

Gráficos con R

Juan Carlos Correa Morales

Freddy Hernández Barajas

2016-12-01

Índice general

Prefacio	5
1. Introducción	7
1.1. Orígenes	7
1.2. Descarga e instalación	8
1.3. Apariencia del programa	9
1.4. Tipos de objetos	9
1.5. Guía de estilo para la escritura en R	12
2. Gráficos para una variable	17
2.1. Función <code>stem</code>	17
2.2. Función <code>boxplot</code>	18
2.3. Función <code>hist</code>	19

Prefacio

Este libro ...

Capítulo 1

Introducción

1.1. Orígenes

R es un lenguaje de programación usado para realizar procedimientos estadísticos y gráficos de alto nivel, este lenguaje fue creado en 1993 por los profesores e investigadores Robert Gentleman y Ross Ihaka. Inicialmente el lenguaje se usó para apoyar los cursos que tenían a su cargo los profesores, pero luego de ver la utilidad de la herramienta desarrollada, decidieron colocar copias de R en StatLib. A partir de 1995 el código fuente de R está disponible bajo licencia GNU GPL para sistemas operativos Windows, Macintosh y distribuciones Unix/Linux. La comunidad de usuarios de R en el mundo es muy grande y los usuarios cuentan con diferentes espacios para interactuar, a continuación una lista no exhaustiva de los sitios más populares relacionados con R:

- Rbloggers: <https://www.r-bloggers.com/>.
- Comunidad hispana de R: <http://r-es.org/>.
- Nabble: <http://r.789695.n4.nabble.com/>.
- Foro en portugués: <http://r-br.2285057.n4.nabble.com/>.
- Stackoverflow: <http://stackoverflow.com/questions/tagged/r>.
- Cross Validated: <http://stats.stackexchange.com/questions/tagged/r>.
- R-Help Mailing List: <https://stat.ethz.ch/mailman/listinfo/r-help>.
- Revolutions: <http://blog.revolutionanalytics.com/>.
- R-statistics blog: <https://www.r-statistics.com/>.
- RDataMining: <https://rdatamining.wordpress.com/>.



Figura 1.1: Robert Gentleman (izquierda) y Ross Ihaka (derecha) creadores de R.

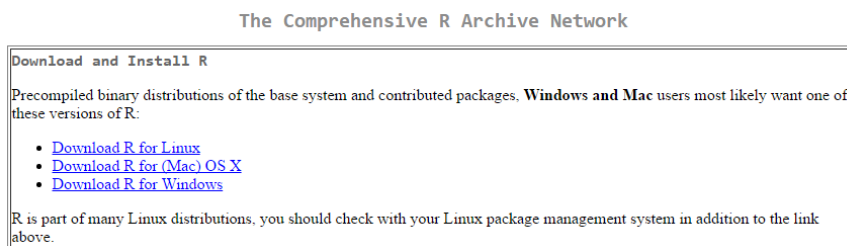


Figura 1.2: Página del Cran.

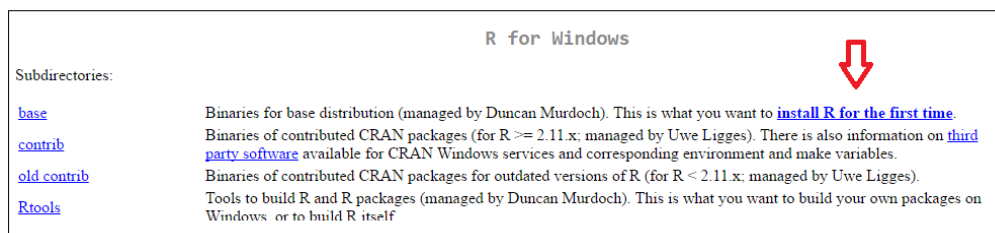


Figura 1.3: Página de instalación para la primera ocasión.

1.2. Descarga e instalación

Para realizar la instalación de R usted debe visitar la página del CRAN (*Comprehensive R Archive Network*) disponible en este enlace <https://cran.r-project.org/>. Una vez ingrese a la página encontrará un cuadro similar al mostrado en la Figura 1.2 donde aparecen los enlaces de la instalación para los sistemas operativos Linux, Mac y Windows.

