x_i)$$

Esta función tiene las siguientes propiedades:

- Su valor siempre es un número real positivo o cero: $0 \leq f(x)$
- El área total bajo su curva es 1: $\int_{-\infty}^{\infty} f(x)dx = 1$
- La probabilidad de que X tome un valor en el intervalo $[a, b]$ es el área bajo la curva de la función de densidad en ese intervalo: $P\,(a \leq X \leq b) = \int_{a}^{b} f(x)dx, \qquad \forall a \leq b$
- Teniendo en cuenta que la probabilidad de que X tome un valor exacto es cero ($P(X = x) = 0$),

$$f(x) \neq P(X = x)$$
$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

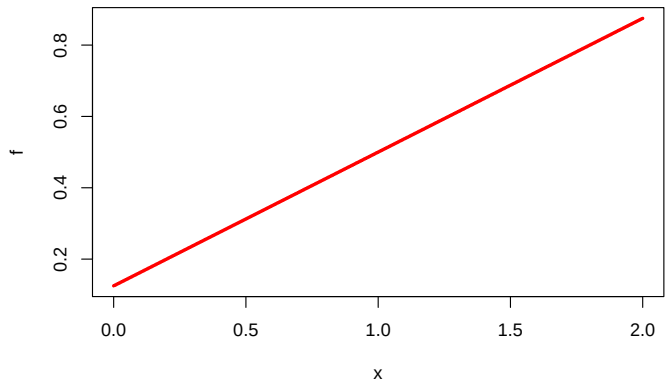
Ejemplo: Consideremos X la variable aleatoria que indica la carga dinámica en un puente (en newtons), cuya función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \leq x \leq 2 \\ 0 & \text{resto} \end{cases}$$

La función de densidad de una VAC se puede representar gráficamente por medio de una curva continua. En **R**, se puede hacer ejecutando las siguientes instrucciones:

```
rm(list=ls())  Elimina todos los objetos del espacio de trabajo
graphics.off()  Elimina las gráficas creadas con anterioridad
```

```
x = seq(0,2,0.0001)  Posibles resultados
f = 1/8+3/8*x  función de densidad
plot(x, f, type="l", col=red", lwd=3)
```



Nótese que, a diferencia de la gráfica de función de densidad de VAD, el vector de posibles resultados o espacio muestral (x) contiene muchos valores entre 0 y 2 (tendiendo a infinito) igualmente espaciados. Por otra parte, la función plot() utiliza el parámetro type=l para especificar que se dibujará una línea que une todos los puntos, al contrario que type=h, que dibuja líneas verticales como en el caso de VAD.

Si se quiere representar gráficamente la función de probabilidad incluyendo también algunos valores de X en que la probabilidad es cero, se puede ejecutar el siguiente código:

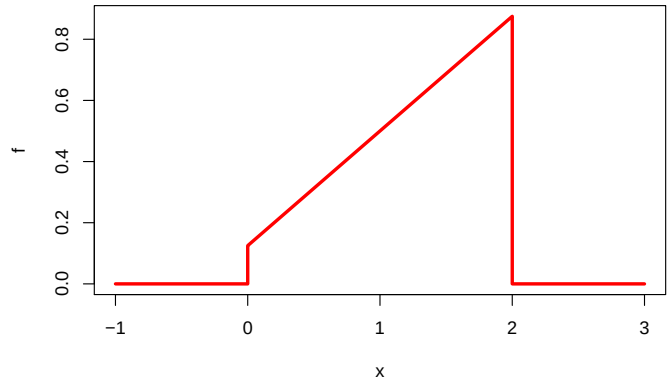
```
LowLim = 0  Mínimo valor que toma X
UppLim = 2  Máximo valor que toma X
delta = 0.001  Resolución, diferencia entre valores de X

Valores parciales de X  x.less0 = seq(LowLim-1, LowLim, delta)
-1 <X <0
x.less2 = seq(LowLim, UppLim, delta)  0 <X <2
x.greater2 = seq(UppLim, UppLim+1, delta)  2 <X <3

Valores parciales de la función de densidad
fx.less0 = rep(0, length(x.less0))  f(x) para -1 <X <0
fx.less2 = 1/8+3/8*x.less2  f(x) para 0 <X <2
fx.greater2 = rep(0, length(x.greater2))  f(x) para 2 <X <3

Vectores finales
x = c(x.less0, x.less2, x.greater2)  -1 <X <3
f = c(fx.less0, fx.less2, fx.greater2)  f(x) para -1 <X <3

Representación gráfica
plot(x, f, type="l", col=red", lwd=3);
gra.fx = recordPlot()
```

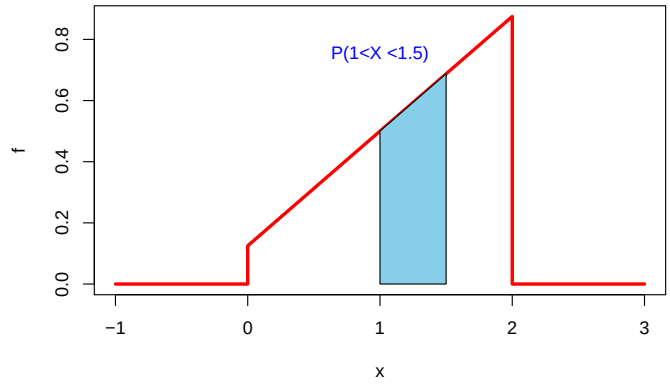


Finalmente, si se quiere resaltar en la gráfica una probabilidad específica, por ejemplo $P(1 < X < 1.5)$, o sea:

$$P(1 < X < 1.5) = \int_1^{1.5} f(x)dx = \int_1^{1.5} \left(\frac{1}{8} + \frac{3}{8}x\right) dx = 0.297,$$

Se puede agregar un polígono a la gráfica de la función de densidad especificando sus vértices.

```
x.i = 1  Límite inferior
x.f = 1.5  Límite superior
x.x = seq(x.i, x.f, delta)  Valores de X dentro de los límites
coord.x = c(x.i, x.x, x.f)  Vector de vértices en la coordenada x
coord.y = c(0, 1/8+3/8*x.x, 0)  Vector de vértices en la coordenada y
polygon(coord.x, coord.y, col="skyblue")  Área bajo la curva
text(x.i, 0.7, "P(1 <X <1.5)", pos=3, col="blue")
```



5.2.2. Función de distribución

La función de distribución o función de probabilidad acumulada asociada a una variable aleatoria continua X, definida en un espacio muestral Ω, es la aplicación F que asigna a cada elemento x_i de $X(\Omega)$ la probabilidad de que la variable X tome cualquier valor menor o igual que x_i:

$$F : X(\Omega) \to \mathbb{R}$$

$$x_i \to F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x} f(u)du$$

Por tanto:

$$P(a \leq X \leq b) = F(b) - F(a), \qquad \forall a \leq b$$

$$f(x) = F'(x)$$

En el ejemplo, la función de distribución viene dada por:

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}x\right) dx = \frac{1}{8}x + \frac{3}{16}x^2 \qquad \text{para} \qquad 0 \leq x \leq 2,$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8}x + \frac{3}{16}x^2 & 0 \leq x \leq 2 \\ 1 & 2 < x \end{cases}$$

La función de distribución se representa gráficamente mediante una curva continua que indica la probabilidad $F(x) = P(X \leq x)$, donde x puede ser cualquier valor real. Nótese que esta función comienza en 0 y termina en 1. La función de distribución del ejemplo se representa gráficamente de la siguiente manera:

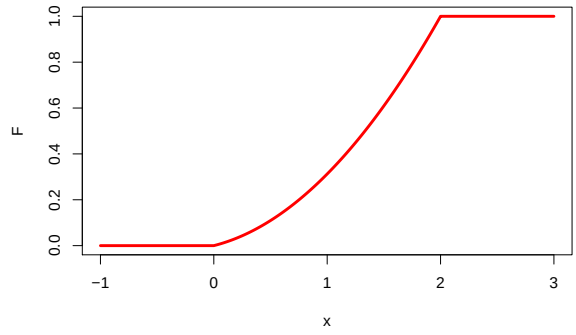
```
LowLim = 0  Mínimo valor que toma X
UppLim = 2  Máximo valor que toma X
delta = 0.001  Resolución, diferencia entre valores de X

Valores parciales de X
x.less0 = seq(LowLim-1, LowLim, delta)  -1 <X <0
x.less2 = seq(LowLim, UppLim, delta)  0 <X <2
x.greater2 = seq(UppLim, UppLim+1, delta)  2 <X <3

Valores parciales de la función de distribución
Fx.less0 = rep(0, length(x.less0))  F(x) para -1 <X <0
Fx.less2 = 1/8*x.less2+3/16*x.less2^2  F(x) para 0 <X <2
Fx.greater2 = rep(1, length(x.greater2))  F(x) para 2 <X <3
```

```
Vectores finales
x = c(x.less0, x.less2, x.greater2)  -1 <X <3
F = c(Fx.less0, Fx.less2, Fx.greater2)  F(x) para -1 <X <3

Representación gráfica
plot(x, F, type="l", col=red", lwd=3)
```

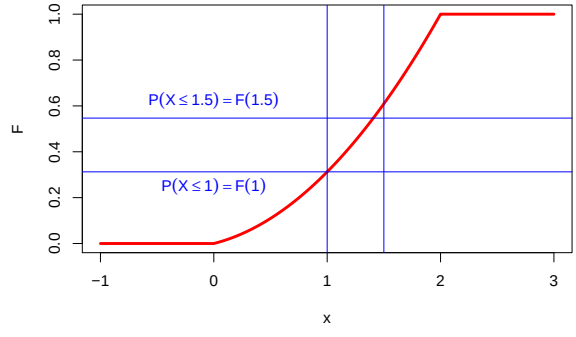


Finalmente, si se quiere resaltar en la gráfica una probabilidad específica, por ejemplo la probabilidad de que la carga dinámica esté entre 1 y 1.5 $P(1 < X < 1.5)$ que viene dada por:

$$P(1 < X < 1.5) = F(1.5) - F(1) = \\ = \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] = 0.297$$

Se agregan dos líneas horizontales: una en $F(1.5)$ y otra en $F(1)$.

```
a = 1; b = 1.5 Fa = 1/8* a+3/16*a^2;Fb = 1/8*a +3/16*b^2
abline(v=b,col="blue")
abline(h=Fb,col="blue")
text(0,Fb, expression(P(X <= 1.5) == F(1.5)), pos=3, col="blue")
abline(v=a,col="blue") abline(h=Fa,col="blue")
text(0,Fa, expression(P(X <= 1) == F(1)), pos=1, col="blue")
```



Tips & Tricks!

La diferencia entre la definición de VAD y VAC radica en el vector del espacio muestral x=seq(). Un vector con muchos (acercándose al infinito) valores de x en que la distancia entre valores consecutivos es muy pequeña, puede considerarse un vector de espacio muestral de una VAC. Cuanto más grande es este vector, mejor es la aproximación. Sin embargo, el cálculo computacional es también mayor. Por tanto, se aconseja seleccionar una longitud de vector adecuada, de tal manera que pueda considerarse VAC pero sin sobrecargar el trabajo del ordenador.

La instrucción polygon(coord.x, coord.y) dibuja un polígono cuyos vértices se especifican en los vectores coord.x y coord.y.

5.2.3. Medidas características de las VAC

Estos parámetros o características, como sucede en el caso de las VAD, cuantifican la tendencia central y la variabilidad o dispersión de la variable aleatoria continua. De hecho, conociendo estas cantidades, a parte de la distribución completa, pueden darnos una idea de la naturaleza del sistema.

Valor esperado

Este parámetro explica cómo "esperamos" que la variable se comporte por término medio a largo plazo (es lo que también se denomina *teoría frecuencial de la probabilidad*). El valor esperado $E(X)$ de la VAC X viene dado por:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

De esta forma, el valor esperado de la carga dinámica del puente del ejemplo (X) viene dada por:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{2} x \left(\frac{1}{8} + \frac{3}{8} x \right) dx = \frac{5}{4} = 1.25,$$

Varianza

Este parámetro mide la dispersión de los posibles valores de X. La varianza es el promedio (esperado) de la distancia al cuadrado (o desviación) de la media:

$$\sigma^2 = Var(X) = E\left[(X - E[X])^2 \right] = E[X^2] - E[X]^2 =$$
$$= \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - E[X]^2$$

Por tanto, la varianza de la carga dinámica del ejemplo (X) es:

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{0}^{2} \left(x - \frac{5}{4}\right)^2 \left(\frac{1}{8} + \frac{3}{8}x\right) dx = \frac{13}{48} = 0.2708.$$

5.2.4. Uso de sample() para generar simulaciones

Con **R**, también se pueden generar observaciones de una VAC utilizando la función sample() (como ya se ha explicado en la sesión anterior). El vector x que contiene los valores de los cuales se quiere hacer el muestreo (espacio muestral) ha de ser suficientemente grande para representar el mayor número posible de valores del intervalo continuo. El vector de probabilidades con que se rige cada elemento del espacio muestral se genera usando la función de densidad de la VAC.

Considerando el ejemplo, un experimento u observación se puede simular de la siguiente manera:

```
x = seq(0,2,0.01) f = 1/8+3/8*x sample(x, size=1, prob=f)
```

```
## [1] 0.57
```

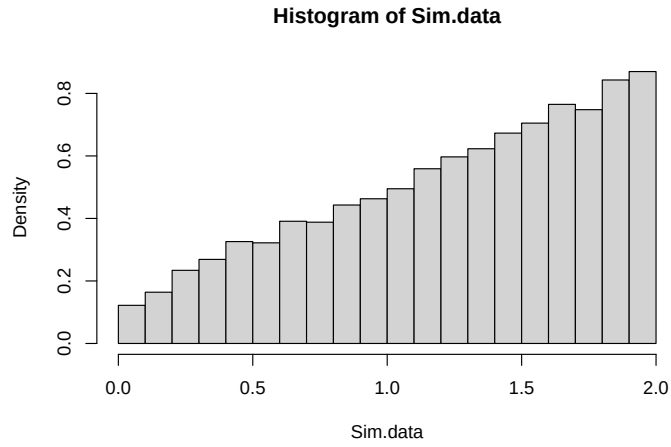
Para simular 100 observaciones:

```
sample(x, size=100, replace=T, prob=f)
```

```
##   [1] 0.68 1.28 1.72 1.07 1.92 1.76 0.60 1.71 1.69 1.70 0.92 0.66
        1.64 0.80 1.87
##  [16] 1.28 1.62 1.97 1.96 0.33 1.99 0.31 0.54 1.59 0.75 0.49 1.37
        1.68 0.53 1.75
##  [31] 0.38 1.82 1.71 1.44 1.82 1.41 1.22 0.42 1.75 1.92 1.21 1.76
        1.31 0.84 1.31
##  [46] 1.58 0.88 0.96 1.54 1.73 1.12 1.46 1.36 1.66 0.59 1.20 0.13
        1.70 1.71 0.84
##  [61] 1.99 1.94 1.57 1.84 1.73 0.48 1.13 1.79 1.58 1.74 1.86 1.61
        1.22 1.75 1.01
##  [76] 1.40 0.63 1.65 1.00 1.99 0.81 1.49 1.18 1.27 0.97 1.39 0.93
        1.91 1.95 1.65
##  [91] 0.50 0.01 1.62 1.34 1.68 1.54 0.35 1.99 1.14 1.63
```

Esta simulación no es útil, debido a la gran cantidad de posibles resultados que no están incluidos en el vector x. Una buena simulación de una VAC debería incluir miles de posibles resultados, no solo 200. Para incrementar la longitud del vector x donde la separación entre valores sea solo de 0.0001, simula 1000 observaciones y ver la densidad de los resultados (histograma), se puede ejecutar el siguiente código:

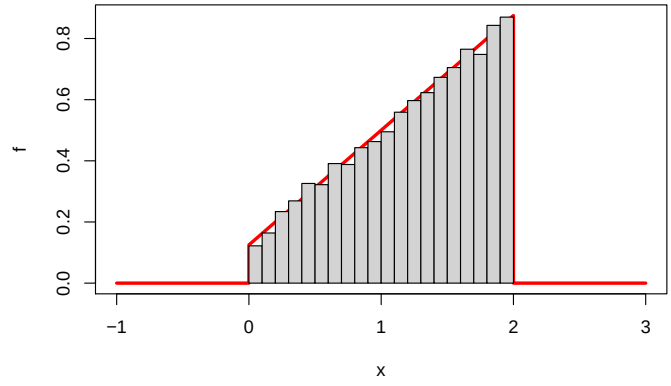
```
x = seq(0,2,0.0001)  f = 1/8+3/8*x set.seed(10)
Sim.data = sample(x, size=10000, replace=T, prob=f)
hist(Sim.data, freq=FALSE); gra.his = recordPlot()
```



5.2.5. Validación de los experimentos simulados y su distribución de probabilidad

La comparación de la distribucción de una VAC con las simulaciones realizadas se hace de igual forma que con una VAD, superponiendo la función de densidad y el histograma de los resultados de las simulaciones (tabla de frecuencias).

```
gra.fx hist(Sim.data, freq=FALSE, add=T)
```



Por otro lado, también se pueden comparar la media y la varianza de los experimentos simulados con el valor esperado y la varianza de la función de densidad; estos valores han de ser similares. Recordemos que, para el ejemplo, el valor esperado es 1.25 y la varianza es 0.2708. La media y la varianza de los datos simulados en dicho ejemplo son:

```
mean(Sim.data)
```

```
## [1] 1.251668
```

```
var(Sim.data)
```

```
## [1] 0.2707881
```

5.3. Distribuciones de probabilidad continuas más comunes

Como ya se ha mencionado en la sesión anterior, existen funciones de probabilidad tanto en VAD como en VAC que son típicas y ampliamente utilizadas en estadística. En esta sesión, se estudia dos de las funciones de distribución o modelos de probabilidad continuos más importantes en ingeniería. Lo que se vea aquí podrá extrapolarse al resto de modelos de probabilidad continuos, teniendo en cuenta los parámetros para cada caso.

Distribución uniforme

Si una VAC toma cualquier valor en un intervalo $[a, b]$ con la misma probabilidad, decimos que esta variable sigue una distribución uniforme continua. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = \frac{1}{b-a} \qquad \forall x \in [a, b]$$

$$E(X) = \frac{a+b}{2}$$

$$V(X) = \sigma^2 = \frac{(b-a)^2}{12}$$

Distribución exponencial

La familia de distribuciones exponenciales proporciona modelos que se utilizan mucho en la ciencia y en la ingeniería. La VAC, que es igual a la distancia en los sucesivos eventos de un proceso de Poisson con una media de eventos $\lambda > 0$ por intervalo la unidad, sigue una distribución exponencial con parámetro λ. Su función de densidad, su media y su varianza vienen dadas por:

$$f(x) = \lambda e^{-\lambda x} \qquad \forall x \in \mathbb{R}^+$$

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \sigma^2 = \frac{1}{\lambda^2}$$

Distribución normal

La distribución de probabilidad continua más importante en todo el campo de la estadística es la distribución normal. Su representación gráfica, denominada **curva normal**, es la curva con forma de campana, la cual describe aproximadamente el fenómeno que se presenta en la naturaleza, en la industria y en la investigación. Las medidas físicas en áreas tales como los experimentos meteorológicos o los estudios pluviales, y las medidas de partes manufacturadas suelen quedar mejor explicadas utilizando una distribución normal.

