VNIVERSITAT DE VALÈNCIA Facultat d'Economia

POSTGRAU

MÁSTER EN CIENCIAS ACTUARIALES Y FINANCIERAS

Estadística Avanzada para Actuarios

Introducción a R

R :un lenguaje de programación y un entorno para análisis estadístico .

Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda.

R actualmente es el resultado de un esfuerzo de colaboración de personas del todo el mundo.

Ya que tiene la posibilidad de modificación directa del código fuente.

Por otra parte, R es un proyecto GNU similar a S, desarrollado éste por los Laboratorios Bell. Las diferencias entre R y S son importantes, pero la mayoría del código escrito para S corre bajo R sin modificaciones.

La pagina principal del proyecto es

http://www.r-project.org.

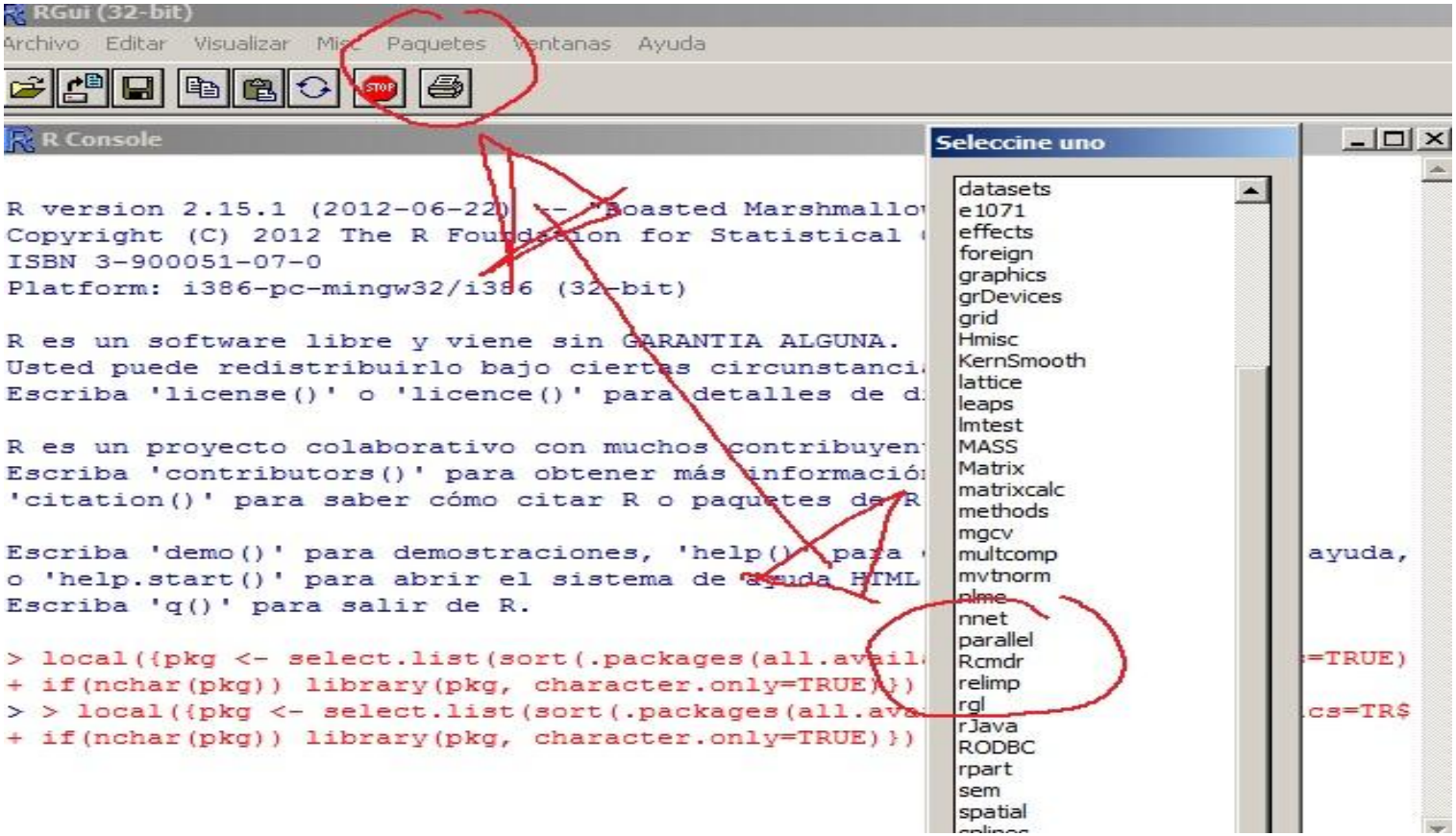
En España http://cran.es.r-project.org/

Característica principal de R

Cualquier expresión evaluada por R tiene como resultado un objeto.

Cada objeto pertenece a una clase, de forma que las funciones pueden tener comportamientos diferentes en función de la clase a la que pertenece su objeto

Instalar R- Commander



Luego se procede con la ejecución, siguiendo las instrucciones. Para la instalación de Rcmdr, se arranca R desde Inicio→Todos los programas→ R. A con□nuación, Paquetes→Instalar Paquete(s) y elegido el mirror desde el cual se quiere instalar el paquete, por ejemplo Spain (Madrid), se selecciona Rcmdr.

Si se cierra Rcmdr (sin cerrar R), para volver a cargarlo se debe ejecutar la instrucción Commander().

Los datos : Análisis Exploratorio de Datos

En una primera instancia , los datos se supondrán obtenidos sobre un conjunto de n individuos físicos, de los que se conoce una serie de k caracteres u observaciones de igual o distinta naturaleza