Supongamos que se desea instalar R en Windows, para esto se debe dar clic sobre el hipervínculo **Download R for Windows** de la Figura 1.2. Una vez hecho esto se abrirá una página con el contenido mostrado en la Figura 1.3. Una vez ingrese a esa nueva página usted debe dar clic sobre el hipervínculo **install R for the first time** como es señalado por la flecha roja en la Figura 1.3.

Luego de esto se abrirá otra página con un encabezado similar al mostrado en la Figura 1.4, al momento de capturar la figura la versión actual de R era 3.2.5 pero seguramente en este momento usted tendrá disponible una versión actualizada. Una vez allí usted debe dar clic sobre **Download R 3.2.5 for Windows** como es señalado por la flecha verde. Luego de esto se descargará el instalador R en el computador el cual deberá ser instalado con las opciones que vienen por defecto.

Se recomienda observar el siguiente video didáctico de instalación de R disponible en este enlace <http://tinyurl.com/jd7b9ks> para facilitar la tarea de instalación.

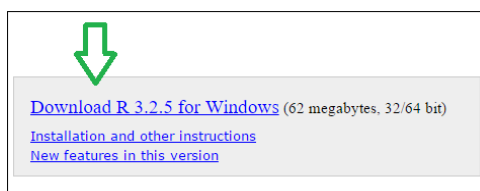


Figura 1.4: Página de descarga.



Figura 1.5: Apariencia del acceso directo para ingresar a R.

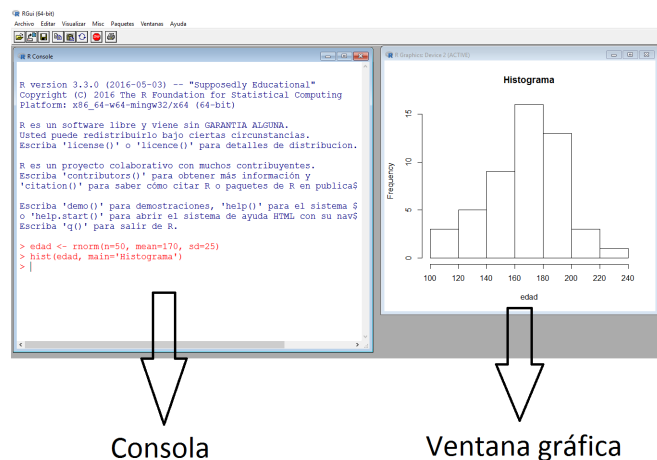


Figura 1.6: Apariencia de R.

1.3. Apariencia del programa

Una vez que esté instalado R en su computador, usted podrá acceder a él por la lista de programas o por medio del acceso directo que quedó en el escritorio, en la Figura 1.5 se muestra la apariencia del acceso directo para ingresar a R.

Al abrir R lo que aparecerá en la pantalla de su computador será algo muy similar a lo que está en la Figura 1.6. La ventana izquierda se llama consola y es donde se ingresan las instrucciones, una vez que se construye un gráfico se activa otra ventana llamada ventana gráfica. Cualquier usuario puede modificar la posición y tamaños de estas ventanas, puede cambiar el tipo y tamaño de las letras en la consola, para hacer esto se deben explorar las opciones de *editar* en la barra de herramientas.

1.4. Tipos de objetos

En R existen varios tipos de objetos que permiten que el usuario pueda almacenar la información para realizar procedimientos estadísticos y gráficos. Los principales objetos en R son vectores, matrices, arreglos, marcos de datos y listas. A continuación se presentan las características de estos objetos y la forma para crearlos.

1.4.1. Vectores

Los vectores son arreglos ordenados en los cuales se puede almacenar información de tipo numérico (variable cuantitativa), alfanumérico (variable cualitativa) o lógico (TRUE o FALSE), pero no mezclas de éstos. La función de R para crear un vector es `c()` y que significa concatenar; dentro de los paréntesis de esta función

se ubica la información a almacenar. Una vez construido el vector se acostumbra a etiquetarlo con un nombre corto y representativo de la información que almacena, la asignación se hace por medio del operador `<-` entre el nombre y el vector.