Si una VAC X tiene una distribución normal con parámetros μ y σ $(X \hookrightarrow N(\mu, \sigma))$ o $(N(\mu, \sigma^2))$, donde $-\infty < \mu < \infty$ y $0 < \sigma$, su función de densidad, su media y su varianza vienen dadas por:

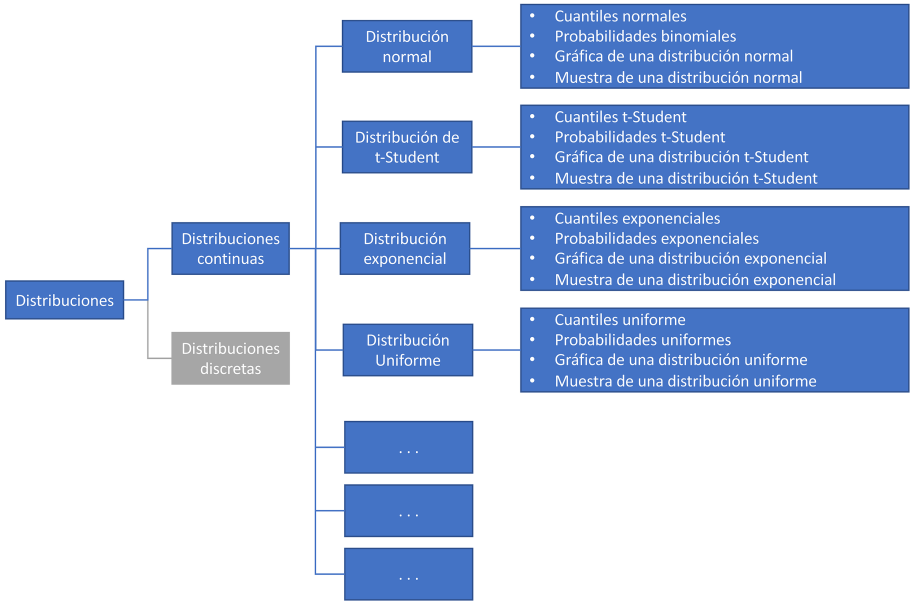
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \forall x \in \mathbb{R}$$

$$E(X) = \mu, \qquad V(X) = \sigma^2$$

Cada curva de densidad es simétrica con respecto a μ y tiene forma de campana, de modo que el centro de la campana (punto de simetría) es a la vez la media y la mediana de la distribución. El valor de σ es la distancia desde μ hasta los puntos de inflexión de la curva (los puntos en que la curva pasa de ser cóncava hacia arriba a ser cóncava hacia abajo). Los valores altos de σ extienden la forma de la curva a lo largo de μ, mientras que los valores pequeños de σ producen curvas con un alto pico en μ.

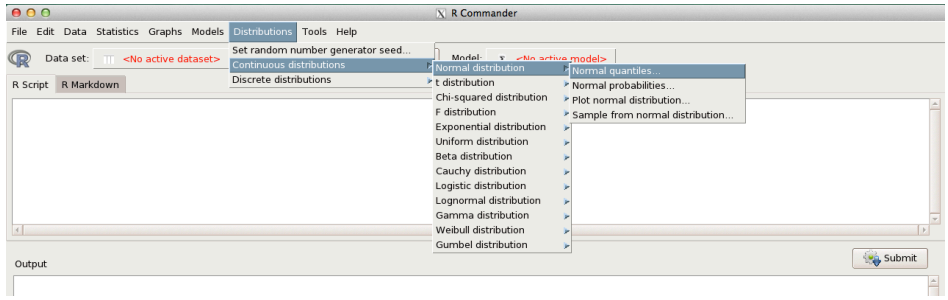
Dada una función de distribución continua comúnmente utilizada en ingeniería (uniforme, exponencial, normal, t-Student, etc.), con sus respectivos parámetros, aplicando **R** y/o **R-Commander** se pueden obtener muy fácilmente la probabilidad de un intervalo, los cuantiles y los gráficos de distribuciones estadísticas estándar (que pueden usarse, por ejemplo, como sustituto de las tablas estadísticas); además, se pueden generar muestreos o simulaciones de estas distribuciones.

Si se dispone de **R-Commander**, en la barra superior podemos encontrar en la ruta *Distribuciones > Distribuciones continuas* las distribuciones más utilizadas. El árbol del submenú completo para distribuciones de probabilidad discretas se muestra en la siguiente figura. La mayoría de opciones del menú nos llevan a diferentes cuadros de diálogo. Las opciones del menú están inactivas (en gris) si no se pueden aplicar al contexto actual.



Si no se dispone de **R-Commander**, o si se prefiere usar la **R-Console** o el **Rstudio**, o simplemente se está creando un *script* para su posterior uso, también se indicarán cuáles son las instrucciones adecuadas para cada caso. Por otra parte, para no ser repetitivos, en esta guía solo se explican todas las opciones concernientes a la distribución normal; las demás distribuciones tienen las mismas opciones, con la diferencia de que los parámetros son diferentes para cada caso. Por ejemplo, la distribución normal tiene como parámetros μ y σ, mientras que la distribución exponencial solo tiene λ.

Dada una VAC X que sigue una distribución normal de parámetros μ y σ, es decir $X \hookrightarrow N(\mu, \sigma^2)$, se pueden calcular la función de densidad y probabilidades entre dos valores de X, realizar los gráficos de las funciones de densidad y de distribución, calcular los cuantiles normales y realizar un muestreo o simulación de experimentos normales, tal como se muestra en la figura.

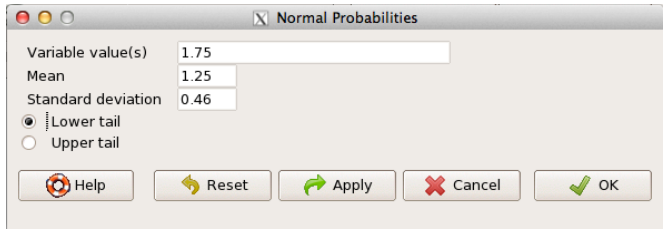


5.3.1. Probabilidades

Calcula la probabilidad de que la variable X sea como máximo x, $P(X \leq x)$.

Ejemplo: El tiempo de reacción en tráfico de una señal de freno (luces de freno) puede modelizarse como una distribución normal con una media de 1.25 segundos y una desviación típica de 0.46 segundos. ¿Cuál es la probabilidad de que el tiempo de reacción sea inferior a 1.75 segundos?

Si definimos la VAC X que denota el tiempo de reacción de las luces de freno, entonces $X \hookrightarrow N(1.25, 0.46^2)$. Por tanto, $P(X \leq 1.75)$ se obtiene al introducir los siguientes parámetros en el cuadro de diálogo que se genera al ejecutar *Distribuciones -> Distribuciones continuas -> Distribución normal -> Probabilidades normales.*.



En la ventana de instrucciones, se genera el siguiente código, con su respectiva respuesta en la ventana de resultados:

```
pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=TRUE)
```

```
## [1] 0.861472
```

Por tanto, la probabilidad de que el tiempo sea inferior a 1.75 segunsdos es $P(X \leq 1.75) = 0.861472$. Nótese que la probabilidad se calcula mediante la función pnorm(x, mean=mu, sd=sigma, lower.tail=TRUE), donde x es el valor de que se quiere calcular la probabilidad acumulada, mu es la media, sigma es la desviación típica y lower.tail=TRUE indica que se calculará la probabilidad de la cola izquierda. De esta manera, si se prefiere usar directamente los comandos en la **R-Console** o **Rstudio**, se puede ejecutar la instrucción dada.

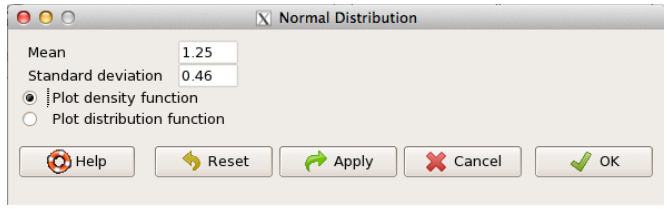
Además, $P(X > x)$ también se puede calcular seleccionando la opción "Cola derecha" en lugar de "Cola izquierda" o cambiando la opción lower.tail=TRUE. De esta forma, continuando con el ejemplo, $P(X > 1.75) = P(X \geq 1.75) = 1 - P(X \leq 1.75)$

```
pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=FALSE)
```

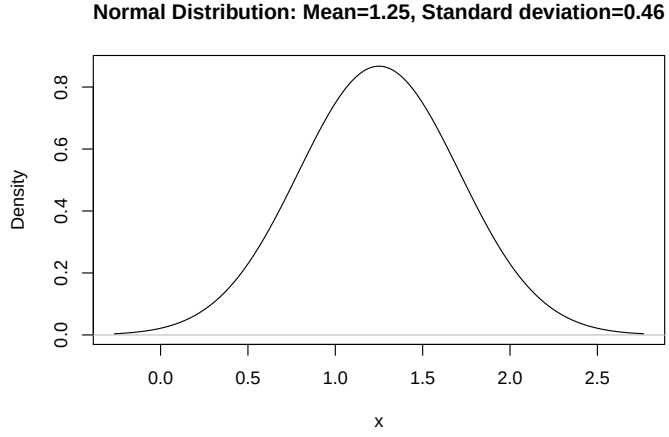
```
## [1] 0.138528
```

5.3.2. Gráfica de una distribución

Esta opción permite generar la representación gráfica de la función de densidad o la función de distribución de la distribución dada. La función de densidad del ejemplo se genera de la siguiente manera:

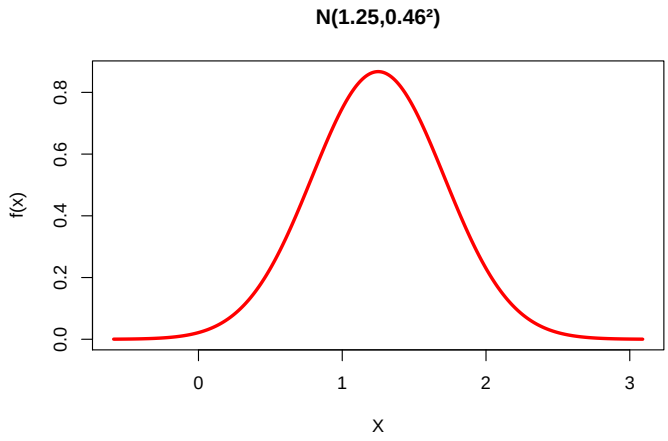



```
local({ .x <- seq(-0.264, 2.764, length.out=1000) plotDistr(.x, dnorm(.x,
mean=1.25, sd=0.46), cdf=FALSE, xlab="x", ylab="Density", main=paste("Normal
Distribution: Mean=1.25, Standard deviation=0.46")) })
```



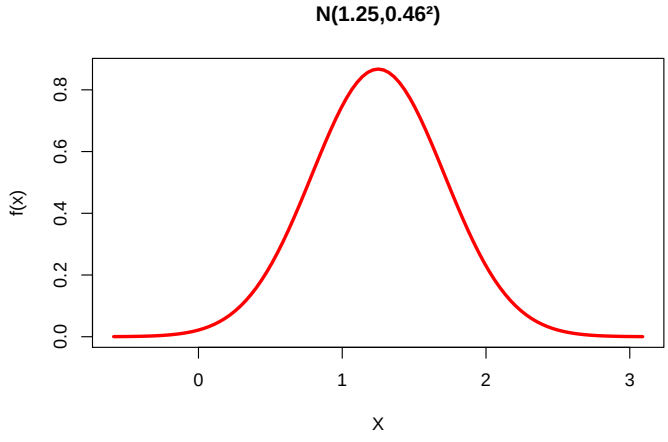
Nótese que los valores de la función de densidad ($f(x)$) se calculan mediante la función dnorm(x, mean=mu, sd=sigma) donde x es el vector de valores de X del cual se quiere calcular la función de densidad. Por tanto, para hacer la gráfica directamente desde **R-Console** o **Rstudio** se pueden utilizar las intrucciones explicadas al principio de la guía de la siguiente forma:

```
mu = 1.25 sigma = 0.46
x = seq(mu-4*sigma, mu+4*sigma, length=1000) Posibles resultados
f = dnorm(x, mean=mu, sd=sigma)  función de densidad
plot(x, f, type="l", col=red", lwd=3, main="N(1.25,0.46²)",
xlab="X", ylab="f(x)")
```



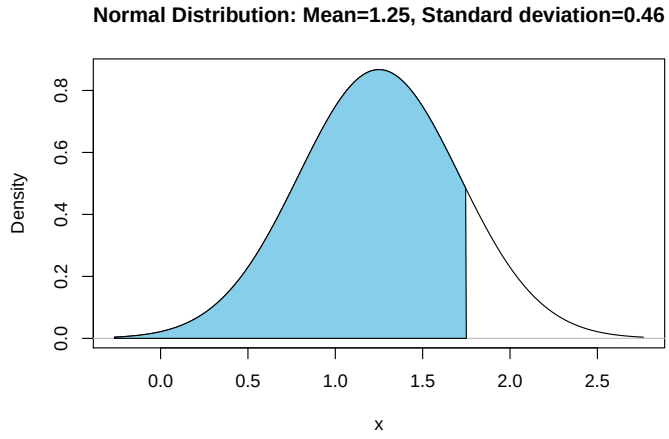
O utilizando la función curve()

```
mu = 1.25
sigma = 0.46
curve(dnorm(x, mean=mu, sd=sigma), xlim =c(mu-4*sigma, mu+4*sigma),
col=red", lwd=3, main="N(1.25,0.46²)", xlab="X", ylab="f(x)")
```

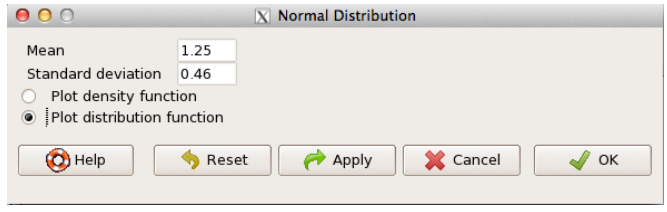


Independientemente de cómo se ha generado la gráfica de la función de densidad, es posible incluir o resaltar alguna probabilidad dentro de la gráfica (activa). Continuando con el ejemplo de cuál es la probabilidad de que el tiempo de reacción sea inferior a 1.75 segundos, se puede representar agregando el siguiente código:

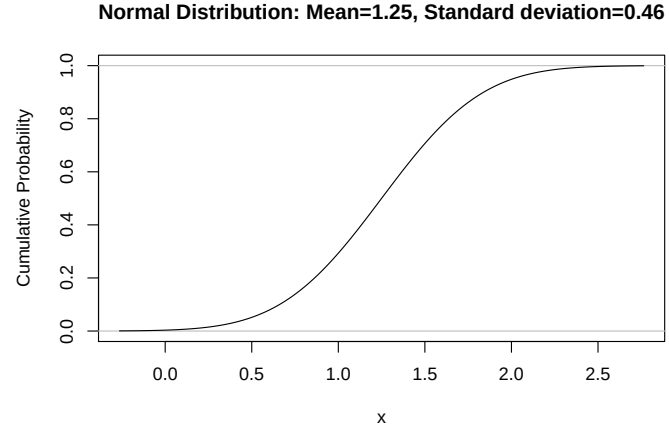
```
cord.x=c(-0.264, seq(-0.264,1.75,0.01),1.75)
Vector de vértices en x para el polígono
cord.y=c(0, dnorm(seq(-0.264,1.75,0.01),1.25,0.46),0)
Vector de vértices en y
polygon(cord.x, cord.y, col='skyblue')
```



Por otra parte, la gráfica de la función de distribución se puede generar al introducir los siguientes parámetros en el cuadro de diálogo.

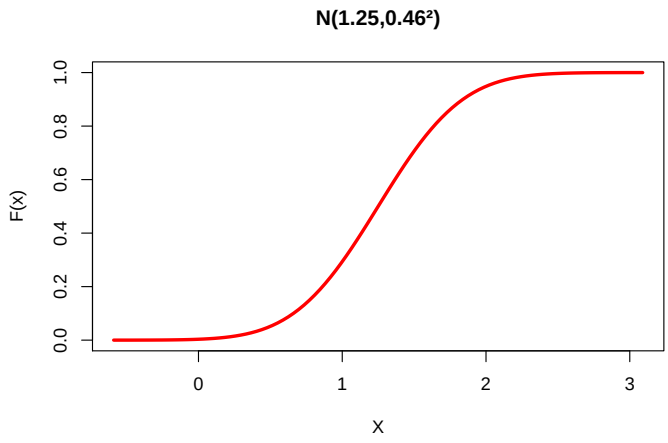



```
local({  .x <- seq(-0.264, 2.764, length.out=1000)
plotDistr(.x, pnorm(.x, mean=1.25, sd=0.46),
cdf=TRUE, xlab="x", ylab=Çumulative Probability",
main=paste("Normal Distribution: Mean=1.25, Standard deviation=0.46"))
})
```



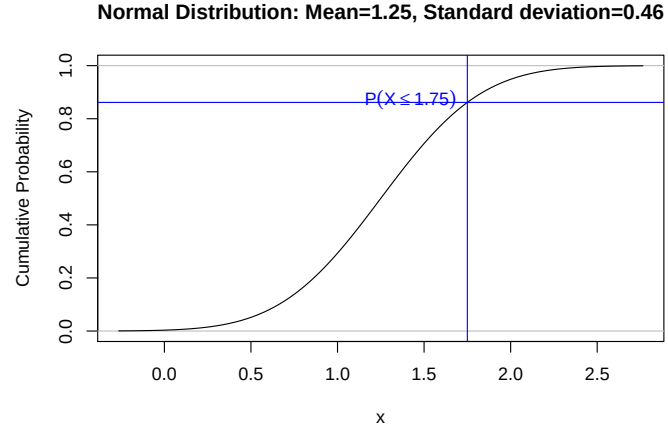
Igualmente, sin necesidad de usar el **R-commander**, mediante la función curve() se puede generar la gráfica de la distribución con la opción de personalizarla completamente:

```
mu = 1.25; sigma = 0.46 curve(pnorm(x, mean=mu, sd=sigma),
xlim = c(mu-4*sigma, mu+4*sigma), col=red", lwd=3,
main="N(1.25,0.46²)", xlab="X", ylab="F(x)")
```



Análogamente, se puede representar la probabilidad de que el tiempo de reacción sea inferior a 1.75 segundos:

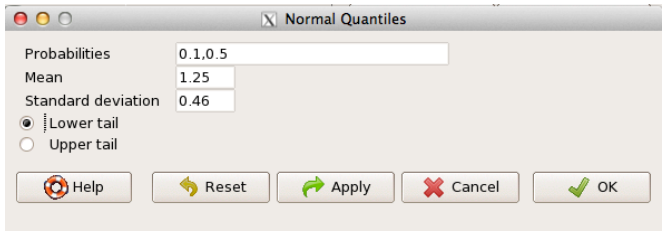
```
x = 1.75 F_x = pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=TRUE)
abline(v=x, col="blue");
abline(h=F_x, col="blue")
text(x, F_x,expression(P(X<=1.75)), pos=2, col="blue")
```



5.3.3. Cuantiles

Esta función calcula la "función inversa" de la función de distribución dada. El p-ésimo cuantil se define como el valor x tal que $F(x) = P(X \leq x) = p$. Este se calcula seleccionando la opción "Cola izquierda". También se puede calcular el valor de x tal que $P(X > x) \geq p$ seleccionando la opción "Cola derecha". En el ejemplo dado, se determinan el décimo cuantil, la mediana y el valor de x tal que $P(X > x) = 0.15$.

El décimo cuantil y la mediana indican el valor de x tal que $P(X \leq x) = 0.1$ y $P(X \leq x) = 0.5$, respectivamente. Por tanto, si seleccionamos la primera opción del submenú "Distribución normal", nos aparece la siguiente ventana:




```
qnorm(c(0.1,0.5), mean=1.25, sd=0.46, lower.tail=TRUE)
```

```
## [1] 0.6604863 1.2500000
```

Entonces, el décimo cuantil es 0.66 ya que $P(X \leq 0.66) = 0.1$ y, como cabría esperar, la mediana es igual a la media, ya que $P(X \leq 1.25) = 0.5$. Para calcular el valor de x tal que $P(X > x) = 0.15$, el parámetro "probabilidades" debe cambirse a 0.15 y seleccionar "cola derecha", y se generan las siguientes instrucciones y su resultado:

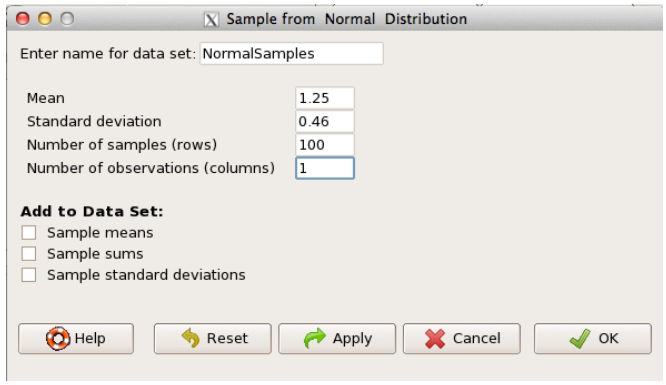
```
qnorm(c(0.15), mean=1.25, sd=0.46, lower.tail=FALSE)
```

```
## [1] 1.726759
```

lo cual indica que $P(X > 1.727) = 0.15$.

