Entorno de programación RStudio: incorporar datos

Román Salmerón Gómez

Abril de 2017

El objetivo del presente documento es el de crear un breve informe con **RStudio** mediante el cual se muestre al alumno su uso básico.

Tras abrir **RStudio** vamos a crear una carpeta y establecerla como el directorio de trabajo (*Session -> Set Working Directory -> Choose Directoy*), abrimos un nuevo proyecto de **R Markdown** (muy interesante consultar su ayuda: *Help -> Markdown Quick Reference*). Seleccionaremos que genere un *word*. De esta forma toda la información que se genere se almacenará en la carpeta establecida como directorio de trabajo.

Para conseguir ayuda sobre cualquier comando escribir en consola *help()* donde entre paréntesis hay que escribir dicho comando.

# Simulación de datos

En primer lugar vamos a generar un conjunto aleatorio de datos procedentes de una normal y vamos a calcular su media y varianza:

obs = 300  
 media = 5  
 desv.tip = 5  
 muestra.normal = rnorm(obs, media, desv.tip) # genera datos según una normal con media y varianzas dadas  
 mean(muestra.normal) # calcula la media de un conjunto de datos

## [1] 5.307081

var(muestra.normal) # calcula la cuasivarianza de un conjunto de datos

## [1] 26.71755

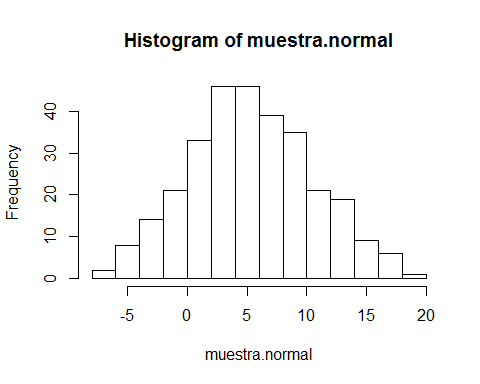
(obs/(obs-1))\*var(muestra.normal) # calculamos varianza: puesto que el número de observaciones es alto...

## [1] 26.8069

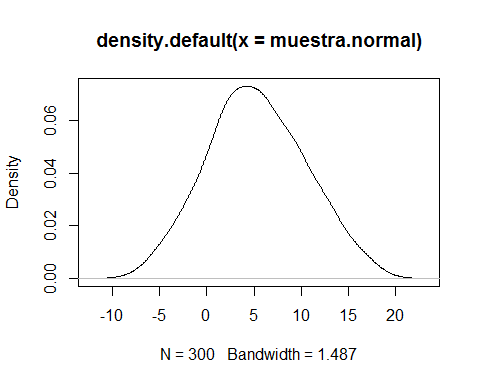
sd(muestra.normal) # help(sd)

## [1] 5.168902

hist(muestra.normal) # representa el histograma, más info help(hist)



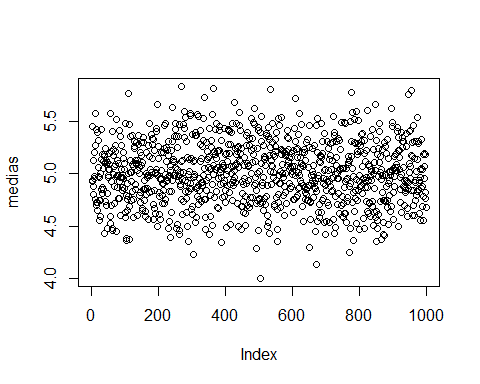
plot(density(muestra.normal)) # representa la densidad



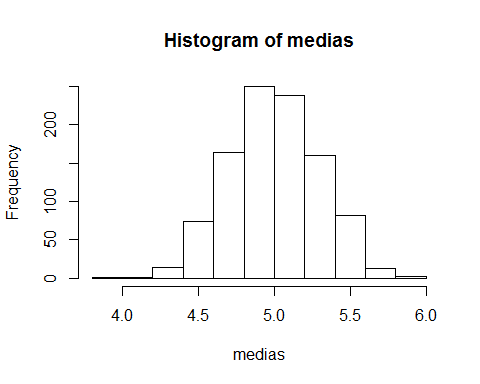
# como se ha puesto de manifiesto la almohadilla sirve para introducir comentarios

A continuación repito el proceso anterior *num.muestras* veces, por lo que estamos generando *num.muestras* muestras aleatorias independientes procedentes de una población normal con una media y varianza concretos (en este caso, con media *media* y varianza *desv.tip^2*). Para cada una de las muestras calculo su media y lo almaceno en *medias*:

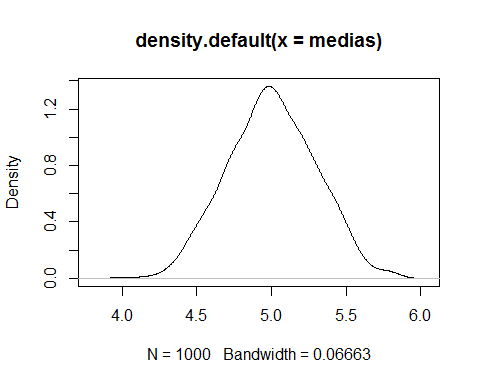
num.muestras = 1000  
 medias = array(,num.muestras) # creo un vector vacío de dimensiones num.muestrasx1  
 for (i in 1:num.muestras) {  
 muestra.normal.it = rnorm(obs, media, desv.tip)   
 medias[i] = mean(muestra.normal.it)   
 }  
 # medias # muestra en pantalla las medias calculadas (quitar el primer comentario)  
 plot(medias) # representa los valores anteriores, help(plot) para más detalles



hist(medias)



plot(density(medias))



# Incorporación de un conjunto de datos existente

## Mediante consola

Si bien ante conjuntos de datos con tamaño grande puede resultar poco práctico, es posible introducir los datos directamente en consola:

notas1 = c(5.75, 6.25, 4.5, 6, 7.75, 5, 5.75, 2.5, 6, 4.5, 4.5, 1, 1.5, 8.5, 7.25, 6, 4, 6.75, 3.25, 1, 7, 1, 6.75)  
 notas1

## [1] 5.75 6.25 4.50 6.00 7.75 5.00 5.75 2.50 6.00 4.50 4.50 1.00 1.50 8.50  
## [15] 7.25 6.00 4.00 6.75 3.25 1.00 7.00 1.00 6.75

summary(notas1)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.625 5.750 4.891 6.500 8.500

notas2 = c(1, 1.25, 6.25, 6.75, 6.25, 4, 8.75, 3.5, 8.5, 7.25, 2.25, 2, 5.75, 1.5, 7.25, 6.25, 2)  
 notas2

## [1] 1.00 1.25 6.25 6.75 6.25 4.00 8.75 3.50 8.50 7.25 2.25 2.00 5.75 1.50  
## [15] 7.25 6.25 2.00

summary(notas2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 5.750 4.735 6.750 8.750

## Mediante comando *read.table*

Sin embargo, la opción habitual será incorporar la información a tratar a partir de una hoja excel (o similar). Para hacer esta incorporación más estable, guardaremos la información como un archivo *.csv* separado por puntos y comas o como un archivo *.txt* separado por tabulaciones. En este sentido hay que tener en cuenta que el delimitador decimal es el punto y no la coma. Los datos se pueden incorporar a **RStudio** siguiendo el camino *Tools -> Import Dataset* o mediante el siguiente comando (se supone que los datos se encuentran en la carpeta que hemos establecido como directorio de trabajo, que es donde se encuentra este archivo *.Rmd*):

datos = read.table("sesion1.csv", header=T, sep=";")   
 # header=T indica que en la primera línea están el nombre de las variables  
 # sep=";" indica que las variables están separadas por ;   
 # help(read.table) para ver más opciones  
 names(datos) # vemos el nombre de cada una de las variables

## [1] "DESCRIPTOR" "M3" "ppi" "CPI" "GDP"   
## [6] "M1NSA" "DDNSA"

# M3 # no la reconoce (quitar el primer comentario)  
 attach(datos) # incorporamos los datos a R para referenciar cada variable por el nombre anterior  
 # M3 # ya sí la reconoce (quitar el primer comentario)

Una vez tengamos las variables incorporadas podremos trabajar con ellas, por ejemplo, obteniendo un resumen estadístico de ellas:

summary(M3)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 615.7 1036.0 2055.0 2731.0 4205.0 8149.0 3

Otra forma de usar la orden anterior es incorporar datos desde la web:

datosWeb <- read.table("http://www.ugr.es/~romansg/material/WebEco/datos\_RM.csv", header=T, sep=";")  
 head(datosWeb) # un vistazo para ver cómo son los datos

## Resultado Llull Chacho Rudy Mirotic PR PorcRT Q1 T1IR L  
## 1 1 1 1 1 1 61 53.84615 25 21 0  
## 2 1 0 1 0 1 79 48.43750 29 20 0  
## 3 1 0 0 0 1 53 64.00000 26 6 1  
## 4 1 1 0 1 1 63 43.54839 16 14 0  
## 5 1 0 1 0 1 72 58.33333 26 18 1  
## 6 1 1 0 0 1 58 59.45946 18 8 1

summary(datosWeb)