A continuación se presenta un ejemplo de cómo crear tres vectores que contienen las respuestas de cinco personas a tres preguntas que se les realizaron.

```
edad <- c(15, 19, 13, NA, 20)
deporte <- c(TRUE, TRUE, NA, FALSE, TRUE)
comic.fav <- c(NA, 'Superman', 'Batman', NA, 'Batman')
```

El vector `edad` es un vector cuantitativo y contiene las edades de las 5 personas. En la cuarta posición del vector se colocó el símbolo `NA` que significa *Not Available* debido a que no se registró la edad para esa persona. Al hacer una asignación se acostumbra a dejar un espacio antes y después del operador `<-` de asignación. El segundo vector es llamado `deporte` y es un vector lógico que almacena las respuestas a la pregunta de si la persona practica deporte, nuevamente aquí hay un `NA` para la tercera persona. El último vector `comic.fav` contiene la información del cómic favorito de cada persona, como esta variable es cualitativa es necesario usar las comillas `' '` para encerrar las respuestas. Cuando se usa `NA` para representar una información *Not Available* NO SE DEBEN usar las comillas `' '`.

Si se desea ver lo que está almacenado en cada uno de estos vectores, se debe escribir en la consola de R el nombre de uno de los objetos y luego se presiona la tecla *enter* o *intro*, al realizar esto lo que se obtiene se muestra a continuación.

```
edad
```

```
## [1] 15 19 13 NA 20
```

```
deporte
```

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

```
comic.fav
```

```
## [1] NA "Superman" "Batman" NA "Batman"
```

1.4.2. Matrices

Las matrices son arreglos rectangulares de filas y columnas con información numérica, alfanumérica o lógica. Para construir una matriz se usa la función `matrix()`. Por ejemplo, para crear una matriz de 4 filas y 5 columnas (de dimensión 4×5) con los primeros 20 números positivos se escribe el código siguiente en la consola.

```
mimatriz <- matrix(data=1:20, nrow=4, ncol=5, byrow=FALSE)
```

El argumento `data` de la función sirve para indicar los datos que se van a almacenar en la matriz, los argumentos `nrow` y `ncol` sirven para definir la dimensión de la matriz y por último el argumento `byrow` sirve para indicar si la información contenida en `data` se debe ingresar por filas o no. Para observar lo que quedó almacenado en el objeto `mimatriz` se escribe en la consola el nombre del objeto seguido de la tecla *enter* o *intro*.

```
mimatriz
```

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

1.4.3. Arreglos

Un arreglo es una matriz de varias dimensiones con información numérica, alfanumérica o lógica. Para construir una arreglo se usa la función `array()`. Por ejemplo, para crear un arreglo de $3 \times 4 \times 2$ con las primeras 24 letras minúsculas del alfabeto se escribe el siguiente código.

```
miarrray <- array(data=letters[1:24], dim=c(3, 4, 2))
```

El argumento `data` de la función sirve para indicar los datos que se van a almacenar en el arreglo y el argumento `dim` sirve para indicar las dimensiones del arreglo. Para observar lo que quedó almacenado en el objeto `miarrray` se escribe en la consola lo siguiente.

```
miarrray
```

```
## , , 1
##
##      [,1] [,2] [,3] [,4]
## [1,] "a"  "d"  "g"  "j"
## [2,] "b"  "e"  "h"  "k"
## [3,] "c"  "f"  "i"  "l"
##
## , , 2
##
##      [,1] [,2] [,3] [,4]
## [1,] "m"  "p"  "s"  "v"
## [2,] "n"  "q"  "t"  "w"
## [3,] "o"  "r"  "u"  "x"
```

1.4.4. Marco de datos

El marco de datos o *data frame* es uno de los objetos más utilizados porque permite agrupar vectores con información de diferente tipo (numérica, alfanumérica o lógica) en un mismo objeto, la única restricción es que los vectores deben tener la misma longitud. Para crear un marco de datos se usa la función `data.frame()`, como ejemplo vamos a crear un marco de datos con los vectores `edad`, `deporte` y `comic.fav` definidos anteriormente.

```
mimarco <- data.frame(edad, deporte, comic.fav)
```