5.3.4. Muestreo

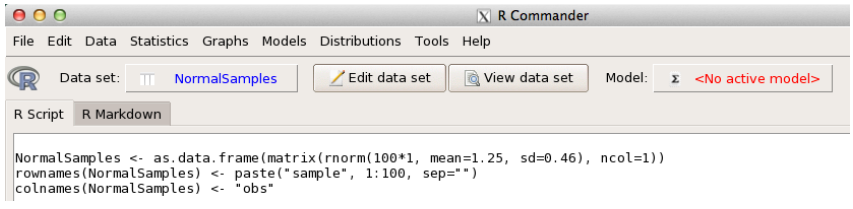
Simula escenarios aleatorios para la distribución dada. Con el ejemplo previo, si queremos simular una observación de 100 muestras, en las cuales cada muestra consista en una carga dinámica sobre el puente, se deben introducir los siguientes parámetros en el cuadro de diálogo:



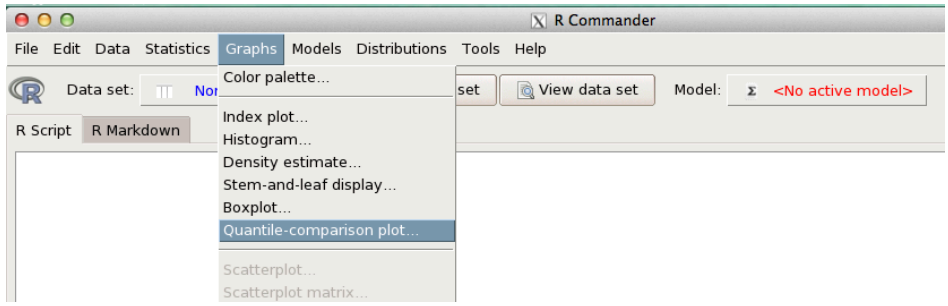
También se pueden añadir algunas medidas de los datos simulados, como la media, la sumatoria y la desviación típica:

```
NormalSamples <- as.data.frame(matrix(rnorm(100*1, mean=1.25,
sd=0.46), ncol=1))
rownames(NormalSamples) <- paste("sample", 1:100, sep=)
colnames(NormalSamples) <- .ºbs"
```

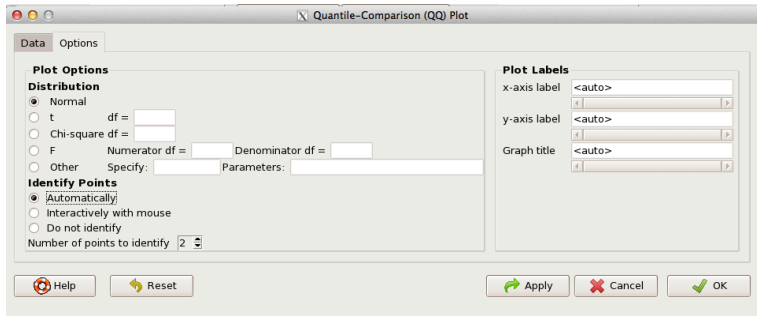
Una vez hemos simulado los datos, estos ya aparecen en el conjunto de datos del **R Commander**. En consecuencia, el nombre del conjunto de datos aparece en el botón que hay junto al extremo izquierdo en la ventana principal. Los datos simulados se guardan en el vector denominado NormalSamples y se pueden ver haciendo clic en *Visualizar conjunto de datos/View data set*.



El conjunto de datos simulados se puede mostrar a través del índice de gráficos, histograma, etc. (v. figura siguiente). Un importante método para comparar si los datos simulados siguen una distribución normal es utilizar la gráfica de comparación de cuantiles (QQ en inglés), un método gráfico para comparar dos distribuciones de probabilidad mostrando sus cuantiles, uno frente al otro. Si las dos distribuciones comparadas son similares, los puntos en el gráfico QQ forman una línea $y = x$, aproximadamente. Si las distribuciones son linealmente dependientes, los puntos en el gráfico QQ forman una línea, pero no necesariamente $y = x$.

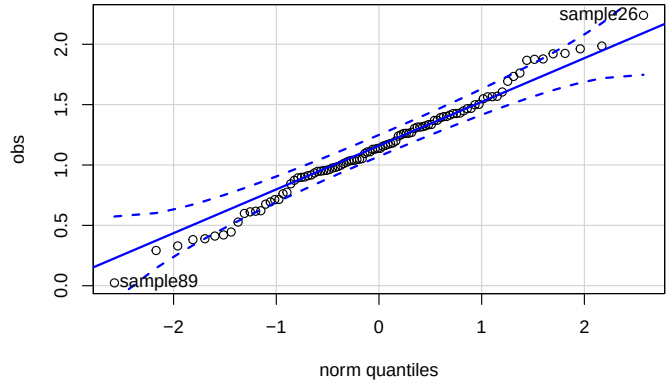


Para comparar los datos simulados con una distribución normal, hemos de seleccionar en "Opciones" el parámetro "Normal", como vemos a continuación:



Aceptando, se generan los siguientes comandos y salida:

```
with(NormalSamples, qqPlot(obs, dist="norm", id=list(method="y", n=2,
labels=rownames(NormalSamples))))
```




```
## sample89 sample26
##       89       26
```

Tips & Tricks!

En caso de no disponer o hacer uso del **R-Commander**, se pueden introducir directamente en **R-Console** o **Rstudio** las siguientes funciones:

- dnorm() calcula la función de distribución de una distribución normal, es decir, $f(x)$ para todos los posibles valores de x.
- pnorm() calcula la probabilidad acumulada de una distribución normal, es decir $F(x) = P(X \leq x)$ o $1 - F(x) = P(X > x)$ para un valor de x dado.
- qnorm() calcula el p-ésimo cuantil de una distribución normal, es decir, el valor de x tal que $F(x) = P\left(X \leq x\right) \geq p$ para un valor de p dado.
- rnorm() realiza un muestreo o simulación de un número determinado de experimentos de una distribución normal.

Nótese que la primera letra indica: función de distribución (d), probabilidad (p), cuantil (q) o muestra aleatoria (r). Las demás letras indican la distribución: normal (normal), exponencial (exp), t-Student (t), etc. De esta manera, si lo que se quiere es calcular un cuantil de una distribución de t-Student, la función correspondiente es qt(), para la distribución de exponencial es qexp(), etc.
Otra forma representar gráficamente una curva continua definida entre dos valores es mediante la función curve(expr,from=a,to=b), donde expr es el nombre de una función o una expresión escrita como función de x, por ejemplo pnorm(x,mean,sd.

5.4. Ejercicios propuestos

1. En una ciudad, se estima que la temperatura máxima en el mes de junio sigue una distribución normal, con una media de 23^{o}C y una desviación típica 5^{o}C.

   a) Calcula la probabilidad de que un día cualquiera la temperatura máxima se encuentre entre 21^{o}C y 27^{o}C.

   b) Representa gráficamente la función de distribución y la probabilidad anteriores.

   c) Simula una muestra de 200 observaciones y compáralas con la función de distribución.

2. Se supone que los resultados de un examen siguen una distribución normal con una media de 78 y una varianza de 36.

   a) ¿Cuál es la probabilidad de que una persona que se presenta al examen obtenga una calificación superior a 72?

   b) Representa gráficamente la función de distribución y la probabilidad anteriores.

   c) Simula una muestra de 10 observaciones y compáralas con la función de distribución.

3. Consideremos X la variable aleatoria cuya función de densidad viene dada por:

$$f(x) = \begin{cases} \dfrac{4}{\pi(1+x^2)} & 0 \leq x \leq 1 \\ 0 & \text{resto} \end{cases}$$

   a) Calcular la probabilidad de que X esté entre 0.4 y 0.6.

   b) Representa gráficamente la función de distribución y la probabilidad anteriores.

   c) Simula una muestra de 100 observaciones y compáralas con la función de distribución.


2
1
3

Muestreo y Teorema del límite central

6.1. Introducción y objetivos

Los resistores que se utilizan en la fabricación de productos electrónicos están etiquetados con una resistencia "nominal" y con un porcentaje de tolerancia. Por ejemplo, se prevé que una resistencia de 330 ohmios (Ω) con una tolerancia del 5 $\%$ tendrá una resistencia real R de entre 313.5 (Ω) y 346.5 (Ω) uniformemente distribuida.

Si se consideran cinco resistencias de este tipo, seleccionadas al azar de la población de todos los resistores con esas especificaciones, que están conectados en serie, la resistencia total R_T del sistema viene dada por $R_T = R_1 + R_2 + ... + R_5$, donde R_i son los valores de las resistencias, valores que son aleatorios, independientes e idénticamente distribuidos uniformemente. Por tanto, la resistencia del sistema R_T también es una variable aleatoria que tiene asociado un valor esperado $E(R_T)$, una varianza $V(R_T)$ y una función de densidad $f(R_T)$. Pero, ¿cuáles son? ¿Cómo se calculan?, ¿Es R_T también uniformemente distribuida? ¿Qué sucede si en lugar de 5 tuviésemos un número suficientemente grande de resistencias conectadas en serie?, ¿Qué cambiaría si el valor real de las resistencias R no estuviera distribuido uniformemente sino normalmente?

En esta sesión, empezaremos formalizando el propósito de la inferencia estadística, concretamente cuál es el comportamiento de la suma, la media y la varianza de una muestra de una variable aleatoria; en otras palabras, el muestreo aleatorio y la aplicación de la teoría de distribución de las muestras. De esta forma, en las siguientes sesiones se discutirá el problema de la estimación de los parámetros de la población y de las pruebas de contraste de hipótesis a partir de una muestra dada.

Al finalizar la sesión, el alumno ha de ser capaz de:

- Entender el comportamiento de la suma y/o promedio de una muestra aleatoria
- Comprender el teorema del límite central y sus aplicaciones en la estimación de parámetros.

– Simular una muestra aleatoria de una población dada y representar los resultados mediante **R**.

– Comprobar por medio de simulaciones los teoremas de muestreo y del límite central.

6.2. Muestreo

Uno de los objetivos más importantes de la estadística es la inferencia estadística. Hacer inferencia sobre algo significa sacar conclusiones de ello a partir del razonamiento y la evidencia. Así pues, la inferencia estadística se puede definir como el conjunto de teorías, métodos y prácticas para formular juicios sobre los parámetros de una población a partir de sus relaciones estadísticas, basadas en una muestra representativa de dicha población. En otras palabras, la inferencia estadística utiliza una muestra para conocer algo relacionado con una población mucho mayor. Ya que la inferencia se basa en las muestras, es conveniente estudiar primero cuál es el comportamiento de la muestra y qué relación tiene con la población.

6.2.1. Muestra aleatoria

Dada una población con variable aleatoria X (discreta o continua), una *muestra aleatoria* es un conjunto de valores o datos aleatorios, independientes e idénticamente distribuidos, obtenidos a partir de la variable aleatoria X y que se distribuyen igual que esta.

Por ejemplo, $X_1, X_2, X_3, ..., X_n$ son muestras aleatoria de una variable aleatoria X que se distribuye normalmente con una media de 100 y una desviación típica de 15, si $X_1, X_2, X_3, ..., X_n$ son independientes y cada una tiene una distribución normal con una media 100 y una desviación típica 15. Similarmente, serán muestras aleatorias de una distribución exponencial con $\lambda = 12$ si son independientes y cada una de ellas es exponencial con el mismo valor de λ.

Consideremos una población con una variable aleatoria X cuya función de densidad es $f_X(x)$, su valor esperado es $E(X)$ o μ_X y su varianza es $V(X)$ o σ_X^2. Sea $x_1 = \{x_{1,1}, x_{1,2}, ..., x_{1,n}\}$ una muestra aleatoria de tamaño n de dicha variable X, donde $x_{1,1}$ es el valor de la variable del primer individuo u objeto seleccionado, $x_{1,2}$ es el valor de la misma variable para el segundo individuo u objeto, etc. De esta muestra, se puede visualizar la tabla de frecuencias (o histograma); también se pueden calcular algunos estadísticos, tales como la suma de los elementos t_1, la media muestral $\bar{x}_1$ y la varianza muestral s_1^2, entre otros, donde:

$$t_1 = \sum_{i=1}^{n} x_{1,i}, \qquad \bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1,i} = \frac{t_1}{n}, \qquad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{1,i} - \bar{x}_1\right)^2.$$

Ejemplo 1: En la introducción, se menciona que la resistencia real de unos resistores etiquetados como de 330 Ω se puede considerar una variable aleatoria R que se distribute uniformemente con una media de 330 Ω entre 313.5 Ω y 346.5 Ω. Por lo tanto, $\mu_R = 330$ y $\sigma_R^2 = (346.5 - 313.5)^2/12 = 90.75$.

Ejemplo 2: La cantidad de tiempo que un paciente que se somete a un procedimiento especial pasa en un determinado centro de cirugía ambulatoria es una variable aleatoria W que se distribuye normalmente con un valor medio de 4.5 h y una desviación típica de 1.4 h. Por tanto: $\mu_W = 4.5$ y $\sigma_W^2 = 1.4^2 = 1.96$.

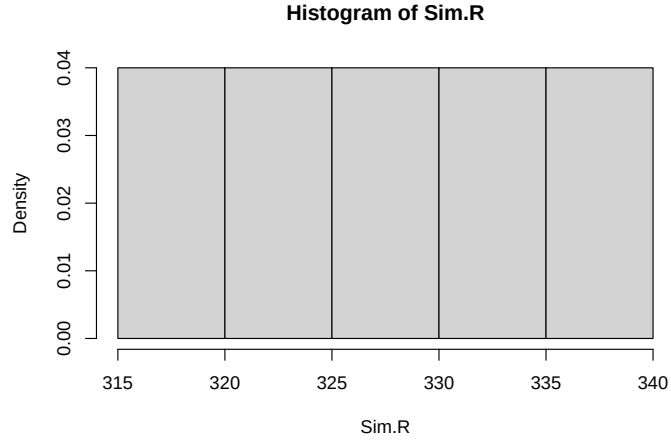
Recordemos de la sesión anterior que, para simular eventos o generar muestras de una población con una función de densidad dada, en *R* se utiliza la función sample(). Si la distribución es de las más utilizadas en ingeniería, existen funciones específicas tales como rnorm(), rbinom(), runif(), entre otras.

A continuación, generaremos una muestra (simular eventos) para cada uno de los ejemplos fijando primero la semilla para reproducir los resultados. Adicionalmente, visualizaremos el histograma y calcularemos la suma de los elementos de la muestra t, la media muestral $\bar{x}$ y la varianza muestral s^2 de cada muestra. Para el ejemplo 1, se simulará la selección aleatoria de 5 resistencias.

```
n.R = 5  Tamaño de la muestra
set.seed(10)  Fijación de la semilla de la aleatoriedad
Sim.R = runif(n.R, min=313.5, max=346.5);
Sim.R  Muestra en el ejemplo 1
```

```
## [1] 330.2468 323.6234 327.5880 336.3724 316.3095
```

```
hist(Sim.R,prob=T)  Histograma
```




```
sum.R = sum(Sim.R); sum.R  Suma de las observaciones de la muestra
```

```
## [1] 1634.14
```

```
mean.R = mean(Sim.R); mean.R  Media de la muestra
```

```
## [1] 326.828
```

```
var.R = var(Sim.R); var.R  Varianza de la muestra
```

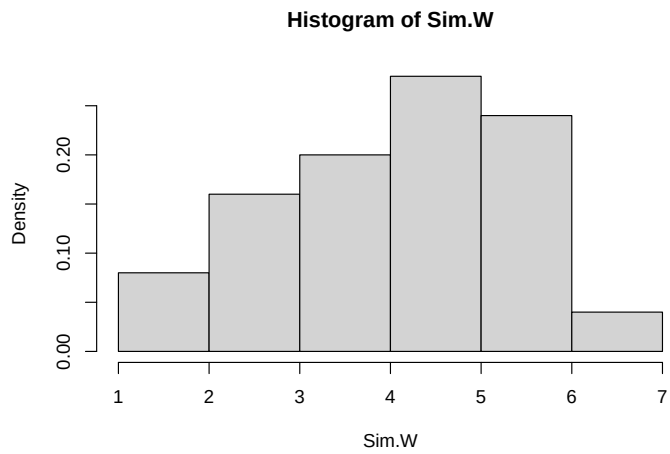
```
## [1] 56.06735
```

Para el ejemplo 2, se simula el tiempo de 25 pacientes.

```
n.W = 25  Tamaño de la muestraset.seed(10)  Fijación de la semilla de
la aleatoriedadSim.W = rnorm(n.W, mean=4.5, sd=1.4); Sim.W  Muestra
en el ejemplo 2
```

```
##  [1] 4.526245 4.242046 2.580137 3.661165 4.912363 5.045712
       2.808693 3.990854
##  [9] 2.222658 4.140930 6.042491 5.558094 4.166473 5.882423
       5.537946 4.625086
## [17] 3.163079 4.226789 5.795730 5.176170 3.665165 1.440598
       3.555188 1.533314
## [25] 2.728723
```

```
hist(Sim.W,prob=T)  Histograma
```




```
sum.W = sum(Sim.W); sum.W  Suma de las observaciones de la muestra
```

```
## [1] 101.2281
```

```
mean.W = mean(Sim.W); mean.W  Varianza de la muestra
```

```
## [1] 4.049123
```

```
var.W = var(Sim.W); var.W
```

```
## [1] 1.740741
```

¿Ha notado alguna relación entre las medias muestrales, las varianzas muestrales y los histogramas con las medias poblacionales, las varianzas poblacionales y las funciones de densidad de la población? ¿Qué pasa si se aumenta el tamaño de la muestra?

Si realizamos otra muestra $x_2 = \{x_{2,1}, x_{2,2}, ..., x_{2,n}\}$, esta tendrá también un histograma, t_2, $\bar{t_2}$, s_2^2, etc. Debido a la aleatoriedad del muestreo, estos histogramas no han de ser idénticos, como tampoco las sumas, ni las medias, ni las varianzas muestrales. De hecho, $(t_1, t_2, ...)$, $(\bar{x}_1, \bar{x}_2, ...)$ y $(s_1^2, s_2^2, ...)$ se pueden considerar valores de las nuevas variables aleatorias T, $\bar{X}$ y S^2, respectivamente.

Como T, $\bar{X}$ y S^2 son variables aleatorias continuas, han de tener:

- Función de densidad: $f_T(x)$, $f_{\bar{X}}(x)$, $f_{S^2}(x)$
- Valor esperado: $E(T)$, $E(\bar{X})$, $E(S^2)$
- Varianza: $V(T)$, $V(\bar{X})$, $V(S^2)$

Pero, ¿cuáles son? y ¿qué relación tienen con la distribución y los parámetros de la población? Para resolver estos interrogantes, estudiaremos la distribución de la suma muestral (T), de la media muestral $\bar{X}$ y de la varianza muestral S^2. Para ello, se simulará no solo una muestra (como hasta ahora), sino muchas muestras (N=300), y se almacenarán en un data.frame donde cada fila representa una muestra y las columnas, las observaciones.

```
Ejemplo 1
N = 300  Número de muestras
n.R = 5  Tamaño de la muestra
min.R = 313.5  Parámetros de la población
max.R = 346.5 set.seed(10) Fijación de la semilla de la aleatoriedad
samples = runif(N*n.R, min=min.R, max=max.R)
Simulación de N * nr muestras
samples.R = as.data.frame(matrix(samples, ncol=n.R))
Organización en un data.frame
```

```
Ejemplo 2
N = 300  Número de muestras
n.W = 25  Tamaño de la muestra
mean.W = 4.5  Parámetros de la población
sd.W = 1.4 set.seed(10)  Fijación de la semilla de la aleatoriedad
samples = rnorm(N*n.W, mean=mean.W, sd=sd.W)
Simulación de N * nr muestras
samples.W = as.data.frame(matrix(samples, ncol=n.W))
Organización en un data.frame
```

6.2.2. Distribución de la suma muestral

Sea $X_1, X_2, ..., X_n$ la muestra aleatoria de tamaño n de una población X cuya función de densidad es $f_X(x)$, su valor esperado es $E(X) = \mu_X$ y su varianza es $V(X) = \sigma_X^2$. De la suma de los elementos de la muestra $T = X_1 + X_2 +, ..., +X_n$, se puede deducir que:

- Su valor esperado viene dado por:

$$\begin{aligned} E(T) &= E\left(X_1 + X_2 +, ..., +X_n\right) = E(X_1) + E(X_2) +, ..., +E(X_n) \\ &= E(X) + E(X) +, ..., +E(X) = n\mu_X \end{aligned}$$

- Su varianza viene dada por:

$$\begin{aligned} V(T) &= V\left(X_1 + X_2 +, ..., +X_n\right) = V(X_1) + V(X_2) +, ..., +V(X_n) \\ &= V(X) + V(X) +, ..., +V(X) = n\sigma_X^2 \end{aligned}$$

- Si las X_i están distribuidas normalmente, entonces T también está distribuida normalmente, es decir:

$$X \hookrightarrow N(\mu_X, \sigma_X^2) \qquad \Rightarrow \qquad T \hookrightarrow N(n\mu_X, n\sigma_X^2)$$

Volviendo al ejemplo 1, la resistencia total de la conexión en serie de 5 resistencias seleccionadas de forma aleatoria $R_T = R_1 + R_2 + ... + R_5$, tiene como valor esperado $E(R_T) = n\mu_R = 5 \times 330 = 1650\ \Omega$ y su varianza es $V(R_T) = n\sigma_R^2 = 5 \times 90.75 = 453.75\ \Omega^2$. Finalmente, como R no está distribuida normalmente, no podemos asegurar cuál es la distribución de R_T.