## Resultado Llull Chacho Rudy   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.0000 Median :0.0000 Median :0.0000   
## Mean :0.8481 Mean :0.5316 Mean :0.4557 Mean :0.4595   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :5   
## Mirotic PR PorcRT Q1   
## Min. :0.000 Min. : 53.00 Min. :32.14 Min. :10.00   
## 1st Qu.:0.000 1st Qu.: 66.00 1st Qu.:47.59 1st Qu.:18.00   
## Median :1.000 Median : 73.00 Median :50.82 Median :20.00   
## Mean :0.519 Mean : 73.78 Mean :51.22 Mean :20.87   
## 3rd Qu.:1.000 3rd Qu.: 80.00 3rd Qu.:54.67 3rd Qu.:25.00   
## Max. :1.000 Max. :110.00 Max. :64.00 Max. :34.00   
##   
## T1IR L   
## Min. : 0.00 Min. :0.0000   
## 1st Qu.:11.00 1st Qu.:0.0000   
## Median :16.00 Median :0.0000   
## Mean :16.13 Mean :0.4684   
## 3rd Qu.:21.00 3rd Qu.:1.0000   
## Max. :34.00 Max. :1.0000   
##

# Bases de datos existentes en R

En *R* existen diversos conjuntos de datos que pueden ser usados sin más que escribir su nombre. Por ejemplo, los datos de la base *cars* (para más información *help(cars)*) se obtendrían como sigue:

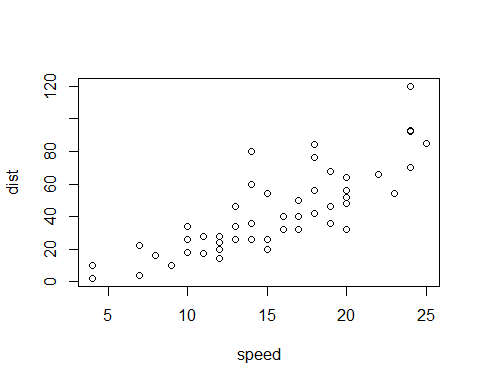
# cars # quitar primera almohadilla si se quieren ver los datos  
 names(cars) # vemos los nombres de las variables que hay en esta base de datos

## [1] "speed" "dist"

attach(cars)

Entonces, ya se puede trabajar con ellos:

plot(speed,dist) # diagrama de dispersión de speed vs dist, ver help(plot) para más opciones



cov(cars) # matriz varianzas-covarianza entre speed y dist

## speed dist  
## speed 27.95918 109.9469  
## dist 109.94694 664.0608

cor(cars) # matriz correlaciones entre speed y dist

## speed dist  
## speed 1.0000000 0.8068949  
## dist 0.8068949 1.0000000

# Crear funciones en R

Puesto que en el comando *summary* no aparecen todas las características estadísticas que nos pueden interesar, creamos nuestra propia función:

# install.packages("moments") # ejecutarlo sólo una vez para instalar el paquete necesario para calcular la asimetría y curtosis  
  
 Resumen <- function(datos)  
 {  
 minimo <- min(datos)  
 maximo <- max(datos)  
 mediana <- median(datos)  
 cuantil1 <- quantile(datos, probs=0.25, name=FALSE)  
 cuantil2 <- quantile(datos, probs=0.5, name=FALSE)  
 cuantil3 <- quantile(datos, probs=0.75, name=FALSE)  
 media <- mean(datos)  
 varianza <- var(datos) # ¡ojo! cuasivarianza  
 desviaciontipica <- sqrt(var(datos)) # otra opción es usar sd(datos)  
 library(moments)  
 asimetria = skewness(datos)  
 curtosis = kurtosis(datos)  
 resultado <- c(minimo, maximo, mediana, cuantil1, cuantil2, cuantil3, media, varianza, desviaciontipica, asimetria, curtosis)  
 names(resultado) <- c("Mínimo", "Máximo", "Mediana", "Primer Cuartil", "Segundo Cuartil", "Tercer Cuartil", "Media", "Varianza", "Desviación Típica", "Coeficiente de asimetría", "Coeficiente de curtosis")  
 return(resultado)  
 }

De manera que:

Resumen(speed)

## Mínimo Máximo Mediana   
## 4.0000000 25.0000000 15.0000000   
## Primer Cuartil Segundo Cuartil Tercer Cuartil   
## 12.0000000 15.0000000 19.0000000   
## Media Varianza Desviación Típica   
## 15.4000000 27.9591837 5.2876444   
## Coeficiente de asimetría Coeficiente de curtosis   
## -0.1139548 2.4228526

Resumen(dist)

## Mínimo Máximo Mediana   
## 2.0000000 120.0000000 36.0000000   
## Primer Cuartil Segundo Cuartil Tercer Cuartil   
## 26.0000000 36.0000000 56.0000000   
## Media Varianza Desviación Típica   
## 42.9800000 664.0608163 25.7693775   
## Coeficiente de asimetría Coeficiente de curtosis   
## 0.7824835 3.2480187

## Ejercicios propuestos

* Incorpora los datos de *longley* a *R*. ¿Cuáles son las variables disponibles en esta base de datos?
* Calcula la media y varianza de la variable correspondiente al pIB.
* Usa la función *Resumen* para clacular las principales características descriptivas de la variable correspondiente al desempleo.