Una vez creado el objeto `mimarco` podemos ver el objeto escribiendo su nombre en la consola, a continuación se muestra lo que se obtiene.

```
mimarco
```

```
##   edad deporte comic.fav
## 1   15     TRUE    <NA>
## 2   19     TRUE  Superman
## 3   13      NA   Batman
## 4   NA    FALSE    <NA>
## 5   20     TRUE   Batman
```

De la salida anterior vemos que el marco de datos tiene 3 variables (columnas) cuyos nombres coinciden con los nombres de los vectores creados anteriormente, los números consecutivos al lado izquierdo son sólo de referencia y permiten identificar la información para cada persona en la base de datos.

1.4.5. Listas

Las listas son otro tipo de objeto muy usado para almacenar objetos de diferente tipo. La instrucción para crear una lista es `list()`. A continuación vamos a crear una lista que contiene tres objetos: un vector con 5 números aleatorios llamado `mivector`, una matriz de dimensión 6×2 con los primeros doce números enteros positivos llamada `matriz2` y el tercer objeto será el marco de datos `mimarco` creado en el apartado anterior. Las instrucciones para crear la lista requerida se muestran a continuación.

```
set.seed(12345)
mivector <- runif(n=5)
matriz2 <- matrix(data=1:12, ncol=6)
milista <- list(E1=mivector, E2=matriz2, E3=mimarco)
```

La función `set.seed` de la línea número 1 sirve para fijar la semilla de tal manera que los números aleatorios generados en la segunda línea con la función `runif` sean siempre los mismos. En la última línea del código anterior se construye la lista, dentro de la función `list` se colocan los tres objetos `mivector`, `matriz2` y `mimarco`. Es posible colocarle un nombre especial a cada uno de los elementos de la lista, en este ejemplo se colocaron los nombres `E1`, `E2` y `E3` para cada uno de los tres elementos. Para observar lo que quedó almacenado en la lista se escribe `milista` en la consola y el resultado se muestra a continuación.

```
milista

## $E1
## [1] 0.7209039 0.8757732 0.7609823 0.8861246 0.4564810
##
## $E2
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    3    5    7    9   11
## [2,]    2    4    6    8   10   12
##
## $E3
##   edad deporte comic.fav
## 1   15     TRUE      <NA>
## 2   19     TRUE   Superman
## 3   13      NA     Batman
## 4   NA    FALSE      <NA>
## 5   20     TRUE     Batman
```

1.5. Guía de estilo para la escritura en R

Así como en el español existen reglas ortográficas, la escritura de códigos en R también tiene unas reglas que se recomienda seguir para evitar confusiones. Tener una buena guía de estilo es importante para que el código creado por usted sea fácilmente entendido por sus lectores (Wickham, 2015). No existe una única y mejor guía de estilo para escritura en R, sin embargo aquí vamos a mostrar unas sugerencias basadas en la guía llamada *Google's R style guide*.

1.5.1. Nombres de los archivos

Se sugiere que el nombre usado para nombrar un archivo tenga sentido y que termine con extensión `.R`. A continuación dos ejemplos de como nombrar mal y bien un archivo.

- Mal: `hola.R`
- Bien: `analisis_icfes.R`

1.5.2. Nombres de los objetos

Se recomienda no usar los símbolos `_` y `-` dentro de los nombres de objetos. Para las variables es preferible usar letras minúsculas y separar las palabras con puntos (`peso.maiz`) o utilizar la notación camello iniciando en minúscula (`pesoMaiz`). Para las funciones se recomienda usar la notación camello iniciando todas las palabras en mayúscula (`PlotRes`). Para los nombres de las constantes se recomienda que inicien con la letra `k` (`kPrecioBus`). A continuación ejemplos de buenas y malas prácticas.

Para variables:

- Bien: `avg.clicks`
- Aceptable: `avgClicks`
- Mal: `avg_Clicks`

Para funciones:

- Bien: `CalculateAvgClicks`
- Mal: `calculate_avg_clicks`, `calculateAvgClicks`

1.5.3. Longitud de una línea de código

Se recomienda que cada línea tenga como máximo 80 caracteres. Si una línea es muy larga se debe cortar siempre por una coma.