Para cada muestra realizada previamente, se calcula la suma de las 5 observaciones y se almacenan en el vector sum.samples.R (suma muestral), y a este vector se le calcula la media, la varianza, y se visualiza el histograma.

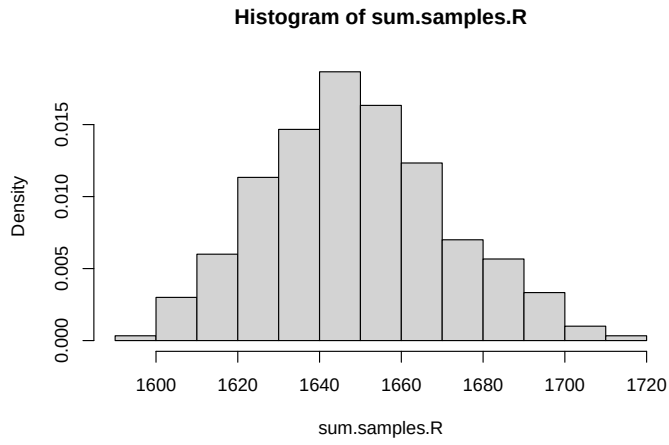
```
sum.samples.R = apply(samples.R,1,sum)
Calcula la suma de cada muestra (fila)
mean(sum.samples.R)  Media de la suma muestral
```

```
## [1] 1649.235
```

```
var(sum.samples.R)  Varianza de la suma muestral
```

```
## [1] 499.0941
```

```
hist(sum.samples.R,prob=T)  Histograma de la suma muestral
```



Por otra parte, el tiempo total que 25 pacientes permanecen en un determinado centro de cirugia $W_T = W_1 + W_2 + ... + W_{25}$, tiene como valor esperado $E(W_T) = n\mu_W = 25 \times 4.5 = 112.5\ s$ y su varianza es $V(W_T) = n\sigma^2_W = 25 \times 1.96 = 49\ s^2$. Finalmente, como W está distribuida normalmente, W_T también se distribuye normalmente.

La suma de las 25 observaciones se almacena en el vector sum.samples.W (suma muestral), y de este vector se calculan la media y la varianza. Finalmente, se visualiza el histograma y la función de densidad de una normal con una media 112.5 ($n\mu_W$) y una varianza 49 ($n\sigma^2_W$) o una desviación típica de 7 ($\sqrt{n}\sigma_W$).

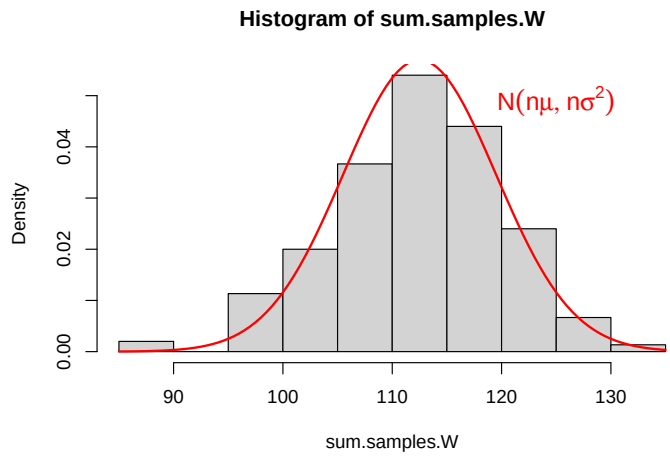
```
sum.samples.W = apply(samples.W,1,sum)
Calcula la suma de cada muestra (fila)
mean(sum.samples.W)  Media de la suma muestral
```

```
## [1] 112.3905
```

```
var(sum.samples.W)  Varianza de la suma muestral
```

```
## [1] 59.88336
```

```
hist(sum.samples.W,prob=T)  Histograma de la suma muestral
curve(dnorm(x,mean=n.W*mean.W,sd=sqrt(n.W)*sd.W),
add=T, lwd=2, col=red")
text(125,0.049,expression(N(n*mu,n*sigma^2)),col=red",cex=1.3)
```



6.2.3. Distribución de la media muestral

Continuando con la muestra $X_1, X_2, ..., X_n$ de una población X, y teniendo en cuenta los resultados del apartado anterior, de la media muestral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n},$$

se puede deducir que:

- Su valor esperado viene dado por:

$$E(\bar{X}) = E\left(\frac{T}{n}\right) = \frac{1}{n}E(T) = \frac{1}{n}n\mu_X = \mu_X$$

- Su varianza viene dada por:

$$V(\bar{X}) = V\left(\frac{T}{n}\right) = \frac{1}{n^2}V(T) = \frac{1}{n^2}n\sigma_X^2 = \frac{\sigma_X^2}{n}$$

– Si las X_i están distribuidas normalmente, entonces $\bar{X}$ también está distribuida normalmente, es decir:

$$X \hookrightarrow N(\mu_X, \sigma_X^2) \qquad \Rightarrow \qquad \bar{X} \hookrightarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

De esta forma, la resistencia media de una muestra de 5 resistencias seleccionadas de forma aleatoria $\bar{R}$ en el ejemplo 1 tiene como valor esperado $E(\bar{R}) = \mu_R = 330\ \Omega$ y varianza $V(\bar{R}) = \sigma_R^2/n = 90.75/5 = 18.15\ \Omega^2$. Finalmente, como R no está distribuida normalmente, no podemos asegurar cuál es la distribución de $\bar{R}$.

Como en el apartado anterior, para cada muestra se calcula la media de las 5 observaciones y se almacenan en el vector mean.samples.R (media muestral); de este vector se calculan la media y la varianza, y se visualiza el histograma.

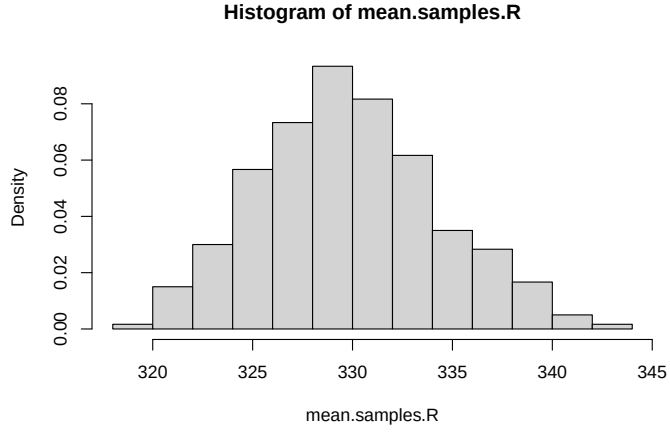
```
mean.samples.R = apply(samples.R,1,mean)
Calcula la media de cada muestra (fila)
mean(mean.samples.R)  Media de la media muestral
```

```
## [1] 329.847
```

```
var(mean.samples.R)  Varianza de la media muestral
```

```
## [1] 19.96376
```

```
hist(mean.samples.R,prob=T)  Histograma de la media muestral
```



Por otra parte, el tiempo medio que 25 pacientes permanecen en un determinado centro de cirugía $\bar{W}$ tiene como valor esperado $E(\bar{W}) = \mu_W = 4.5s$ y su varianza es

$V(\bar{W}) = \frac{\sigma_W^2}{n} = \frac{1.96}{25} = 0.0784s^2$. Finalmente, como W está distribuida normalmente, $\bar{W}$ también se distribuye normalmente.

El promedio de las 25 observaciones se almacena en el vector mean.samples.W (media muestral); de este vector se calculan la media y la varianza. Finalmente, se visualiza el histograma y la función de densidad de una normal con una media de 4.5 (μ_W) y una varianza de 0.0784 (σ_W^2/n) o una desviación típica de 0.28 ($\sigma_W/\sqrt{n}$).

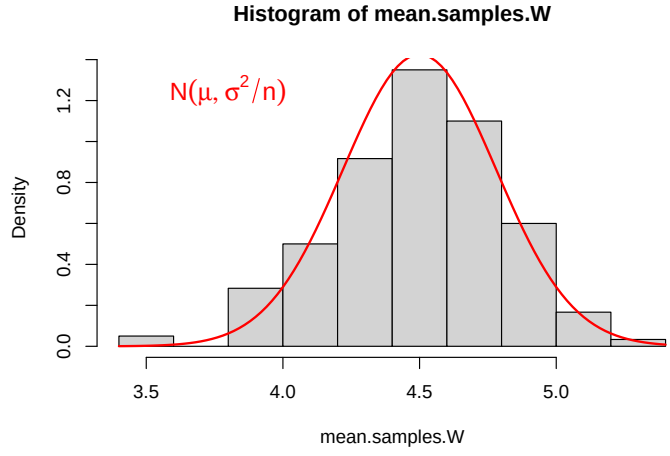
```
mean.samples.W = apply(samples.W,1,mean)
Calcula la media de cada muestra (fila)
mean(mean.samples.W)  Media de la media muestral
```

```
## [1] 4.495621
```

```
var(mean.samples.W)  Varianza de la media muestral
```

```
## [1] 0.09581337
```

```
hist(mean.samples.W,prob=T)  Histograma de la media muestral
curve(dnorm(x,mean=mean.W,sd=sd.W/sqrt(n.W)), add=T, lwd=2, col=red")
text(3.8,1.25,expression(N(mu,sigma^2/n)),col=red",cex=1.3)
```



6.2.4. Distribución de la varianza muestral

Similarmente, considerando la muestra $X_1, X_2, ..., X_n$ de una población X, y teniendo en cuenta los resultados de los apartados anteriores, de la varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2,$$

se puede deducir que:

– Su valor esperado viene dado por:

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right) = \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right)-\sum_{i=1}^{n}\left(2X_i\bar{X}\right)+\sum_{i=1}^{n}\left(\bar{X}^2\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right)-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right)-2n\bar{X}^2+n\bar{X}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right)-n\bar{X}^2\right) \\
&= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}\left(X_i^2\right)\right)-E\left(n\bar{X}^2\right)\right]
\end{aligned}$$

Teniendo en cuenta que $E(X^2)=\sigma_X^2+\mu_X^2$ y $E(\bar{X}^2)=\dfrac{\sigma_X^2}{n}+\mu_X^2$, se tiene:

$$E(S^2)=\frac{1}{n-1}\left[n\sigma_X^2+n\mu_X^2-n\mu_X^2-\sigma_X^2\right]=\sigma_X^2$$

– Cuando el tamaño de la muestra tiende a infinito, su varianza tiende a cero:

$$\lim_{n\to\infty} V(S^2)=0$$

– Si las X_i están distribuidas normalmente, entonces $(n-1)\dfrac{S^2}{\sigma_X^2}$ está distribuida según la *función chi-cuadrado* (χ^2) con $(n-1)$ grados de libertad, es decir:

$$X \hookrightarrow N(\mu_X,\sigma_X^2) \qquad \Rightarrow \qquad (n-1)\frac{S^2}{\sigma_X^2} \hookrightarrow \chi^2_{n-1},$$

Esta distribución, también denominada *distribución de Pearson* o *ji-cuadrada*, es una distribución de probabilidad continua con un parámetro k que representa los grados de libertad de la variable aleatoria y su función de densidad es:

$$f(x;k)=\begin{cases}\dfrac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2} & x>0,\\ 0 & x\leq 0\end{cases}$$

donde Γ es la función gamma.

Teniendo en cuenta el ejemplo 1, la varianza de las resistencias de una muestra de 5 resistencias seleccionadas de forma aleatoria S^2 tiene como valor esperado $E(S^2) = \sigma_R^2 = 90.75\Omega^2$ y, como R no está distribuida normalmente, no podemos asegurar cuál es la distribución de S^2/σ_R^2.

Para cada muestra simulada, se calcula la varianza de las 5 observaciones y se almacenan en el vector var.samples.R (varianza muestral); de este vector se calculan la media y la varianza, y se visualiza el histograma.

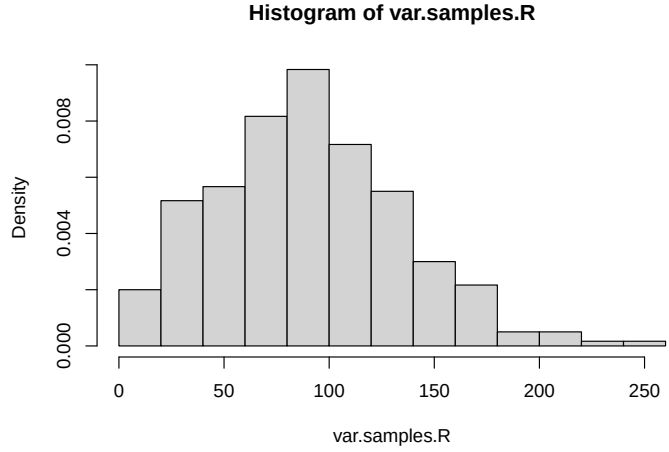
```
var.samples.R = apply(samples.R,1,var)
Calcula la varianza de cada muestra (fila)
mean(var.samples.R)  Media de la varianza muestral
```

```
## [1] 90.39892
```

```
var(var.samples.R)  Varianza de la varianza muestral
```

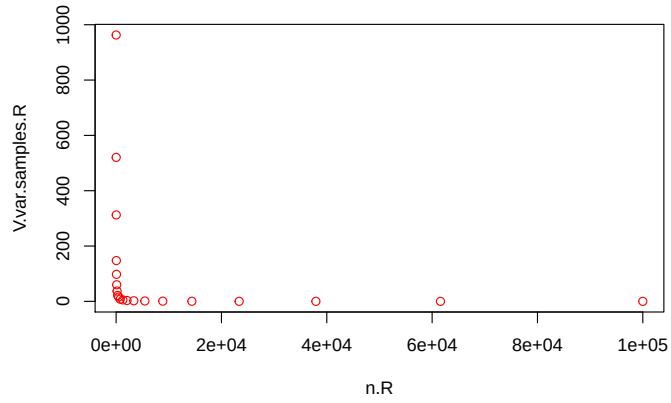
```
## [1] 2014.426
```

```
hist(var.samples.R,prob=T)  Histograma de la varianza muestral
```



Si se desea comprobar qué sucede con la varianza cuando el tamaño de la muestra tiende a infinito, se repite todo el procedimiento anterior para valores de n.R entre 10 y 10^5. En el vector V.var.samples.R, se almacena la varianza de la varianza muestral para cada tamaño de muestra empleado. Finalmente, se visualiza su tendencia en función del tamaño muestral.

```
n.R = round(10^(seq(1,5,length=20)))  Diferentes tamaños de muestra
V.var.samples.R = rep(0,20)  Inicialización del vector con ceros
for (i in 1:20)  Comienzo de las repeticiones{
samples = runif(N*n.R[i], min=min.R, max=max.R)
samples.R = as.data.frame(matrix(samples, ncol=n.R[i]))
var.samples.R = apply(samples.R,1,var)
Calcula la varianza de cada muestra (fila)
V.var.samples.R[i] = var(var.samples.R) }
plot(n.R,V.var.samples.R,type="p",col=red")  Gráfica de la tendencia
```



Con respecto a la varianza del tiempo empleado por los 25 pacientes S^2 del ejemplo 2, se tiene como valor esperado $E(S^2) = \sigma_W^2 = 1.96s^2$.

Para cada muestra simulada, se calcula la varianza de las 25 observaciones y se almacenan en el vector var.samples.W (varianza muestral); de este vector se calculan la media y la varianza, y se visualiza el histograma.

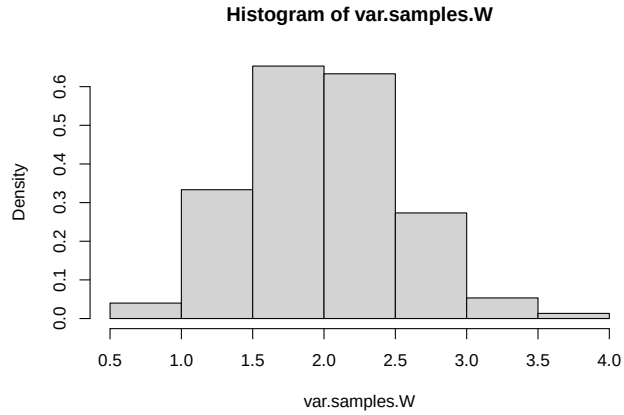
```
var.samples.W = apply(samples.W,1,var)
Calcula la varianza de cada muestra (fila)
mean(var.samples.W)  Media de la varianza muestral
```

```
## [1] 1.99669
```

```
var(var.samples.W)  Varianza de la varianza muestral
```

```
## [1] 0.2928742
```

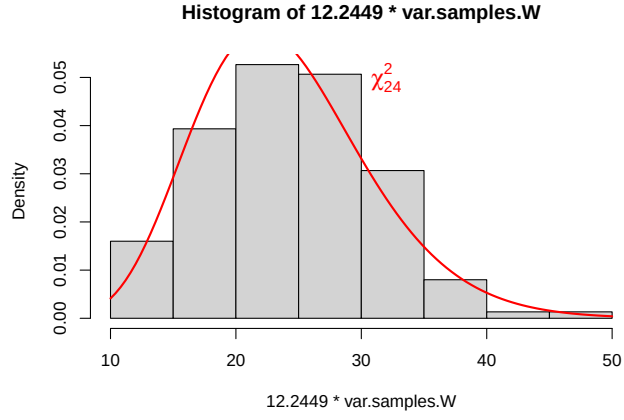
```
hist(var.samples.W,prob=T)  Histograma de la varianza muestral
```



Como W está distribuida normalmente, $(n-1)\dfrac{S^2}{\sigma_W^2}$ está distribuida según la función chi-cuadrado (χ^2) con 24 grados de libertad, es decir, $12.2449S^2 \hookrightarrow \chi^2_{24}$

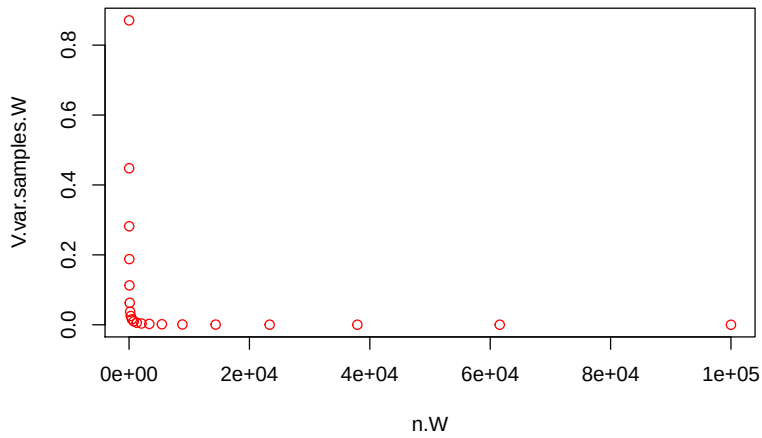
Para corroborarlo, se representa el histograma de la varianza de la varianza muestral y la función de densidad de la función χ^2 con 24 grados de libertad.

```
Histograma de la varianza muestral dividido por la poblacional
y multiplicado por n-1
hist(12.2449*var.samples.W, prob=T)
curve(dchisq(x, df=(n.W-1)), add=T, lwd=2, col=red")
text(32,0.05, expression(chi[24]^2), col=red", cex=1.3)
```



Finalmente, para comprobar qué sucede con la varianza cuando el tamaño de la muestra tiende a infinito, se repite todo el procedimiento anterior para valores de n.W entre 10 y 10^5. En el vector V.var.samples.W, se almacena la varianza de la varianza muestral para cada tamaño de muestra empleado. Finalmente, se visualiza su tendencia en función del tamaño muestral.

```
n.W = round(10^(seq(1,5,length=20)))  Diferentes tamaños de muestra
V.var.samples.W = rep(0,20)  Inicialización del vector con ceros
for (i in 1:20)  Comienzo de las repeticiones{
samples = rnorm(N*n.W[i], mean=mean.W, sd=sd.W)
samples.W = as.data.frame(matrix(samples, ncol=n.W[i]))
var.samples.W = apply(samples.W,1,var)
Calcula la varianza de cada muestra (fila)
V.var.samples.W[i] = var(var.samples.W) }
plot(n.W,V.var.samples.W,type="p",col=red")  Gráfica de la tendencia
```



Tips & Tricks!

