LAB 0. ML CSIC

DataLab CSIC

18/1/2023

## Introducción

Este es el primer LAB. Sirve para recordar conceptos básicos en R. Hay decenas de buenas introducciones. Una muy buena es <https://datanalytics.com/libro_r/>. Una buena intro a Rmarkdown (el curso completo está muy bien) <https://www.uv.es/pjperez/curso_R/tt_8_Rmarkdown_v0.html> Recuerda que R y otra info básica se puede descargar de <http://www.r-project.org/> y de <http://cran.es.r-project.org/>. Se recomienda emplear RStudio. Busca download RStudio en Inet.

Primero algunas operaciones básicas con vectores y recuerda usar help cuando lo necesites. Por ejemplo, recuerda que es runif haciendo help(runif) y que R sirve de potente calculadora.

x<-runif(10,6,9)  
x

## [1] 6.853289 7.725570 7.248047 7.852004 8.098404 7.526406 7.209240 6.992238  
## [9] 6.676118 7.338570

2\*x

## [1] 13.70658 15.45114 14.49609 15.70401 16.19681 15.05281 14.41848 13.98448  
## [9] 13.35224 14.67714

log(x)

## [1] 1.924729 2.044536 1.980732 2.060769 2.091667 2.018418 1.975364 1.944801  
## [9] 1.898537 1.993144

y<-log(x)  
x+y

## [1] 8.778017 9.770105 9.228779 9.912772 10.190071 9.544824 9.184604  
## [8] 8.937038 8.574655 9.331714

x\*y

## [1] 13.19072 15.79520 14.35644 16.18116 16.93916 15.19143 14.24087 13.59851  
## [9] 12.67485 14.62683

z<-x\*y  
z

## [1] 13.19072 15.79520 14.35644 16.18116 16.93916 15.19143 14.24087 13.59851  
## [9] 12.67485 14.62683

Aquí mostramos alguna gráfica (histograma) y recordamos que las funciones tienen parámetros que tenemos que acomodar. Busca información sobre hist.

x<-runif(100,6,9)  
 hist (x)   
x<-runif(1000,6,9)   
hist(x)   
hist(x,breaks=7)

## Intervalos de probabilidad y otras cuestiones con la normal

Usaremos bastante el concepto de intervalos de probabilidad. Un intervalo simétrico de probabilidad =.72 para la normal N(0,2) (ojo, desviación típica 2)

lb<-qnorm(0.14,0,2)  
ub<-qnorm(0.86,0,2)  
lb  
ub

Ahora calculamos la probabilidad de tener una observación mayor que 6 sigma (en valor absoluto) en una N(3,2). Después trabajamos con la densidad normal N(0,1) mostrando su densidad y una concatenación de funciones

pnorm(-9,3,2)+(1-pnorm(15,3,2))  
x<-runif(100,-2,2)  
y<-dnorm(sort(x))  
plot(y)

## Recordamos algunas cosas más

Breve recordatorio de varias funciones y resúmenes numéricos y operaciones que emplearemos frecuentemente

x<-runif(100,0,1)  
 boxplot(x)   
summary(x)   
mean(x)   
var(x)   
sqrt(var(x))   
sd(x)   
median (x)  
 quantile(x,0.25)   
quantile (x,0.75)  
cvx<-sd(x)/mean(x)   
 cvx  
y<-(x-mean(x))/sd(x)   
 mean(y)   
 sd(y)

Aquí recordamos conceptos básicos sobre localización y dispersión y como presentar múltiples gráficos (muy básicos, recuerda la potencia espectacular de R en lo que respecta a gráficos, como iremos viendo en futuros labs).

help(rnorm)  
x<-rnorm(1000,0,1)   
 y<-rnorm(1000,0,10)   
 z<-rnorm(1000,100,1)  
par(mfrow=c(1,3))   
hist(x)   
hist(y)  
 hist(z)

El mundo no es sólo normal. Aquí practicamos algunas nociones sobre la exponencial y el efecto de transformaciones sobre los datos

x<-rexp(1000,3)   
mean(x)   
sd(x)  
mean(x)/sd(x)  
boxplot(x)   
boxplot(log(x))

Algo de análisis exploratorio bivariante con una referencia al modelo lineal. Más cosas más tarde

x<-runif(50,4,12)   
 y<-3\*x+5+rnorm(50,0,3)   
plot(x,y)   
cor(x,y)

Aquí un poco de notación sobre modelos, en particular el modelo lineal,

cov(x,y) #hola  
help(lm)   
 modelo<-lm(y ~ x)   
 summary(modelo)   
plot(modelo)

Un poco sobre detección de outliers

x<-rnorm(50,3,1)  
 x[51]<-100   
 x   
 hist(x)   
boxplot(x)

Usaremos bastantes paquetes de R. Aquí descargamos gtools como ejemplo. Nos pedirá el repositorio del que bajarlo (p.ej Spain, Madrid). Una vez instalado se carga con library. Search nos da los paquetes cargados, library los instalados, data los conjuntos de datos instalados. A través de Tools>Install Packages de RStudio resulta más fácil.

library()  
help(distributions)  
install.packages("gtools")  
library()  
library(gtools)  
search()  
library()  
data()

Finalmente, algo superbásico sobre carga de ficheros. Ojo al directorio. Desde RStudio es más sencillo aún (Session, Set working directory, To source file location). Aquí un poco de inferencia clásica con el modelo exponencial. Tras cargar los datos y hacer un poco de análisis exploratorio, cargamos la librería MASS del libro de Venables y Ripley y obtenemos el MLE y el grado de ajuste

datos<-scan("C:/Users/David/Desktop/clases/cursoMCBI/labs/lab0/data1.txt")  
datos<-scan("data0.txt")  
hist(datos)  
summary(datos)  
cv<-sd(datos)/mean(datos)  
help(library)  
library(MASS)  
fitdistr(datos,"exponential")  
lam<-1/mean(datos)  
lam

Fin de esta parte de intro a R. En los siguientes labs empleamos muchos otros paquetes, funciones. Recuerda usar el help

## Piccolo divertimento finales (ma non troppo)

Algunas estructuras más de datos. Cargamos primero ggplot2 para gráficos

install.packages("ggplot2")  
library(ggplot2)  
pages = c(0.1, 0.9, 0.23, 0.77)  
Topico = c("UL", "SL", "UL", "SL")  
Libro = c("ESL", "ESL", "Bishop", "Bishop")  
data = data.frame(pages, Topico, Libro)  
p = ggplot(data, aes(x=Libro, y=pages, fill=Topico)) +   
 geom\_bar(stat = "identity", position = "dodge")  
p = p + xlab("Libro")+ ylab("Proporción de páginas")  
p

## Observación final

Juega con echo = FALSE,TRUE y eval= FALSE,TRUE para visualizar o no, evaluar o no los resultados.