1.5.4. Espacios

Use espacios alrededor de todos los operadores binarios (`=`, `+`, `-`, `<-`, etc.). Los espacios alrededor del símbolo `"="` son opcionales cuando se usan para ingresar valores dentro de una función. Así como en español, nunca coloque espacio antes de una coma, pero siempre use espacio luego de una coma. A continuación ejemplos de buenas y malas prácticas.

```
tab <- table(df[df$days < 0, 2]) # Bien
tot <- sum(x[, 1])                # Bien
tot <- sum(x[1, ])               # Bien
tab <- table(df[df$days<0, 2])  # Faltan espacios alrededor '<'
tab <- table(df[df$days < 0,2]) # Falta espacio luego de coma
tab <- table(df[df$days < 0 , 2]) # Sobra espacio antes de coma
tab<- table(df[df$days < 0, 2]) # Falta espacio antes de '<-'
tab<-table(df[df$days < 0, 2]) # Falta espacio alrededor de '<-'
tot <- sum(x[,1])                # Falta espacio luego de coma
tot <- sum(x[1,])                # Falta espacio luego de coma
```

Otra buena práctica es colocar espacio antes de un paréntesis excepto cuando se llama una función.

```
if (debug)      # Correcto
if(debug)      # Funciona pero no se recomienda
colMeans (x)   # Funciona pero no se recomienda
```

Espacios extras pueden ser usados si con esto se mejora la apariencia del código, ver el ejemplo siguiente.

```
plot(x    = x.coord,
     y    = data.mat[, MakeColName(metric, ptilas[1], "roi0pt")],
     ylim = ylim,
     xlab = "dates",
```

```
ylab = metric,
main = (paste(metric, " for 3 samples ", sep = "")))
```

No coloque espacios alrededor del código que esté dentro de paréntesis ()' o corchetes []", la única excepción es luego de una coma, ver el ejemplo siguiente.

```
if (condicion)      # Correcto
x[1, ]             # Correcto
if ( condicion )   # Sobran espacios alrededor de condicion
x[1,]              # Se necesita espacio luego de coma
```

Los signos de agrupación llaves \{ \}' se utilizan para agrupar bloques de código y se recomienda que nunca una llave abierta{' esté sola en una línea; una llave cerrada "}" si debe ir sola en su propia línea. Se pueden omitir las llaves cuando el bloque de instrucciones esté formado por una sola línea pero esa línea de código NO debe ir en la misma línea de la condición. A continuación dos ejemplos de lo que se recomienda.

```
if (is.null(ylim)) {                                # Correcto
  ylim <- c(0, 0.06)
}

if (is.null(ylim))                                  # Correcto
  ylim <- c(0, 0.06)

if (is.null(ylim)) ylim <- c(0, 0.06)              # Aceptable

if (is.null(ylim))                                  # No se recomienda
{
  ylim <- c(0, 0.06)
}

if (is.null(ylim)) {ylim <- c(0, 0.06)}            # Frente a la llave { no debe ir nada
                                                    # la llave de cierre } debe ir sola
```

La sentencia else debe ir siempre entre llaves "{ }", ver el siguiente ejemplo.

```
if (condition) {
  one or more lines
} else {
  one or more lines
}

if (condition) {
  one or more lines
}
else {
  one or more lines
}

if (condition)
  one line
else
  one line
```

1.5.5. Asignación

Para realizar asignaciones se recomienda usar el símbolo `<-`, el símbolo de igualdad `=` no se recomienda usarlo para asignaciones.

```
x <- 5 # Correcto
x = 5 # No recomendado
```

1.5.6. Punto y coma

No se recomienda colocar varias instrucciones separadas por `;` en la misma línea, aunque funciona dificulta la revisión del código.

```
n <- 100; y <- rnorm(n, mean=5); hist(y) # No se recomienda

n <- 100                                # Correcto
y <- rnorm(n, mean=5)
hist(y)
```


Capítulo 2

Gráficos para una variable

En este capítulo se presentan funciones para la creación de gráficos con una sola variable.