- matrix(data, nrow=, ncol=) convierte los datos almacenados en *data* en una matriz con el número de filas especificado en nrow y el número de columnas en ncol.
- as.data.frame() verifica si un objeto es un data.frame, o lo convierte si es posible.
- apply() aplica una función a todos los elementos de un objeto en la dirección especificada: 1 para las filas o 2 para las columnas.
- dchisq() calcula la función de densidad de la distribución chi-cuadrado (χ^2).

6.3. Teorema del límite central

En la sección anterior, se ha estudiado el comportamiento probabilístico que tiene la suma de los elementos de una muestra, la media y su varianza. En resumen, estadísticos muestrales como la media y la varianza de las variables aleatorias: media, varianza y suma muestral se relacionan directamente con los parámetros de la población. Sin embargo, la distribución de probabilidad de estas variables aleatorias es desconocida, a excepción del caso en que la población tiene una distribución normal.

Aunque, en aplicaciones reales en ingeniería, muchos procesos o sistemas con incertidumbre (o aleatoriedad) generan variables que siguen una distribución de probabilidad normal, otros tantos no la siguen o simplemente no conocemos a ciencia cierta cómo se distribuyen.

Por otra parte, en estadística, muchos métodos se basan en que la población se distribuye normalmente, como por ejemplo en inferencia. Por tanto, si tenemos un caso en que no se puede garantizar la normalidad de la variable, ¿cómo aplicamos el método? Para superar este obstáculo, recurrimos a uno de los resultados más notables de la teoría estadística, que es el teorema central del límite o teorema del límite central. Se considera el teorema fundamental de la estadística y por ello lleva en su nombre la palabra "central". Y se enuncia a continuación.

Sea $X_1, X_2, X_3, ..., X_n$ un conjunto de n variables aleatorias, independientes e idénticamente distribuidas (con la misma función de densidad) con una media μ y una varianza σ^2 finita. Sea su suma $T_n = X_1 + X_2 + X_3 + ... + X_n$, entonces:

$$\lim_{n\to\infty} P\left(\frac{T_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z),$$

donde $\Phi(z)$ es la función de distribución $N(0,1)$.

Si las n variables aleatorias provienen de la muestra de una población, su suma tiende a seguir de manera asintótica una distribución normal, con media $n\mu$ y varianza $n\sigma^2$ siempre y cuando n sea lo suficientemente grande. Por otra parte, la media muestral que está dada por:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n},$$

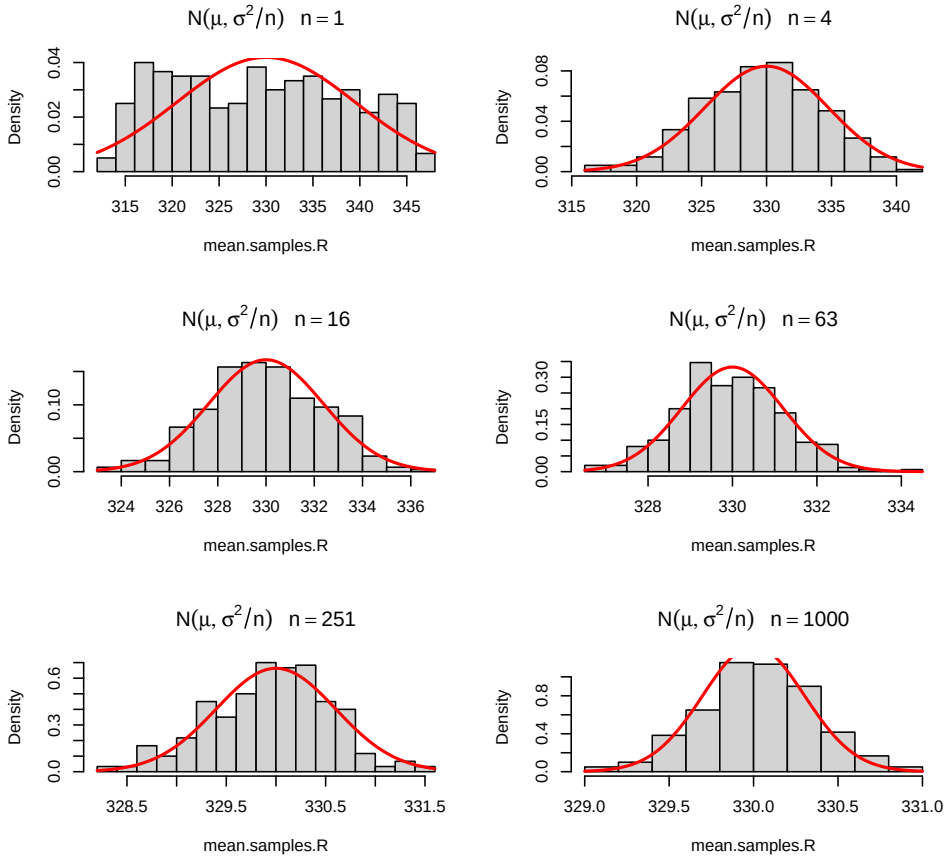
sigue una distribución normal, con una media $\mu_{\bar{X}} = \mu$ y una varianza $\sigma^2_{\bar{X}} = \frac{\sigma}{n}$.

Intuitivamente, ello nos indica que, al tomar una muestra aleatoria de tamaño suficientemente elevado de una población con cualquier distribución probabilística, la distribución de la media de dicha muestra y de otros estadísticos, tales como la proporción o la mediana, puede aproximarse mediante una distribución normal, cuya media (valor esperado) precisamente coincide con el valor del parámetro poblacional. Por tanto, se supera el obstáculo de no tener una población con distribución normal.

Supongamos que se quiere analizar el consumo de energía eléctrica en una ciudad. Es evidente que en cada hogar el consumo de electricidad se produce de manera aleatoria. Así, para cada hogar i, se puede asignar una variable aleatoria X_i, que describe el consumo particular, con su distribución de probabilidad y sus parámetros (media y varianza) respectivos. En cada hogar, se puede consumir energía eléctrica de modo muy diferente y, por tanto, con diferente distribución de probabilidad. Sin embargo, gracias al teorema central del límite, se puede asegurar que la cantidad total de energía eléctrica consumida en dicha ciudad (suma total de numerosos hogares) se aproxima a una distribución normal.

Para corroborar el teorema del límite central, utilizamos el ejemplo 1, en que la población (resistencia real de unos resistores) no se distribuye normalmente, pero sabemos que tiene una media de $\mu_R = 330\ \Omega$ y una varianza de $\sigma_R^2 = 90.75\ \Omega^2$. Se repite el procedimiento anterior simulando varias muestras con diferentes tamaños (n.R entre 1 y 1000) y se compara cada una de los histogramas resultantes con una distribución normal de media μ_R y varianza σ_R^2/n.

```
N = 300  Número de muestras
min.R = 313.5  Parámetros de la población
max.R = 346.5
mean.R = (min.R + max.R)/2
sd.R = sqrt((max.R - min.R)^2/12)
n.R = round(10^(seq(0,3,length=6)))  Diferentes tamaños de muestra
M.mean.samples.R = rep(0,6)  Inicialización del vector con ceros
par(mfrow=c(3,2))  Se divide la figura en 6
for (i in 1:6)  Comienzo de las repeticiones{
samples = runif(N*n.R[i], min=min.R, max=max.R)
samples.R = as.data.frame(matrix(samples, ncol=n.R[i]))
mean.samples.R = apply(samples.R,1,mean)
Calcula la media de cada muestra (fila)
title = bquote(N(mu,sigma^2/n)    n==.(n.R[i])) Cadena de caracteres
para el título
hist(mean.samples.R, prob=T, breaks=12, main=title)
Histograma de la media muestral
curve(dnorm(x,mean=mean.R,sd=sd.R/sqrt(n.R[i])),
add=T, lwd=2, col=red") }
```



6.4. Ejercicios propuestos

1. Se fabrica un cierto tipo de hilo con una resistencia a la tracción media de 78,3 kg y una desviación típica de 5,6 kg.

   a) Calcula la probabilidad de que, si se selecciona aleatoriamente un hilo, este tenga una resistencia a la tracción menor que 79 kg.

   b) Si se selecciona una muestra de 5 hilos, calcula la probabilidad de que la media de la resistencia de esa muestra sea menor que 79 kg.

   c) Si la muestra es de 50 hilos, calcula la probabilidad de que la media de la resistencia de esa muestra sea menor que 79 kg.

2. En una población, hay una tasa de infección de una determinada enfermedad de 1:100000 al año. Se considera X como la VAD que representa el número de habitantes infectados en un núcleo urbano de 3 millones de habitantes; por tanto, X sigue una distribución binomial con parámetros $n = 3 * 10^6$ y $p = 1 \times 10^{-}5$.

   a) Simula una muestra de 10 habitantes; calcula su media, su varianza y el histograma de resultados. ¿Se puede aproximar a una distribución normal? Si la

respuesta es afirmativa, ¿cuáles son los parámetros de la distribución?

b) Repite el numeral anterior con una muestra de 200 habitantes.

3. Los pesos de los hombres adultos de una determinada población se distribuyen normalmente, con una media de 80 kg y una desviación estándar de 15 kg .

   a) Encuentra la probabilidad de que un hombre seleccionado al azar pese más de 85 kg. Realiza la gráfica de la función de distribución y representa la probabilidad calculada.

   b) Simula una muestra de 25 observaciones; calcula su media, la varianza y el histograma de los resultados. ¿Se puede aproximar a una distribución normal? Si la respuesta es afirmativa, ¿cuáles son los parámetros de la distribución?

   c) Un elevador en un gimnasio para hombres tiene un letrero que dice que el peso máximo permitido es de 2125 kg. Si 25 hombres seleccionados al azar entran al elevador, ¿cuál es la probabilidad de que supere el peso máximo permitido? Realiza la gráfica de la función de distribución de la suma y representa la probabilidad calculada.

→7

Estimación

7.1. Introducción y objetivos

Supongamos que queremos analizar el promedio y la varianza de la altura de todos los estudiantes de una universidad. Para conocer estos parámetros, tendríamos que tomar la altura de todos y cada uno de los estudiantes y luego calcular su media y su varianza. Pero, si esto es imposible por el costo que representa, o simplemente porque la población y, por tanto, su espacio muestral son infinitos (como en casi todas las aplicaciones en ingeniería), solo podríamos inferir acerca de estos parámetros a partir de una muestra dada.

La inferencia es uno de los principales objetivos de la estadística y consiste en obtener información sobre parámetros desconocidos de una población a partir de un conjunto de datos obtenidos de una muestra aleatoria. Existen dos formas de hacer esta inferencia: por estimación y por contraste de hipótesis.

La estimación de estos parámetros puede ser de forma puntual, es decir, sugerir un valor de dicho parámetro, por ejemplo: $\hat{\mu}_X = 176$ cm y $\hat{\sigma}_X = 10$ cm, donde el símbolo "sombrero" ($\hat{}$) indica que no es un valor real, sino una estimación del parámetro de la VA X. Como esta estimación es altamente dependiente de la muestra, tendrá un error que no sería fácil de interpretar. Por tanto, en la mayoría de casos se prefiere sugerir no un valor, sino un intervalo que contiene el valor real del parámetro, por ejemplo: $\hat{\mu}_X \in [174, 178]$ cm y $\hat{\sigma}_X \in [9, 11]$ cm. Puesto que este intervalo depende también de la muestra (por cada muestra que se utilice, se tendrá un intervalo diferente), se ha de garantizar que una proporción significativa de los intervalos calculados contiene al valor real del parámetro; esta proporción se denomina *nivel de confianza del intervalo* y,para simplificar, este intervalo se denomina **intervalo de confianza del** $100(1-\alpha)\,\%$.

En esta sesión, se analizan e implementan los métodos clásicos para el cálculo del intervalo de confianza de los parámetros más comunes: media y varianza. Por tanto al finalizar, el alumno ha de ser capaz de:

- Comprender el concepto de estimación: puntual y por intervalos.
- Realizar en **R** la estimación por intervalos de la media y la varianza de una población a partir de una muestra.
- Comprender la influencia del tamaño de la muestra en la estimación por intervalos.
- Comprender el concepto del nivel de confianza de un intervalo.

7.2. Estimación de la media de una población

Para ilustrar el procedimiento para la estimación de la media poblacional, se utiliza un conjunto de datos de una muestra de 237 estudiantes de estadística en una universidad australiana. Este conjunto de datos, llamado *survey* pertenece al *package* MASS, que está incluido en la instalación básica de **R** pero debe cargarse con anterioridad de la siguiente manera:

```
library(MASS)  Carga la librería MASS
head(survey)  Visualiza las primeras observaciones
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never
##   Height      M.I       Age
## 1 173.00     Metric    18.250
## 2 177.80   Imperial    17.583
## 3     NA       <NA>    16.917
## 4 160.00     Metric    20.333
## 5 165.00     Metric    23.667
## 6 172.72   Imperial    21.000
```

Este conjunto de datos contiene, entre otra información, el sexo de cada estudiante seleccionado (Sex), su frecuencia cardiaca (Pulse), su altura (Height) y su edad (Age). En el transcurso de esta sesión, se harán las estimaciones de los parámeros de la variable altura, aunque el procedimiento es extensible a cualquiera de las variables cuantitativas/numéricas del conjunto de datos. Como algunos de los estudiantes encuestados no contestaron todas las preguntas, existen algunas observaciones con valores faltantes, por lo que, debemos filtrarlos usando la función na.omit(), convertirlos a un vector numérico con la función as.numeric() y guardarlos para su posterior uso en el objeto height.

```
height = as.numeric(na.omit(survey$Height))
Vector de datos sin valores faltantes (NA)
```

7.2.1. Estimación puntual de la media

El objetivo de un estimador puntual es obtener un valor numérico único que se aproxime a algún parámetro (único y desconocido) de la población a partir de un estadístico de la muestra (aleatoria pero conocida). De esta forma, el estadístico $\hat{\theta}$ se denomina estimador puntual del parámetro θ.

Recordando las propiedades de la media muestral de la sesión anterior $E(\bar{X}) = \mu_X$, es decir, el valor que esperamos que se obtenga al calcular la media de una muestra, es igual al valor de la media de la población. Por tanto, la media de la muestra es un buen estimador de la media poblacional ($\hat{\mu}_X = \bar{X}$). Además, este estimador posee propiedades esenciales tales como insesgadez, eficiencia, convergencia y robustez.

```
est_mu = mean(height); est_mu
```

```
## [1] 172.3809
```

Si queremos ir un poco más allá de una estimación puntual (un solo valor plausible) de la media poblacional, necesitaremos un modo de cuantificar su precisión. Es decir, definir un intervalo en torno a esta estimación puntual con un nivel de confianza dado, es decir, el nivel de seguridad que tenemos de que el intervalo incluye el parámetro. El método, aunque es general, presenta pequeñas diferencias en función de la distribución de la población, del conocimiento o desconocimiento de la varianza de la población y del tamaño de la muestra.

7.2.2. Intervalo de confianza de la media de una población con distribución normal y varianza conocida

Como se sabe que la población se distribuye normalmente, de acuerdo con las propiedades de la media muestral estudiadas en la sesión anterior, podemos afirmar que la media muestral también se distribuye normalmente y su varianza es igual a la varianza de la población dividida por n:

$$\bar{X} \hookrightarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

donde μ_X es el parámetro desconocido y, por tanto, por estimar.

Para cualquier muestra aleatoria, los puntos extremos del intervalo estimado para la media de la población con un nivel de confianza de $(1-\alpha)$ % vienen dados por:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

donde $z_{\alpha/2}$ denota el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribución normal estándar, $\frac{\sigma}{\sqrt{n}}$ es la desviación típica de la media muestral (aquí la denominamos *error estándar*) y $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ es el margen de error.

Asumiendo que la altura de todos los estudiantes del ejemplo tiene una distribución normal y que la desviación típica es conocida $\sigma = 9.48$ cm (desviación típica poblacional), el intervalo de estimación para la media de la altura de los estudiantes con un $95\,\%$ de confianza se calcula como sigue:

```
sigma = 9.48  sigma conocida
alpha = 0.05  ya que el nivel de confianza (1-alpha) es 0.95,
alpha = 0.05
n = length(height)  tamaño de la muestra
SE = sigma/sqrt(n); SE  error estándar
```

```
## [1] 0.6557453
```

Debido a que el intervalo está centrado en $\bar{x}$, el $100(1-\alpha)\,\% = 95\,\%$ de confianza implica el 97.5 $(1-\alpha/2 = 0.975)$ percentil de la distribución normal en la cola superior. Por tanto, $z_{\alpha/2}$ viene dado por qnorm(0.975) (qnorm(1-alpha/2)). Lo multiplicamos por el error estándar SE y obtenemos el margen de error E.

```
E = qnorm(1-alpha/2)*SE; E  margen de error
```

```
## [1] 1.285237
```

A la media muestral, le sumamos y le restamos este valor para obtener los extremos del intervalo.

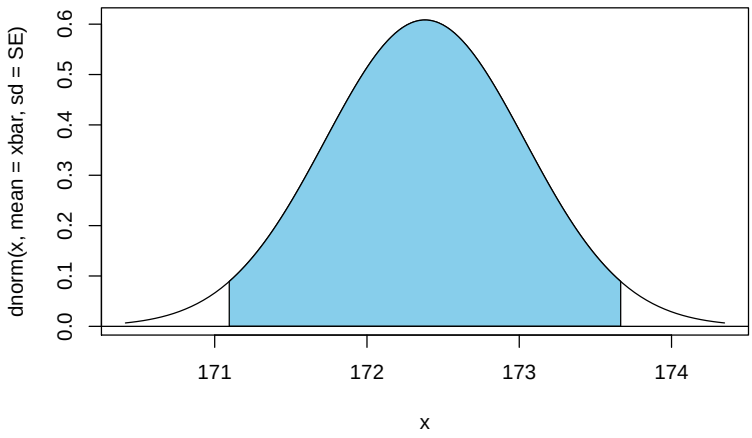
```
xbar = mean(height)  media muestral
IC = xbar + c(-E,E); IC  Intervalo estimado
```

```
## [1] 171.0956 173.6661
```

Finalmente, representamos gráficamente el intervalo.

```
Gráfica de la función de densidad de la media muestral
curve(dnorm(x,mean=xbar,sd=SE),from=xbar-3*SE, to=xbar+3*SE)

Gráfica de la región del intervalo
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dnorm(seq(IC[1],IC[2],length=100),mean=xbar,sd=SE),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



Conclusión: Asumiendo la desviación típica poblacional σ como 9.48, el margen de error para la media de la altura de los estudiantes con un 95 % de confianza es 1.2852 centímetros, y el intervalo para la media poblacional se halla entre 171.10 y 173.67 centímetros.

Hasta el momento, hemos usado la fórmula general para el cálculo del intervalo de confianza de la media; sin embargo, podemos aplicar la función z.test del paquete TeachingDemos. No es un paquete que se instala por defecto en **R**, sino que se debe instalar y cargar previamente:

```
install.packages("TeachingDemos")
instalación del paquete (solo se hace una vez)
```

```
library(TeachingDemos)
cargar el paquete (se ha de cargar siempre que se utilice)
IC = z.test(height, sd=sigma); IC
Cálculo del intervalo de confianza para la media
```

```
##  One Sample z-test
##
## data:  height
## z = 262.88, n = 209.00000, Std. Dev. = 9.48000,
## Std. Dev. of the sample
## mean = 0.65575, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 171.0956 173.6661
## sample estimates:
## mean of height
## 172.3809
```

Si queremos cambiar el nivel de confianza del intervalo, podemos añadirlo mediante la opción conf.level=:

```
IC = z.test(height, sd=sigma, conf.level=0.9); IC
```

```
##  One Sample z-test
##
## data:  height
## z = 262.88, n = 209.00000,
## Std. Dev. = 9.48000, Std. Dev. of the sample
## mean = 0.65575, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
## 171.3023 173.4595
## sample estimates:
## mean of height
## 172.3809
```

De acuerdo con los intervalos calculados, se verifica que, cuanto mayor es el nivel de confianza, más amplio es el intervalo.

El resultado, almacenado en una variable de tipo lista que contiene, entre otros, los siguientes componentes:

- statistic: valor del estadístico $z = \dfrac{\bar{x}}{\sigma/\sqrt{n}}$.
- data.name: cadena de carácteres con el nombre de la variable.
- data.parameter: vector con los grados de libertad del estadístico $n-1$, la desviación estándar de la población σ y de la media muestral s.
- conf.int: vector con los extremos del intervalo.
- estimates: media muestral $\bar{x}$ (estimación puntual de la media poblacional $\hat{\mu} = \bar{x}$).

De esta manera, para visualizar solamente el intervalo, basta con ejecutar:

```
IC$conf.int
```

```
## [1] 171.3023 173.4595
## attr(,"conf.level")
## [1] 0.9
```

7.2.3. Intervalo de confianza de la media de una población con distribución normal y varianza desconocida

En el caso anterior, si se calcula el mismo intervalo de confianza para la media a partir de otra muestra de igual tamaño, solo puede cambiar el centro o punto medio del intervalo

($\bar{x}$) pero no su amplitud ($2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$), ya que la varianza de la población σ es conocida e invariante.

Si esta varianza es desconocida (es el caso más factible en aplicaciones reales), entonces se tendría que estimar. Un buen estimador puntual de σ es la varianza de la muestra corregida $\hat{\sigma} = S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$. Este estimador posee las características ideales: insesgadez, eficiencia, convergencia y robustez (consistencia). Dado que esta estimación depende directamente de la muestra, al calcular el mismo intervalo de confianza para la media a partir de otra muestra de igual tamaño, no solo puede cambiar el centro del intervalo ($\bar{x}$) sino también su amplitud ($2z_{\alpha/2}\frac{S}{\sqrt{n}}$). Por tanto, para calcular el intervalo de confianza no se asume que $\bar{x}$ está distribuida normalmente, sino que sigue una distribución muy similar a la normal, que se denomina *distribución t-Student*, con $n-1$ de grados de libertad, donde n es el tamaño de la muestra.

La distribución de *t-Student* es similar a la distribución normal estándar $N(0,1)$: las dos son simétricas, con una media de cero y forma de campana; la principal diferencia radica en la varianza. La varianza de la distribución *t-Student* con $n-1$ grados de libertad es $V(T) = \frac{n-1}{n-2}$, es decir, mayor que 1, que es la varianza de la normal. Si observamos el valor de la varianza de la distribución de *t-Student* cuando los grados de libertad tienden a infinito, tenemos:

$$\lim_{n\to\infty}\frac{n-1}{n-2} = 1,$$

es decir, cuando los grados de libertad tienden a infinito, la distribución *t-Student* tiende a la distribución normal estándar.

Teniendo en cuenta lo anterior, para cualquier muestra aleatoria, los puntos extremos del intervalo estimado para la media de la población con un nivel de confianza de $(1-\alpha)$ % vienen dados por:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},$$

donde $t_{\alpha/2,n-1}$ denota el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribución de *t-Student* con $n-1$ grados de libertad, S es la desviación típica de la muestra, $\frac{S}{\sqrt{n}}$ es la desviación típica de la media muestral (error estándar) y $t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$ es el margen de error.

En el ejemplo dado, asumiendo distribución normal y varianza poblacional desconocida, el intervalo de confianza al 95 % utilizando únicamente los 10 primeros elementos de la muestra se calcula de la siguiente forma:

```
alpha = 0.05  ya que el nivel de confianza (1-alpha) es 0.95,
alpha = 0.05
height.10 = height[1:10]
Se seleccionan una muestra de 10 observaciones
```

```
n = length(height.10)  tamaño de la muestra
S = sd(height.10)  desviación típica d ela muestra
SE = S/sqrt(n); SE  error estándar
```

```
## [1] 2.874441
```

Igualmente que en el caso anterior, el $100(1-\alpha)\,\% = 95\,\%$ de confianza implica el 97.5 $(1-\alpha/2 = 0.975)$ percentil de la distribución *t-Student* con 9 grados de libertad $(n-1)$ en la cola superior. Por tanto, $t_{\alpha/2,n-1}$ viene dado por qt(0.975,df=9) (qt(1-alpha/2,n-1)). Lo multiplicamos por el error estándar SE y obtenemos el margen de error E.

```
E = qt(1-alpha/2,n-1)*SE; E  margen de error
```

```
## [1] 6.502436
```

A la media muestral le sumamos y le restamos este valor para obtener los extremos del intervalo.

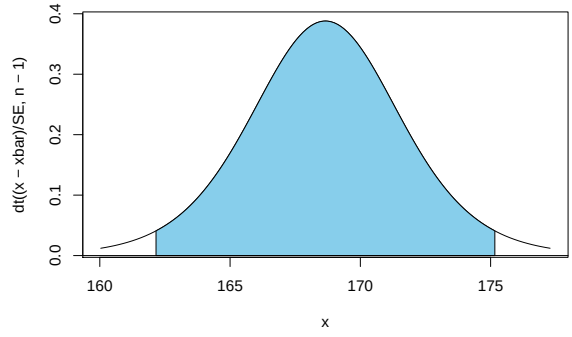
```
xbar = mean(height.10)  media muestral
IC = xbar + c(-E,E); IC  Intervalo estimado
```

```
## [1] 162.1576 175.1624
```

Finalmente, representamos gráficamente el intervalo.

```
Gráfica de la función de densidad de la media muestral, distribución
de t-Student  centrada en xbar
curve(dt((x-xbar)/SE,n-1),from=xbar-3*SE, to=xbar+3*SE)

Gráfica de la región del intervalo
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dt(seq((IC[1]-xbar)/SE,(IC[2]-xbar)/SE,length=100),n-1),0)
polygon(cord.x,cord.y,col="skyblue")  abline(h=0)
```



Conclusión: Desconociendo la desviación típica poblacional σ y teniendo en cuenta las 10 primeras observaciones, el margen de error para la media de la altura de los estudiantes con un $95\,\%$ de confianza es 6.502 centímetros, y el intervalo para la media poblacional se halla entre 162.16 y 175.16 centímetros. Si cambiamos la muestra, ¿qué resultados se obtienen? Calcula varios intervalos utilizando diferentes muestras de tamaño menor a 30.

En **R** también se puede encontrar la función que calcula este intervalo, t.test, del paquete stats, que normalmente está integrado en la instalación básica de **R**.

```
IC = t.test(height.10); IC  Cálculo del intervalo de confianza para la
media
```

```
##  One Sample t-test
##
## data:  height.10
## t = 58.676, df = 9, p-value = 6.111e-13
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 162.1576 175.1624
## sample estimates:
## mean of x
## 168.66
```

De igual forma que con z.test, se puede cambiar el nivel de confianza y el objeto IC es una lista que contiene las mismas componentes.

7.2.4. Intervalo de confianza de la media de una población con distribución desconocida

Recordando el teorema del límite central de la sesión anterior, al tomar una muestra aleatoria de tamaño suficientemente elevado de una población con cualquier distribución probabilística, la distribución de la media de dicha muestra se puede aproximar mediante una distribución normal cuyo valor esperado coincide con el valor de la media poblacional.

Por tanto, si tenemos una muestra suficientemente grande de una población con distribución desconocida pero su varianza es conocida, el intervalo de confianza al $(1-\alpha)\,\%$ se calcula como en la sección 2.2:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Por otra parte, si el valor de la varianza poblacional es desconocido, entonces el intervalo se calcula como en la sección 2.3:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

Pero, como la distribución de *t-Student* es similar a la normal estándar para altos valores de grados de libertad, es decir, para una muestra suficientemente grande (típicamente $n > 30$), entonces el intervalo se puede aproximar de la siguiente forma:

$$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

Continuando con el ejemplo dado, teniendo la muestra de 209 observaciones y considerando que la distribución y la varianza de altura de los estudiantes de una determinada universidad es desconocida, el intervalo de confianza al $95\,\%$ de la altura media de dichos estudiantes se calcula de la siguiente forma:

```
alpha = 0.05  ya que el nivel de confianza (1-alpha) es 0.95,
alpha = 0.05
n = length(height)  tamaño de la muestra
S = sd(height)  desviación típica de la muestra
SE = S/sqrt(n); SE  error estándar
```

```
## [1] 0.6811677
```

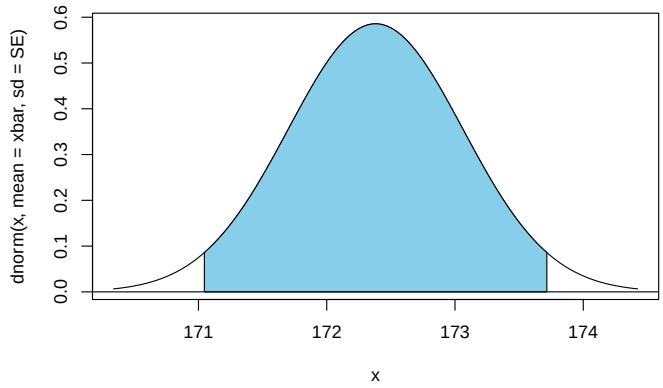
```
E = qnorm(1-alpha/2)*SE; E  margen de error
```

```
## [1] 1.335064
```

```
xbar = mean(height)  media muestral
IC = xbar + c(-E,E); IC  Intervalo estimado
```

```
## [1] 171.0458 173.7159
```

```
Gráfica de la región del intervalo
curve(dnorm(x, mean=xbar, sd=SE),from=xbar-3*SE, to=xbar+3*SE)
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dnorm(seq(IC[1],IC[2],length=100), mean=xbar, sd=SE),0)
polygon(cord.x,cord.y,col="skyblue")  abline(h=0)
```



o usando la función z.test.

```
IC = z.test(height, sd=sd(height)); IC
```

```
## One Sample z-test
##
## data:  height
## z = 253.07, n = 209.00000, Std. Dev. = 9.84753, Std. Dev. of the sample
## mean = 0.68117, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 171.0458 173.7159
## sample estimates:
## mean of height
## 172.3809
```

Teniendo en cuenta que la distribución *t-Student* es similar a la normal estándar para una muestra grande, entonces una aproximación puede ser:

```
E = qt(1-alpha/2,n-1)*SE; E  margen de error
```

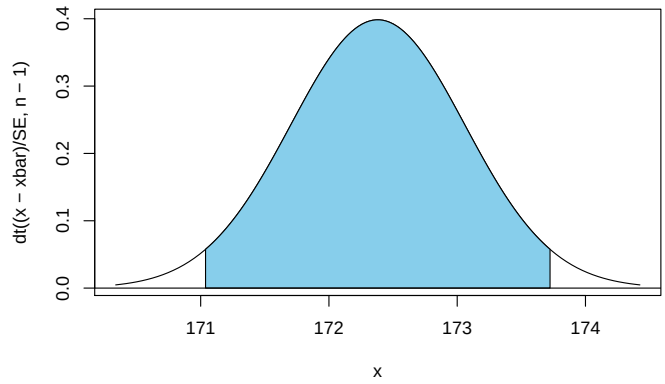
```
## [1] 1.342878
```

```
IC = xbar + c(-E,E); IC  Intervalo estimado
```

```
## [1] 171.0380 173.7237
```

```
Gráfica de la región del intervalo
curve(dt((x-xbar)/SE,n-1),from=xbar-3*SE, to=xbar+3*SE)
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dt(seq((IC[1]-xbar)/SE,(IC[2]-xbar)/SE,length=100),n-1),0)

polygon(cord.x,cord.y,col="skyblue")  abline(h=0)
```



o usando la función t.test.

```
IC = t.test(height); IC
```

```
##  One Sample t-test
##
## data:  height
## t = 253.07, df = 208, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 171.0380 173.7237
## sample estimates:
## mean of x
## 172.3809
```

Conclusión: Desconociendo la distribución de la población, pero con una muestra grande ($n = 209$), el margen de error para la media de la altura de los estudiantes con un 95 % de confianza es de 1.34 centímetros, y el intervalo para la media poblacional se halla entre 171.05 y 173.72 centímetros. Nótese la pequeña diferencia que hay en el intervalo calculado usando la normal y usando la distribución de *t-Student*. Si la muestra es pequeña, por ejemplo 10, ¿cuál es la diferencia al usar la distribución normal? Recuerda que esta aproximación solo es posible si la muestra es grande.

7.2.5. Tamaño de la muestra

Se ha podido observar la influencia directa que tiene el tamaño de la muestra n en la longitud del intervalo de confianza resultante. El tamaño de la muestra necesario para cumplir con los requerimientos del intervalo de confianza al $(1 - \alpha)$ % de la media poblacional, es decir, para un margen de error E viene dado por:

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2}.$$

En el ejemplo, conociendo que la desviación estándar poblacional σ de la altura de los estudiantes es 9.48, para encontrar el tamaño necesario de la muestra para tener un margen de error de 1.2 centímetros con un 95 % de confianza, se puede calcular como sigue:

```
z.alpha2 = qnorm(.975)
sigma = 9.48 E = 1.2 n = z.alpha2^2*sigma^2/ E^2
```

Conclusión: Basándonos en que conocemos la desviación estándar poblacional, la muestra ha de tener un tamaño mínimo de 240 observaciones para un margen de error de 1.2 centímetros con un 95 % de confianza. Si la varianza de la población es desconocida, ¿se puede usar la ecuación anterior?

7.2.6. ¿Qué representa el nivel de confianza?

Se ha definido en todo momento que el estimador intervalar tiene un nivel de confianza de $(1-\alpha)$ %. Por otra parte, se puede apreciar que, si se conoce la varianza poblacional, la longitud del intervalo $\left(2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ solo depende del tamaño de la muestra n. Por tanto, si realizamos otra muestra del mismo tamaño, la longitud del intervalo se mantiene, pero la posición del mismo no, ya que depende de $\bar{x}$ y este nuevo intervalo puede o no incluir el valor real de la media de la población. Teniendo en cuenta lo anterior, se puede definir que el nivel de confianza del intervalo es la proporción de intervalos estimados que incluyan el valor real del parámetro. En otras palabras, el intervalo estimado es un valor del intervalo aleatorio (variable aleatoria) que tiene 100α % de probabilidad de que no incluya el valor real del parámetro.

Para verificarlo, suponemos que la altura media de la población del ejemplo es 172 y su desviación estándar es 9.48. Se simulan 100 muestras de 20 observaciones cada una. Calculamos para cada muestra un intervalo de confianza del 95 % ($\alpha = 0.05$) y lo representamos gráficamente. Finalmente, resaltamos aquellos intervalos estimados que no contienen el valor real de la media de la población.

```
mu = 172 ; sigma = 9.48 Parámetros de la población
alpha = 0.05 Nivel de confianza
N = 100 Número de muestras
n = 20 Tamaño de la muestra

Simulación de las muestras
set.seed(12) Fijación de la semilla de la aleatoriedad
sim = rnorm(N*n, mean=mu, sd=sigma) Simulación de N * n muestras
samples = as.data.frame(matrix(sim, ncol=n))
 Organización en un data.frame

Cálculo de los intervalos de confianza
mean.samples = apply(samples,1,mean)
Media de cada muestra (fila)
sd.samples = apply(samples,1,sd) Desviación típica de cada muestra
(fila)
E = qnorm(1-alpha/2)*sigma/sqrt(n) Margen de error
IC = rbind(mean.samples -E,mean.samples+E) Intervalos de confianza
IC[,1:7] Visualiza los 7 primeros intervalos
```

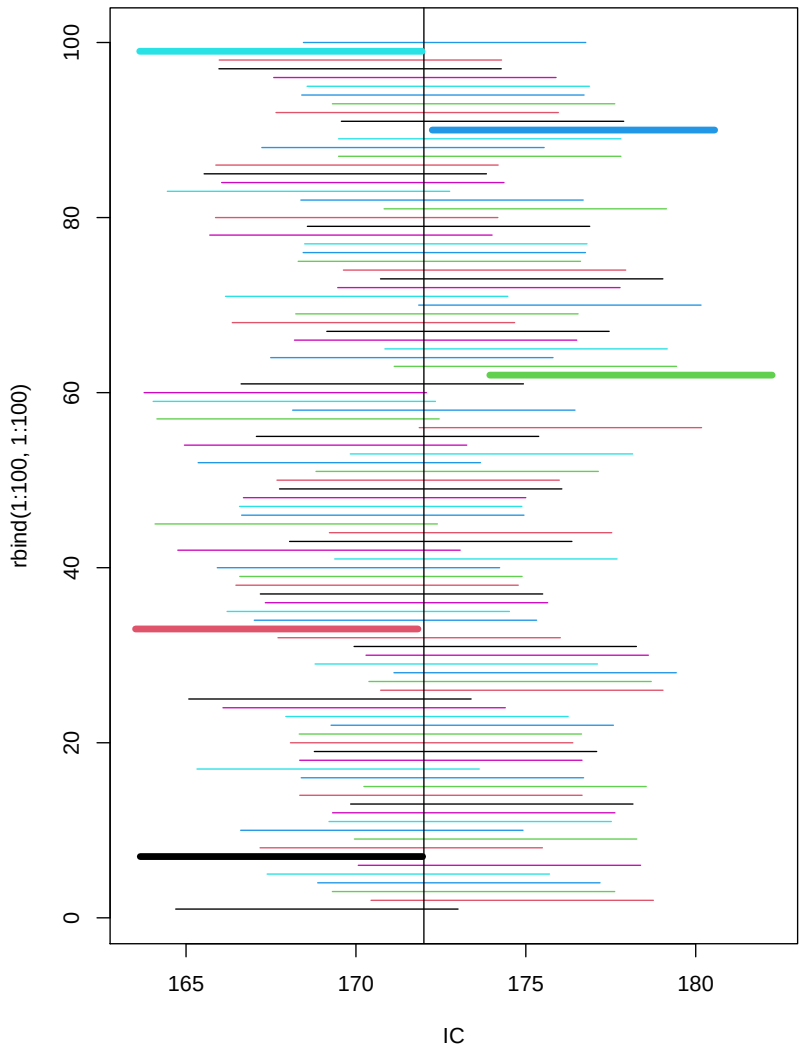
```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 164.6960 170.4440 169.3063 168.8746 167.3864 170.0693 163.6530
## [2,] 173.0054 178.7534 177.6157 177.1840 175.6958 178.3787 171.9624
```

```
Representación gráfica
matplot(IC,rbind(1:100,1:100),type="l",lty=1)
Línea horizontal por intervalo
```

```
abline(v=mu) Línea vertical que representa el valor poblacional real
out=which(!(IC[1,]<mu  mu<IC[2,]))
Detección de intervalos que NO contienen a mu
Error = length(out);
Error Cuántos intervalos NO contienen el valor real
```

```
## [1] 5
```

```
matplot(IC[,out],rbind(out,out),type="l",lty=1,add=T,lwd=5)
Resalta los intervalos
```



7.3. Intervalo de confianza para la varianza de una población con distribución normal

En la sesión anterior, se ha analizado y verificado el comportamiento probabilístico de la varianza de una muestra y su relación con la varianza de la población cuando esta se distribuye normalmente. Se ha llegado a la conclusión de que la varianza de la muestra es una variable aleatoria continua, ya que depende de la muestra seleccionada. Por tanto, tiene un valor esperado, una varianza y la relación $(n-1)\frac{S^2}{\sigma_X^2}$ está distribuida de acuerdo con la función de chi-cuadrado (χ^2) con $(n-1)$ grados de libertad.

Teniendo en cuenta que esta distribución no es simétrica alrededor de ningún punto, un $(1-\alpha)\,\%$ de intervalo de confianza para la varianza de la población normal significa que se han de buscar los valores $\chi^2_{1-\frac{\alpha}{2},n-1}$ y $\chi^2_{\frac{\alpha}{2},n-1}$ de modo que:

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha,$$

donde $\chi^2_{1-\frac{\alpha}{2},n-1}$ y $\chi^2_{\frac{\alpha}{2},n-1}$ denotan los $100(\alpha/2)$ y $100(1-\alpha/2)$ percentiles de la distribución chi-cuadrado (χ^2) con $(n-1)$ grados de libertad.

De esta forma, el intervalo de confianza para la varianza viene dado por:

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$$

Asumiendo que la altura de todos los estudiantes del ejemplo tiene una distribución normal, el intervalo de confianza al $95\,\%$ para la varianza de la altura de los estudiantes se calcula como sigue:

```
alpha = 0.05
n = length(height)  Tamaño de la muestra
Ssq = var(height)  Varianza de la muestra
chi.alpha.lower = qchisq(alpha/2, df=n-1, lower.tail = TRUE)
100(alpha/2) percentil
chi.alpha.upper = qchisq(alpha/2, df=n-1, lower.tail = FALSE)
100(1-alpha/2) percentil
IC = (n-1)*Ssq*c(1/chi.alpha.upper,1/chi.alpha.lower);
IC  Intervalo de confianza
```

```
## [1]  80.73552 118.68446
```

Como no hay una distribución de probabilidad concreta para la varianza muestral (solo para la relación entre las dos varianzas), no tiene sentido hacer una gráfica del intervalo. Para el cálculo del intervalo, también se puede usar la función sigma.test del paquete TeachingDemos.