2.1. Función `stem`

Nos permite obtener el gráfico llamado de tallo y hoja debido a su apariencia. Este gráfico fue propuesto por Tukey (1977) y a pesar de no ser un gráfico para presentación definitiva se utiliza a la vez que el analista recoge la información para ver rápidamente la distribución de los datos.

¿Qué muestra este gráfico?

1. El centro de la distribución.
2. La forma general de la distribución:
 - Simétrica: Si las porciones a cada lado del centro son imágenes espejos de las otras.
 - Sesiada a la izquierda: Si la cola izquierda (los valores menores) es mucho más larga que los de la derecha (los valores mayores).
 - Sesiada a la derecha: Opuesto a la sesiada a la izquierda.
3. Desviaciones marcadas de la forma global de la distribución.
 - Outliers: Observaciones individuales que caen muy por fuera del patrón general de los datos.
 - Gaps: Huecos en la distribución

Ventajas del gráfico:

1. Muy fácil de realizar y puede hacerse a mano.
2. Fácil de entender.

Desventajas del gráfico:

1. El gráfico es tosco y no sirve para presentaciones definitivas.
2. Funciona cuando el número de observaciones no es muy grande.
3. No permite comparar claramente diferentes poblaciones

Ejemplo

Como ilustración vamos a crear el gráfico de tallo y hoja para la variable altura (cm) de un grupo de estudiantes de la universidad. Primero se leerán los datos disponibles en github y luego se usará la función `stem` para obtener el gráfico. A continuación el código usado.

```
url <- 'https://raw.githubusercontent.com/fhernanb/datos/master/medidas_cuerpo'
datos <- read.table(file=url, header=T)
stem(datos$altura)
```

```
##
## The decimal point is 1 digit(s) to the right of the |
##
## 14 | 7
## 15 | 3
## 15 | 679
## 16 | 0001
## 16 | 68888
## 17 | 001334
## 17 | 5678899
## 18 | 000033
## 18 | 88
## 19 | 1
```

De este gráfico sencillo se puede ver que la variable presenta una mayor frecuencia para alturas ente 170 y 179 cm y que no tiene una distribución simétrica.

2.2. Función boxplot

La función `boxplot` sirve para crear un diagrama de cajas y bigote para una variable cuantitativa. La estructura de la función `boxplot` con los argumentos más comunes de uso se muestran a continuación.

```
boxplot(x, formula, data, subset, na.action,
        range, width, varwidth, notch, names,
        plot, col, log, horizontal, add, ...)
```

Los argumentos de la función `boxplot` son:

- **x**: vector numérico con los datos para crear el boxplot.
- **formula**: fórmula con la estructura $x \sim g$ para indicar que las observaciones en el vector **x** van a ser agrupadas de acuerdo a los niveles del factor **g**.
- **data**: marco de datos con las variables.
- **subset**: un vector opcional para especificar un subconjunto de observaciones a ser usadas en el proceso de ajuste.
- **na.action**: una función la cual indica lo que debería pasar cuando los datos contienen “NA’s”.
- **range**: valor numérico que indica la extensión de los bigotes. Si es positivo, los bigotes se extenderán hasta el punto más extremo de tal manera que el bigote no supere **range** veces el rango intercuatílico (*IQ*). Un valor de cero hace que los bigotes se extiendan hasta los datos extremos.
- **width**: un vector con los anchos relativos de las cajas.
- **varwidth**: Si es **TRUE**, las cajas son dibujadas con anchos proporcionales a las raíces cuadradas del número de observaciones en los grupos.
- **notch**: si es **TRUE**, una cuña es dibujada a cada lado de las cajas. Cuando las cuñas de dos gráficos de caja no se traslapan, entonces las medianas son significativamente diferentes a un nivel del 5%.
- **names**: un con las etiquetas a ser impresas debajo de cada boxplot.
- **plot**: si es **TRUE** (por defecto) entonces se produce el gráfico, de lo contrario, se producen los resúmenes de los boxplots.
- **col**: vector con los colores a usar en el cuerpo de las cajas.
- **log**: para indicar si las coordenadas **x** o **y** o serán graficadas en escala logarítmica.
- **...**: otros parámetros gráficos que pueden ser pasados como argumentos para el boxplot.