```
library(TeachingDemos)  Cargar el paquete
IC = sigma.test(height); IC
IC para la varianza de una población normal
```

```
##  One sample Chi-squared test for variance
##
## data:  height
## X-squared = 20171, df = 208, p-value < 2.2e-16
## alternative hypothesis: true variance is not equal to 1
## 95 percent confidence interval:
## 80.73552 118.68446
## sample estimates:
## var of height
## 96.9738
```

Tips & Tricks!

- na.omit() devuelve el objeto eliminando las observaciones con valores faltantes.
- as.numeric() convierte cualquier tipo de objeto a un objeto tipo numérico.
- z.test(data, sd=) calcula el intervalo de confianza para la media de la población (a partir de la muestra data) cuando la varianza de la población es conocida o la muestra es grande. Pertenece a la librería TeachingDemos, que se ha de cargar previamente.
- t.test(data) calcula el intervalo de confianza para la media de la población (a partir de la muestra data) cuando la población se distribuye normalmente con varianza desconocida.
- sigma.test(data) calcula el intervalo de confianza para la varianza de una población que se distribuye normalmente (a partir de la muestra data). Pertenece a la librería TeachingDemos, que se ha de cargar previamente.

7.4. Ejercicios propuestos

1. Un fabricante de vehículos sabe que el consumo de gasolina de sus vehículos se distribuye normalmente. Se selecciona una muestra aleatoria simple de coches, se observa su consumo cada cien kilómetros y se obtienen las siguientes observaciones: (19.2, 19.4, 18.4, 18.6, 20.5, 20.8). Halla el intervalo de confianza para el consumo medio de gasolina de todos los vehículos de este fabricante, con un nivel de confianza del 99 % y represéntalo gráficamente.

2. Un supervisor de control de calidad en una planta enlatadora sabe que la cantidad exacta en cada lata varía, pues hay ciertos factores imposibles de controlar que afectan la cantidad de llenado. El llenado medio por lata es importante, pero igualmente importante es la variación de la cantidad de llenado. Si la varianza es grande, algunas latas contendrán muy poco contenido, y otras, demasiado. A fin de estimar la variación del llenado en la enlatadora, el supervisor escoge al azar 9 latas y pesa el contenido de cada una de ellas y obtiene el siguiente pesaje (en onzas): 7.96, 7.90, 7.98, 8.01, 7.97, 8.03, 8.02, 8.04, y 8.02. Sabiendo que el peso se distribuye normalmente, establece un intervalo de confianza del $95\,\%$ para la varianza poblacional.

3. Los siguientes datos son las puntuaciones obtenidas para 45 personas de una escala de depresión (mayor puntuación significa mayor depresión).

   2; 5 ; 6; 8 ; 8; 9; 9; 10; 11; 11; 11; 13; 13; 14; 14; 14; 14; 14; 14; 15; 15; 16; 16; 16; 16; 16; 16; 16; 16; 17; 17; 17; 18; 18; 18; 19; 19; 19; 19; 19; 19; 19; 19; 20; 20

   a) Halla el intervalo de confianza para la demanda media diaria con un nivel de confianza del $80\,\%$

   b) Halla el intervalo de confianza para la demanda media diaria con un nivel de confianza del $95\,\%$, conociendo que la desviación típica es de 4.5

Contraste de hipótesis

8.1. Introducción y objetivos

En la sesión anterior, se ha explicado que la inferencia estadística consiste en obtener información sobre parámetros desconocidos de una población a partir de un conjunto de datos obtenidos de una muestra aleatoria. Esta información no solo se obtiene a través de la estimación, sino que muchos de los problemas en ingeniería se refieren a la formulación de procedimientos de decisión. Se expone una información y el procedimiento basado en una muestra nos conducirá al rechazo o a la aceptación de dicha hipótesis.

Por ejemplo, un fabricante de agua embotellada afirma que cada botella contiene 50 cl de agua. Como esta información está en la etiqueta, asumimos que es verdadera, pero ¿lo es?

Para responder a esta pregunta, podemos aplicar el procedimiento del contraste estadístico de hipótesis, o simplemente contraste de hipótesis. Se postula una hipótesis inicial, que es una conjetura acerca de una población (nunca de la muestra), y se contrasta con las observaciones de la muestra, es decir, se decide si la propiedad de la población que ha sido postulada (hipótesis) es compatible con lo observado en la muestra de dicha población.

Este método está muy relacionado con los intervalos de confianza, pero con un enfoque diferente. En lugar de estimar el parámetro desconocido, suponemos un valor y analíticamente rechazamos o no esta hipótesis.

En esta sesión, se analiza e implementa el procedimiento para el contraste de hipótesis de la media de una población. Por tanto, al finalizar, el alumno ha de ser capaz de:

- Comprender el concepto de contraste de hipótesis.
- Realizar en **R** el contraste de la media de una población a partir de una muestra.
- Comprender la influencia de la formulación correcta de las hipótesis.
- Comprender el concepto del nivel de significancia de un intervalo.

8.2. Planteamiento general del problema de contraste

En este apartado, se describe el procedimiento general para la realización de un contraste de hipótesis de la media de una población. Aunque es general, este planteamiento también se puede extender al contraste de otros parámetros como, la proporción, la varianza, etc. La diferencia radica principalmente en la elección del estadístico de contraste y su respectiva distribución. Para ilustrarlo, al principio tomaremos como ejemplo el que se ha presentado en la introducción en que un fabricante de agua embotellada afirma que cada botella contiene 50 cl de agua. Para definir los casos particulares, se definirán tres ejemplos adicionales.

8.2.1. Formular las hipótesis

Como punto de partida, tenemos la afirmación del fabricante de que la media de la población es de 50 cl. Esta afirmación la llamamos **hipótesis inicial** o **hipótesis nula** y se denota típicamente como H_0. Para contrastarla, necesitamos una **hipótesis alternativa** que contradiga la nula, eso es,que sean mutuamente excluyentes,que denominaremos H_1.

El resultado del contraste nos llevará a decidir si las observaciones (muestra) son consistentes con la hipótesis inicial. Si no lo son, ello significa que la muestra refuta la hipótesis inicial y, por tanto, se acepta la hipótesis alternativa. Por el contrario, si las observaciones son consistentes, no se puede demostrar que esta sea verdadera, porque puede existir otra muestra que la contradiga. *"En un caja de manzanas, una manzana sana no demuestra que TODAS las manzanas de la caja están sanas, pero una manzana podrida sí demuestra que NO TODAS las manzanas de la caja están sanas."*

Teniendo en cuenta lo anterior, podemos concluir que la hipótesis nula se puede rechazar, pero NUNCA se puede aceptar. Por tanto, la elección de la hipótesis alternativa (que sí se acepta, al rechazar la nula) depende de lo que queramos demostrar, y es por ello que se suele llamar **hipótesis de investigación**.

Por ejemplo, si representamos al consumidor de agua embotellada y **queremos demostrar que el fabricante miente**, suponemos que la cantidad de agua media es igual a 50 cl ($H_0 : \mu = 50cl$) y la contrastamos con la hipótesis de que esta cantidad es diferente a 50 cl ($H_1 : \mu \neq 50cl$). De esta manera, el resultado del test puede ser:

- Rechazar H_0, lo cual implica que aceptamos que la cantidad media de agua es **diferente a 50 cl**, que es justo lo que deseamos.
- No rechazar H_0, lo cual implica que no aceptamos que la cantidad de agua es igual ni diferente a 50 cl. Podríamos decir que, con la muestra aportada, la prueba queda inconclusa.

Por otra parte, como representantes del consumidor pero que **queremos demostrar que el fabricante no solo miente, sino que además la cantidad es menor**, suponemos que la cantidad de agua media es igual a 50 cl ($H_0 : \mu = 50cl$) y la contrastamos con la

hipótesis de que esta cantidad es menor que 50 cl ($H_1 : \mu < 50cl$). De esta manera, el resultado del test puede ser:

- Rechazar H_0, lo cual implica que aceptamos que la cantidad media de agua es **menor que 50 cl**, que es justo lo que deseamos.
- No rechazar H_0, lo cual implica que no aceptamos ninguna hipótesis y la prueba queda inconclusa.

Finalmente, si representamos al fabricante, **queremos demostrar que la cantidad es incluso mayor a la estipulada en la etiqueta**. Por tanto, suponemos que la cantidad de agua media es igual a 50 cl ($H_0 : \mu = 50cl$) y la contrastamos con la hipótesis de que esta cantidad es mayor que 50 cl ($H_1 : \mu > 50cl$). De esta manera, el resultado del test puede ser:

- Rechazar H_0, lo cual implica que aceptamos que la cantidad media de agua es **mayor que 50 cl**, que es justo lo que deseamos.
- No rechazar H_0, lo cual implica que no aceptamos ninguna hipótesis y no demostramos de que el fabricante miente.

8.2.2. Especificar el nivel de significancia α

Considerando lo descrito hasta ahora, se ha planteado la existencia de un problema de rechazo/no-rechazo de la hipótesis nula. Esto implica una posibilidad de fracasar en la decisión tomada, es decir, rechazar H_0 cuando esta es verdadera (error de tipo I), o no rechazarla cuando realmente es falsa (error de tipo II).

Como no sabemos el valor real del parámetro μ, la única forma de cuantificar este error es por medio de probabilidades, por tanto, se define el error de tipo I como la probabilidad de rechazar una hipótesis que en realidad es verdadera:

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ es cierta}),$$

y el error de tipo II como la probabilidad de no rechazar una hipótesis que en realidad es falsa:

$$\beta = P(\text{No rechazar } H_0 | H_0 \text{ es falsa}).$$

α es el **nivel de significancia del contraste**. Como es la probabilidad de cometer un error, se le asigna un valor pequeño, típicamente 0.05 o 0.01. Por otra parte, $1-\beta$ (probabilidad de rechazar una hipótesis falsa) se denomina **la potencia del contraste**.

8.2.3. Seleccionar el tipo de contraste

Dependiendo de la formulación de las hipótesis, se pueden tener tres tipos de contrastes:

Contraste bilateral (dos colas):

Es el caso en que las hipótesis nula y alternativa para la prueba de la media poblacional se formulan de la siguiente forma:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu \neq \mu_0. \end{aligned}$$

donde μ_0 es el valor supuesto de la media poblacional μ. Como el objetivo final es decidir si la media poblacional es *diferente* del valor supuesto, basta con que la muestra tenga una media estadísticamente diferente de μ_0, es decir, que sea menor o mayor (bilateral).

Contraste unilateral inferior (cola izquierda):

Es el caso en que las hipótesis nula y alternativa para la prueba de la media poblacional se formulan de la siguiente forma:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu < \mu_0. \end{aligned}$$

Ahora, como el objetivo final es decidir si la media poblacional es *menor* al valor supuesto, basta con que la muestra tenga una media estadísticamente menor a μ_0 (lateral inferior).

Contraste unilateral superior (cola derecha):

Es el caso en que las hipótesis nula y alternativa para la prueba de la media poblacional se formulan de la siguiente forma:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu > \mu_0. \end{aligned}$$

De la misma forma, como el objetivo final es decidir si la media poblacional es *mayor* al valor supuesto, basta con que la muestra tenga una media estadísticamente mayor a μ_0 (lateral superior).

8.2.4. Determinar el estadístico de contraste

En general, todo número que, obtenido a partir de las observaciones de una muestra, sirve para tomar la decisión sobre rechazar o no H_0, se denomina *estadístico de contraste*. Como esta sesión se enfoca al contraste de la media poblacional, el estadístico de contraste es la media de la muestra. Esta media muestral se ha de tipificar para poder hacer la comparación que permita tomar la decisión. Para tipificar una variable, recordemos que

se le ha de restar la media $E(\bar{X})$ y dividir por su desviación típica $\sigma_{\bar{X}}$, es decir:

$$\bar{X}_{tipificada} = \frac{\bar{X} - E(\bar{X})}{\sigma_{\bar{X}}}.$$

Como ya se ha analizado en la sesión anterior, las propiedades de la distribución de la media muestral, $E(\bar{X}) = \mu$, donde μ es la media de la población, que no conocemos, pero hemos asumido en la hipótesis inicial de que es μ_0. En el ejemplo de la embotelladora de agua $\mu_0 = 50$. También se comprobó de que $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ donde n es el tamaño de la muestra y σ es la desviación típica de la población, que puede ser conocida o no. En el caso de que sea desconocida, se ha de estimar, osea $\hat{\sigma} = S$.

Finalmente, la distribución de la media muestral (normal o t-student) depende de si la población está normalmente distribuida o no, si se conoce la varianza poblacional o no y además, del tamaño de la muestra. Teniendo en cuenta todo lo anterior, en el contraste de hipótesis de la media de una población se pueden tener tres estadísticos de contraste.

Población con distribución normal y varianza conocida

Como la población está normalmente distribuida, la media muestral también lo está y, por tanto:

$$\bar{X} \hookrightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \Rightarrow \qquad z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

donde z_{obs} es el estadístico de contraste.

Población con distribución normal, varianza desconocida y muestra pequeña

Como la población está normalmente distribuida, la media muestral también lo está. Pero, al no conocerse la varianza poblacional σ, se estima con la varianza de la muestra S. Como esta estimación depende de la muestra y su tamaño n, la media muestral no sigue exactamente una distribución normal, sino una distribución similar llamada de *t-Student*, con $n - 1$ grados de libertad, por lo que:

$$\bar{X} \hookrightarrow t_{n-1}\left(\mu, \frac{S^2}{n}\right) \qquad \Rightarrow \qquad t_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}},$$

donde t_{obs} es el estadístico de contraste.

Población con cualquier distribución, varianza desconocida y muestra grande

De acuerdo con el teorema del límite central, independientemente de la distribución de la población, al tener una muestra grande, la distribución de la media muestral se puede aproximar a una normal. Por otra parte, el hecho de no conocer σ determina que la distribución de la media muestral sea la de *t-Student* con $n - 1$ grados de libertad. Sin embargo, como la distribución de *t-Student* tiende a una normal cuando los grados de libertad tienden a infinito, al ser la muestra grande, la distribución de la media muestral se puede aproximar a una normal. Por tanto:

$$\bar{X} \hookrightarrow N\left(\mu, \frac{S^2}{n}\right) \qquad \Rightarrow \qquad z_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}},$$

donde z_{obs} es el estadístico de contraste.

Como se ha visto hasta ahora, nos podemos encontrar con 9 posibles casos dependiendo del tipo de contraste (bilateral, lateral superior o lateral inferior) y del estadístico de contraste (z_{obs} a partir de una población normal y varianza conocida, t_{obs} a partir de una población normal, varianza desconocida y muestra pequeña, o z_{obs} a partir de cualquier distribución, varianza conocida y muestra suficientemente grande). Para continuar con la sesión, nos basaremos en tres ejemplos diferentes, pero que sirven para entender los 9 posibles casos:

Ejemplo 1: Contraste bilateral de una población normal y varianza conocida Una compañía petrolera afirma que, en un yacimiento de petróleo, el área media de los poros de la roca es de 7000 píxeles. Para demostrar que el fabricante se equivoca, se toma una muestra de 12 núcleos cortados en 4 secciones transversales y a cada sección transversal se le midió el área de los poros en píxeles. Los datos se encuentran almacenados en la estructura de datos data.frame llamada *rock* de la librería datasets, que está instalada por defecto en **R**. Suponiendo que esta área está distribuida normalmente con una desviación típica de 1500 píxeles, realizaremos un contraste de hipótesis con un nivel de significancia de 0.05.

Las hipótesis nula y alternativa del test de la media poblacional son:

$$\begin{array}{ll} H_0: & \mu = 7000, \\ H_1: & \mu \neq 7000. \end{array}$$

Por tanto, es un contraste bilateral y el estadístico de contraste es:

$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 7000}{1500/\sqrt{48}}.$$

```
Se inicializa el script
data(rock)  Carga datos
area = rock$area  Define el vector con los datos a utilizar
sigma.area = 1500  Desviación típica de la población
mu.area = 7000  Área media de que se asume que es cierta
alpha.area = 0.05  Nivel de significancia
```

Ejemplo 2: Contraste unilateral inferior de una población normal, varianza desconocida y muestra pequeña El científico Edgar Anderson recogió datos para cuantificar la variación morfológica de la flor *Iris* de tres especies relacionadas: setosa, virgínica y versicolor. Midió cuatro rasgos de cada muestra: el largo y ancho del sépalo y del pétalo, en centímetros. Mediante el contraste de hipótesis con un nivel de significancia del 0.1, demuestra que la longitud media del sépalo de la variedad setosa es mayor que 1.5 cm, suponiendo que dicha longitud se distribuye normalmente. Utiliza las 10 primeras observaciones de la estructura de datos llamada *Iris* de la librería datasets.

Las hipótesis nula y alternativa del test de la media poblacional son:

$$\begin{array}{ll} H_0: & \mu = 1.5, \\ H_1: & \mu < 1.5. \end{array}$$

Por tanto, es un contraste unilateral inferior y el estadístico de contraste es:

$$t_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - 1.5}{S/\sqrt{10}}.$$

```
Se inicializa el script
data(iris")  Carga datos
sepal = iris$Petal.Length[iris$Species=="setosa"]
sepal = sepal[1:10]  Define vector con los datos a utilizar
mu.sepal = 1.5  Longitud media que se asume que es cierta
alpha.sepal = 0.1  Nivel de significancia
```

Ejemplo 3: Contraste unilateral superior de una población con cualquier distribución, varianza desconocida y muestra suficientemente grande El Ayuntamiento de Nueva York afirma que la concentración de ozono en el aire de la ciudad es menor que 37 partes por billón. Un grupo ambientalista quiere refutar esta afirmación mediante un contraste con un nivel de significancia de 0.01 y toma 116 medidas de la concentración de ozono. Los datos se encuentran almacenados en la estructura de datos llamada *airquality* de la librería datasets.

Las hipótesis nula y alternativa del test de la media poblacional son:

$$\begin{array}{ll} H_0: & \mu = 37, \\ H_1: & \mu > 37. \end{array}$$

Por tanto, es un contraste unilateral superior y el estadístico de contraste es:

$$z_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - 37}{S/\sqrt{116}}.$$