Ejemplo

Como ilustración vamos a crear tres boxplot para la variable altura (cm) de un grupo de estudiantes de la universidad, el primer boxplot será sobre la variable altura, el segundo será un boxplot para altura diferenciando por sexo y el tercer boxplot será igual que el primero pero modificando los nombres a imprimir en el eje horizontal. Primero se leerán los datos disponibles en github y luego se usará la función `boxplot` para obtener ambos gráfico. A continuación el código usado.

```
url <- 'https://raw.githubusercontent.com/fhernanb/datos/master/medidas_cuerpo'
datos <- read.table(file=url, header=T)

par(mfrow=c(1, 3))
boxplot(x=datos$altura, ylab='Altura (cm)')

boxplot(altura ~ sexo, data=datos,
        xlab='Sexo', ylab='Altura (cm)')

boxplot(altura ~ sexo, data=datos, horizontal=TRUE,
        ylab='Género', xlab='Altura (cm)',
        names=c('Masculino', 'Femenino'))
```

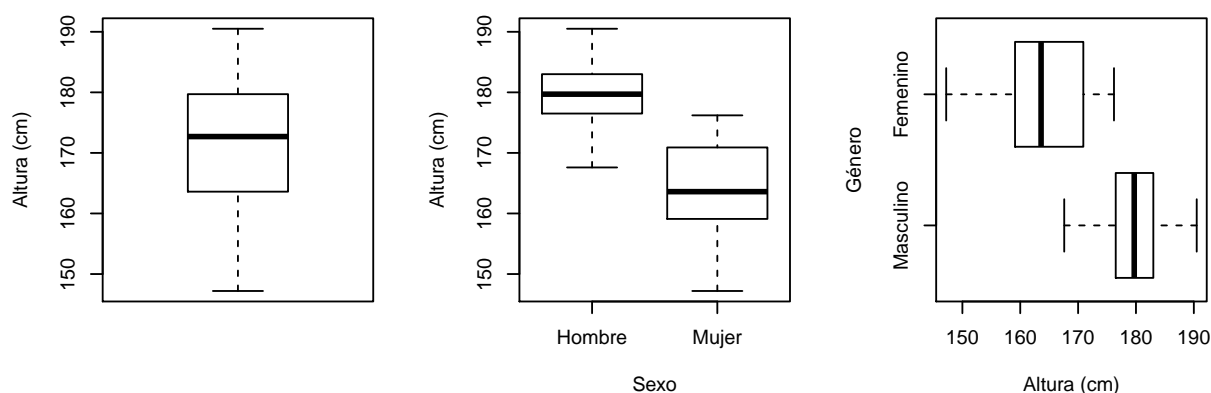


Figura 2.1: Boxplot para la variable altura.

En la Figura 2.1 se presentan los boxplots obtenidos con las instrucciones anteriores. El segundo y tercer boxplot son el mismo, lo único que se modificó fueron los nombres o etiquetas a colocar debajo de cada boxplot por medio del argumento `names` y la orientación.

2.3. Función hist

La función `hist` sirve para crear el histograma para una variable cuantitativa. Como argumentos esta función recibe un vector con los datos y opcionalmente puede ingresarse como argumento el número de intervalos a ser graficados o en su defecto el número de clases se determina con la fórmula de Sturges.

Nota: los programas de computador usualmente construyen los histogramas automáticamente, sin embargo, un buen programa debe permitirnos elegir el número de intervalos del histograma. Si usted posee un programa que no le permite hacer cambios, cambie de programa.

La estructura de la función `hist` con los argumentos más comunes de uso se muestran a continuación.

Bibliografía

Wickham, H. (2015). *R Packages*. O'Reilly Media, Inc.

Índice alfabético

arreglo, 11

boxplot, 18

guía de estilo, 12

hist, 19

histograma, 19

lista, 12

marco de datos, 11

matrices, 10

objetos, 9

vectores, 9