```
Se inicializa el script
data(.ªirquality")  Carga datos
ozono = as.numeric(na.omit(airquality[,1]))  Define vector con
los datos a utilizar
mu.ozono = 37  Concentración media que se asume que es cierta
alpha.ozono = 0.01  Nivel de significancia
```

8.2.5. Definir el criterio de decisión

La pregunta planteada hasta el momento es si la media de la muestra $\bar{x}$ tiene un valor igual, mayor o menor (según el tipo de contraste) al valor que se ha supuesto de la media de la población. Esta es una comparación estadística o probabilística, lo cual implica que, teniendo en cuenta el nivel de significancia, la distribución de probabilidad y el tipo de contraste, se definen el intervalo o los límites en que se considera si estas medias son iguales o cuál es mayor o menor. Este criterio se denomina *criterio de decisión basado en el valor crítico*, donde se determinan las zonas de rechazo o fallo al rechazo de la hipótesis planteada.

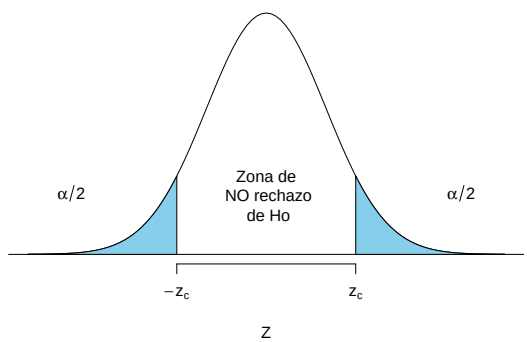
Otra forma de tomar la decisión de la comparación es calcular la probabilidad de que, si tomamos otra muestra, su estadístico correspondiente sea mayor o menor, mayor que, menor que (según el tipo de contraste) del estadístico observado. Esta probabilidad se denomina **p-valor** y se compara con el nivel de significancia del contraste (α). Es el *criterio de decisión basado en el p-valor*.

Como se ha dicho anteriormente, la definición de las zonas de rechazo / no rechazo y el cálculo del *p*-valor dependen de la distribución del estadístico, entre otros factores. Para el contraste de la media poblacional, el estadístico de contraste puede tener una distribución normal o una de *t-Student* con $n-1$ grados de libertad, dependiendo de la distribución de la población, de si se conoce o no la varianza de la población y del tamaño de la muestra. En cualquier caso, la forma de la distribución es similar: la campana gaussiana. Por tanto, para definir estas zonas, nos basamos en que el estadístico de contraste es z_{obs}, pero es igualmente válido si el estadístico de contraste es t_{obs}.

Por otra parte, se ha mencionado que las zonas y el *p*-valor dependen también del tipo de contraste. Así pues, se explica a continuación cómo se definen estos criterios según cada caso, utilizando los ejemplos descritos anteriormente.

Criterio de decisión basado en el valor crítico

Contraste bilateral (dos colas) Como la hipótesis alternativa es $H_1 : \mu \neq \mu_0$, y teniendo en cuenta el nivel de significancia α, los límites del intervalo de la zona de rechazo de H_0 y, por tanto, de la de aceptación de H_1, están definidos por el valor crítico $\pm z_c$ tal que:



$$P(-z_c < Z < z_c) = 1 - \alpha.$$

Es decir, se rechaza H_0 cuando el estadístico observado es menor que el valor crítico negativo ($z_{obs} < -z_c$) o mayor que el valor crítico positivo ($z_{obs} > z_c$).

En el ejemplo 1, se tiene un contraste bilateral de una población normal con varianza conocida y el estadístico de contraste es z_{obs} ya que la media muestral está distribuida normalmente. Por lo tanto, el valor crítico y la zona de no rechazo son:

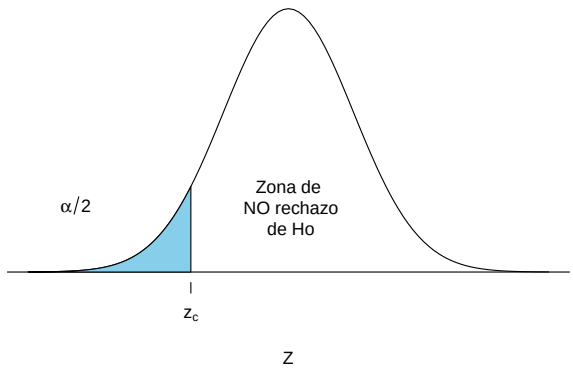
```
Intervalo de no rechazo para el área media de los poros de la roca
zc.area = qnorm(alpha.area/2, lower.tail=FALSE)
No_rec = c(-zc.area,zc.area); No_rec
```

```
## [1] -1.959964  1.959964
```

Por tanto, H_0 se rechaza si el estadístico de la muestra tipificado z_{obs} es menor que -1.96 o mayor que 1.96.

Contraste unilateral inferior (cola izquierda) Como la hipótesis alternativa es $H_1 : \mu < \mu_0$, y teniendo en cuenta el nivel de significancia α, los límites del intervalo de la zona de rechazo de H_0 y, por tanto, de la de aceptación de H_1, están definidos por el valor crítico z_c tal que:

$$P(z_c < Z) = 1 - \alpha.$$



Es decir, se rechaza H_0 cuando el estadístico observado es menor que el valor crítico ($z_{obs} < z_c$).

En el ejemplo 2, se tiene un contraste unilateral inferior de una población normal con varianza desconocida, el estadístico de contraste es t_{obs}, ya que la media muestral sigue una distribución de *t-Student* con 9 grados de libertad (el tamaño de la muestra es 10). Por tanto, el valor crítico y la zona de no rechazo son:

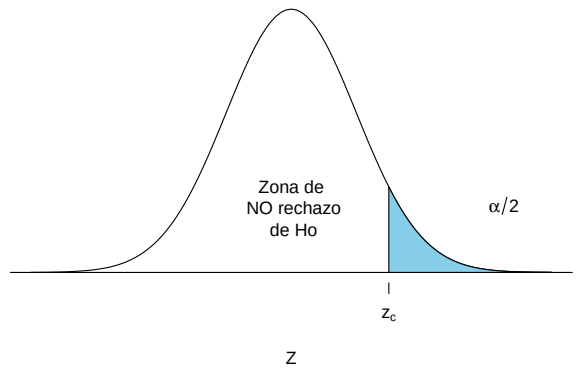
```
Intervalo de no rechazo para la longitud media del sépalo
de la variedad setosa
n.sepal = length(sepal)  Tamaño de la muestra
tc.sepal = qt(alpha.sepal, df=n.sepal-1, lower.tail=TRUE)
No_rec = c(tc.sepal, Ïnf"); No_rec
```

```
## [1] "-1.38302873839663" "Inf"
```

Por tanto, H_0 se rechaza si el estadístico de la muestra tipificado t_{obs} es menor que -1.383.

Contraste unilateral superior (cola derecha): Como la hipótesis alternativa es H_1 : $\mu > \mu_0$, y teniendo en cuenta el nivel de significancia α, los límites del intervalo de la zona de rechazo de H_0 y, por tanto, de la de aceptación de H_1, están definidos por el valor crítico z_c tal que:

$$P(Z < z_c) = 1 - \alpha.$$



Es decir, se rechaza H_0 cuando el estadístico observado es mayor que el valor crítico ($z_{obs} > z_c$).

En el ejemplo 3, se tiene un contraste unilateral superior de una población con distribución y varianza desconocida pero con una muestra suficientemente grande. El estadístico de contraste es z_{obs} ya que, según el teorema del límite central, la distribución de la media muestral se aproxima a una normal. Por tanto, el valor crítico y la zona de no rechazo son:

```
Intervalo de no rechazo para la concentración media de ozono
zc.ozono = qnorm(alpha.ozono, lower.tail=FALSE)
No_rec = c(Inf", zc.ozono); No_rec
```

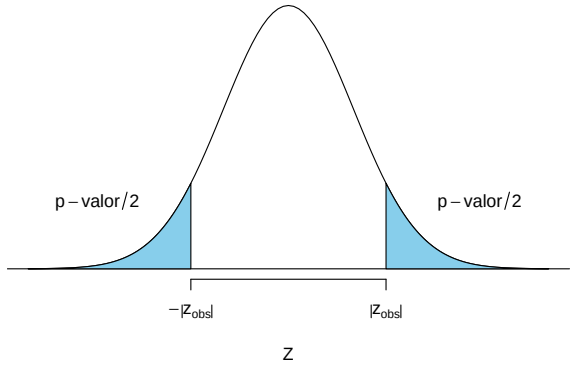
```
## [1] "-Inf"      "2.32634787404084"
```

Por tanto, H_0 se rechaza si el estadístico de la muestra tipificado z_{obs} es mayor que 2.326.

Criterio de decisión basado en el p-valor

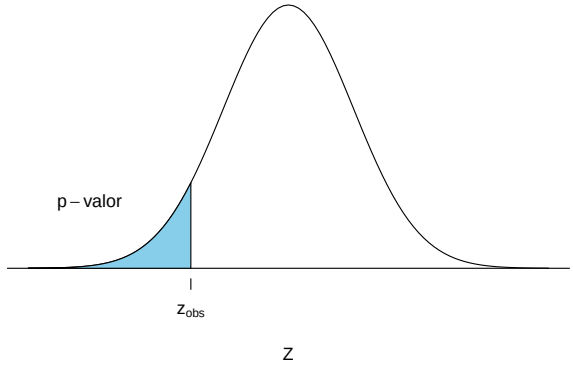
Contraste bilateral (dos colas): Como la hipótesis alternativa es $H_1 : \mu \neq \mu_0$, el p-valor de la muestra observada viene dado por:

$$p - valor = P(-|z_{obs}| < Z < |z_{obs}|) = 2P(Z > |z_{obs}|) = 2P(Z < -|z_{obs}|).$$



Contraste unilateral inferior (cola izquierda): Como la hipótesis alternativa es $H_1 : \mu < \mu_0$, el p-valor de la muestra observada viene dado por:

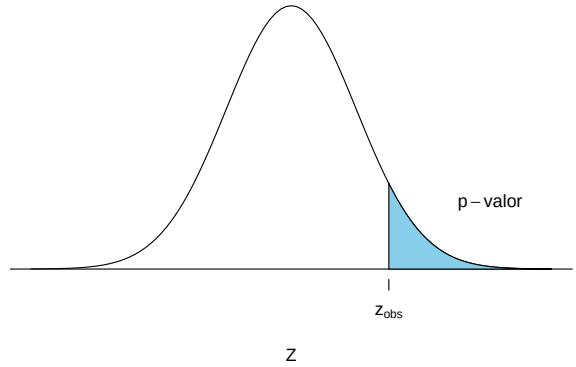
$$p - valor = P(Z < z_{obs}).$$



Contraste unilateral superior (cola derecha): Como la hipótesis alternativa es $H_1 : \mu > \mu_0$, el p-valor de la muestra observada está dado por:

$$p - valor = P(Z > z_{obs}).$$

En cualquiera de los casos, se rechaza H_0 cuando el p-valor es menor que el nivel de significancia ($p - value < \alpha$).



8.2.6. Calcular el estadístico observado (de la muestra) y su p-valor

En el ejemplo 1, el estadístico de contraste es $z_{obs} = \dfrac{\bar{x} - 7000}{1500/\sqrt{48}}$ ya que la media muestral está distribuida normalmente. Teniendo en cuenta la muestra dada, el estadístico observado y su *p*-valor vienen dados por:

```
Estadístico y p-valor para el ejemplo 1
n.area = length(area)  Tamaño de la muestra
xbar.area = mean(area)  Media de la muestra
zob.area = (xbar.area-mu.area)/(sigma.area/sqrt(n.area));
zob.area  estadístico observado
```

```
## [1] 0.8670839
```

```
pvalue.area = 2*pnorm(zob.area,lower.tail=FALSE);
pvalue.area  p-valor
```

```
## [1] 0.3858961
```

En el ejemplo 2, el estadístico de contraste es $t_{obs} = \dfrac{\bar{x} - 1.5}{S/\sqrt{10}}$ ya que la media muestral está distribuida según la distribución *t-Ctudent* con 9 grados de libertad. Teniendo en cuenta la muestra dada, el estadístico observado y su *p*-valor vienen dados por:

```
Estadístico y p-valor para el ejemplo 2
xbar.sepal = mean(sepal)  Media de la muestra
S.sepal = sd(sepal)  Desviación típica de la muestra
tob.sepal = (xbar.sepal-mu.sepal)/(S.sepal/sqrt(n.sepal)); tob.sepal
```

```
## [1] -1.46385
```

```
pvalue.sepal = pt(tob.sepal,df=n.sepal-1,lower.tail=TRUE); pvalue.sepal
```

```
## [1] 0.08863385
```

En el ejemplo 3, el estadístico de contraste es $z_{obs} = \dfrac{\bar{x} - 37}{S/\sqrt{116}}$ ya que, según el teorema del límite central, la media muestral se aproxima a una distribución normal. Teniendo en cuenta la muestra dada, el estadístico observado y su p-valor vienen dados por:

```
Estadístico y p-valor para el ejemplo 3
n.ozono = length(ozono)  Tamaño de la muestra
xbar.ozono = mean(ozono)  Media de la muestra
S.ozono = sd(ozono)  Desviación típica de la muestra
zob.ozono = (xbar.ozono-mu.ozono)/(S.ozono/sqrt(n.ozono)); zob.ozono
```

```
## [1] 1.674686
```

```
pvalue.ozono = pnorm(zob.ozono, lower.tail=FALSE); pvalue.ozono
```

```
## [1] 0.04699788
```

El estadístico observado y el p-valor del contraste de hipótesis para la media poblacional también se pueden calcular por medio de las funciones z.test() de la librería TeachingDemos o t.test() de la librería stats que hemos estudiado en la sesión anterior para el cálculo de los intervalos de confianza. Para calcular el intervalo, solo se introducen los datos de la muestra y la desviación típica de la población (solo para z.test()) y si se quiere cambiar el nivel de significancia que está por defecto. Ahora, para realizar el contraste, también se ha de estipular el valor de la media para la hipótesis nula (mu=) y el tipo de contraste o la hipótesis alternativa (alternative=) cuyas opciones son: two.sided, less, greater. De esta forma:

Ejemplo 1:

```
library(TeachingDemos)
z.test(area, sd=sigma.area, mu=mu.area) Contraste para el ejemplo 1
```

```
## One Sample z-test
##
```

```
## data:  area
## z = 0.86708, n = 48.00, Std. Dev. = 1500.00, Std. Dev. of the sample
## mean = 216.51, p-value = 0.3859
## alternative hypothesis: true mean is not equal to 7000
## 95 percent confidence interval:
## 6763.385 7612.074
## sample estimates:
## mean of area
## 7187.729
```

Entre otras cosas, la función nos retorna el $z_{obs} = 0.86708$ y $p - valor = 0.3859$, que son iguales a los calculados previamente.

Ejemplo 2:

```
t.test(sepal, mu=mu.sepal, alternative="less", conf.level=0.9)
Contraste para el ejemplo 2
```

```
## One Sample t-test
##
## data:  sepal
## t = -1.4639, df = 9, p-value = 0.08863
## alternative hypothesis: true mean is less than 1.5
## 90 percent confidence interval:
## -Inf 1.497239
## sample estimates:
## mean of x
## 1.45
```

Observamos también que el estadístico observado y el p-valor coinciden con los calculados anteriormente $t_{obs} = -1.4639$, $p - valor = 0.08863$.

Ejemplo 3:

```
t.test(ozono, mu=mu.ozono, alternative="greater", conf.level=0.99)
```

```
## One Sample t-test
##
## data:  ozono
## t = 1.6747, df = 115, p-value = 0.04836
## alternative hypothesis: true mean is greater than 37
## 99 percent confidence interval:
## 34.9034     Inf
## sample estimates:
## mean of x
## 42.12931
```

```
z.test(ozono, sd=S.ozono, mu=mu.ozono, alternative="greater",
conf.level=0.99)
```

```
## One Sample z-test
##
## data:  ozono
## z = 1.6747, n = 116.0000,
## Std. Dev. = 32.9879,
## Std.Dev. of the sample
## mean = 3.0628,
## p-value = 0.047
## alternative hypothesis: true mean is greater than 37
## 99 percent confidence interval:
## 35.00406      Inf
## sample estimates:
## mean of ozono
## 42.12931
```

Como la muestra es grande, se puede observar la buena aproximación entre los estadísticos observados y su *p*-valor ($z_{obs} = 1.6747, p - valor = 0.047$).

8.2.7. Rechazar o no la hipótesis inicial (resultado del contraste)

El resultado del contraste consiste en rechazar o no la hipótesis inicial de acuerdo con el criterio de decisión adoptado y el estadístico calculado a partir de la muestra.

En el ejemplo 1, se tiene que $\alpha = 0.05$ y se ha calculado $z_c = 1.96$, $z_{obs} = 0.867$ y $p - valor = 0.386$. Si tenemos en cuenta el criterio del valor crítico, H_0 no se puede rechazar ya que $-z_c < z_{obs} < z_c$, es decir, z_{obs} está en la zona de no rechazo. Por otra parte, si consideramos el criterio del *p*-valor, como $p - valor > \alpha$, el resultado es el mismo: **NO rechazar** H_0.

En el ejemplo 2, se tiene que $\alpha = 0.1$ y se ha calculado $t_c = -1.38$, $t_{obs} = -1.464$ y $p - valor = 0.089$. Si tenemos en cuenta el criterio del valor crítico, H_0 se rechaza ya que $t_{obs} < t_c$, es decir, t_{obs} está en la zona de rechazo. Por otra parte, si consideramos el criterio del *p*-valor, como $p - valor < \alpha$, el resultado es el mismo, **rechazar** H_0. Realiza el contraste tomando otra muestra también de 10 observaciones. ¿Se obtiene el mismo resultado?

Finalmente, en el ejemplo 3, se tiene que $\alpha = 0.01$ y se ha calculado $z_c = 2.326$, $z_{obs} = 1.675$ y $p - valor = 0.047$. Si tenemos en cuenta el criterio del valor crítico, H_0 no se rechaza ya que $z_{obs} < z_c$, es decir, z_{obs} está en la zona de no rechazo. Por otra parte, si consideramos el criterio del *p*-valor, como $p - valor > \alpha$, el resultado, como es de esperar, es el mismo: **NO rechazar** H_0.

8.2.8. Concluir

Para finalizar, es altamente recomendable hacer una declaración para interpretar la decisión en el contexto del problema original. Si el resultado del contraste es rechazar H_0, la conclusión es que hay suficiente evidencia para rechazar que la media es igual a μ_0 (mayor que o menor que μ_0, según el caso). Por tanto, hay suficiente evidencia para defender que la media es diferente que μ_0 (menor que o mayor que μ_0, según el caso) .

Si, por el contrario, el resultado es no rechazar H_0, se concluye que no hay suficiente evidencia para rechazar que la media sea igual a μ_0 (mayor que o menor que μ_0, según el caso), pero tampoco para defender que la media es diferente (menor que o mayor que μ_0, según el caso) que μ_0. En este supuesto, ya que no se puede afirmar nada, ¿qué crees que se debería hacer en una situación real?

Según los resultados obtenidos en los ejemplos, se concluye lo siguiente: para el ejemplo 1, no hay suficiente evidencia para rechazar que la media del área de los polos de la roca del yacimiento sea igual a 7000 píxeles, pero tampoco la hay para defender que la media del área sea diferente de 7000 píxeles.

Por otra parte, en el ejemplo 2 se concluye que hay suficiente evidencia para rechazar que la media de la longitud del sépalo de la variedad *setosa* es igual a 1.5 cm o mayor; por tanto, hay suficiente evidencia para defender que la media de la longitud es menor que 1.5 cm.

Finalmente, en el ejemplo 3, la conclusión es que no hay suficiente evidencia para rechazar que la media de la concentración de ozono sea igual o menor que 37, pero tampoco la hay para defender que la media de la concentración sea mayor que 37.

Tips & Tricks!

- En la formulación de las hipótesis hay que tener en cuenta:
  - Fallar en el rechazo de H_0 no significa que H_0 se acepte como cierto
  - Si se quiere reafirmar algo, hay que ponerlo en H_1
  - Si se quiere desmentir, hay que ponerlo en H_0
- z.test(data, sd=, alternative=) realiza el contraste de hipótesis para la media de la población (a partir de la muestra data) cuando la varianza de la población es conocida o la muestra es grande. Se ha de especificar el tipo de contraste en el parámetro alternative = y las opciones son: two.sided, less o greater. Pertenece a la librería TeachingDemos, que se ha de cargar previamente.
- t.test(data, alternative=) realiza el contraste de hipótesis para la media de la población (a partir de la muestra data) cuando la población se distribuye normalmente con varianza desconocida. Igualmente, se ha de especificar el tipo de contraste.

8.3. Ejercicios propuestos

1. Se especifica que un cierto tipo de hierro debe contener 0.85 g de silicio por cada 100 g de hierro (0.85 %). Se ha determinado el contenido de silicio (normalmente distribuido) de cada una de las 25 muestras de hierro seleccionadas al azar y se ha hallado que la media es de 0.888 y la desviación típica es de 0.1807. ¿Admite que el contenido de silicio es diferente al deseado con un nivel de significación de 0.05?

2. De una muestra de 50 lentes utilizadas en gafas, se obtiene un espesor medio de 3.05 mm y una desviación típica de la muestra de 0.34 mm. El grosor promedio verdadero deseado de dichas lentes es de 3.20 mm. ¿Los datos sugieren fuertemente que el grosor medio real de tales lentes es algo distinto de lo que se desea? Demuéstralo utilizando un nivel de significancia de 0.05.

3. Un fabricante afirma que la resistencia a la rotura media de sus bandas de goma es de 3 kg. Un investigador sospecha que la afirmación del fabricante es demasiado alta. Una muestra de 15 bandas de goma produce una resistencia a la rotura media de 2.6 kg y una desviación estándar de 1.1 kg. Sabiendo que la resistencia a la rotura se distribuye normalmente, al realizar una prueba de contraste con un nivel de significancia de 0.01, ¿decidimos que la afirmación del fabricante es demasiado alta? Y si la prueba se hace con un nivel de significancia de 0.1, ¿qué decidimos?

4. Un ingeniero mide la dureza Brinell de 25 piezas de hierro dúctil recocidas subcríticamente. Los datos resultantes son:

   170 167 174 179 179 156 163 156 187 156 183 179 174 179 170 156 187 179 183 174 187 167 159 170 179

   El ingeniero plantea la hipótesis de que la dureza Brinell media de todas esas piezas de hierro dúctil es mayor que 170. Por tanto, le interesa en probar las hipótesis con un nivel de significancia de 0.05. ¿A qué conclusión llega?