Los datos obtenidos se organizarán en una matriz $n \times k$, donde cada fila representa a un individuo o registro y las columnas a las carácterísticas observadas. Las columnas tendrán naturaleza homogénea, pudiendo tratarse de caracteres nominales, dicotómicos o politómicos, presencias–ausencias, ordenaciones, conteos, escalas de intervalo, razones,…ratios , densidades

Si se consideran los individuos identificados por los términos $I_1, I_2, \ldots, I_n$ y los caracteres por $C_1, C_2, \ldots, C_k$, la casilla x_{ij} representa el comportamiento del individuo I_i respecto al carácter C_j .

Los huecos que queden en la matriz se referirán como valores omitidos o, más comunmente, como valores missing. En R estos valores se representan con NA (Not Available). En función del tipo de análisis que se esté realizando, el procedimiento desestimará sólo el dato o todo el registro completo.

Tipos de datos/información y su " aporte" informativo

Tipos de medidas y gráficos habituales según escala

Escala de Medida	Medidas centrales	Medidas de dispersión	Representaciones gráficas
Atributo	Moda Porcentajes		Diagrama de sectores
Ordenación	Mediana Percentiles	Recorrido Intercuartílico	Diagrama de barras
Recuento	Media	Desviación típica	Diagramas de barras
Intervalo	Media	Desviación típica	Histograma
Razón	Media geométrica	Coeficiente de variación	Histograma Diagrama de dispersión Diagrama de cajas

R como calculadora , funciones

`> 2+2`	Escribimos 2+2 y damos enter
`[1] 4`	Sale el resultado. [1] indica que es el primer (y único resultado) de nuestra orden.
`>`	Al terminar el comando, el sistema vuelve a presentar su indicador.

`> 2*5` `[1] 10`	Multiplicación de dos números
`> 5/2` `[1] 2.5`	División real de dos números
`> 5%/%2` `[1] 2`	División entera: se devuelve la parte entera solamente
`> 5%%2` `[1] 1`	Módulo: resto de dividir un número por otro
`> 11%%3` `[1] 2`	Otro ejemplo de la operación módulo
`> 5^2` `[1] 25`	Potenciación
`> 5^2.3` `[1] 40.51641`	Potenciación, exponente real
`> exp(1)` `[1] 2.718282`	El número *e*

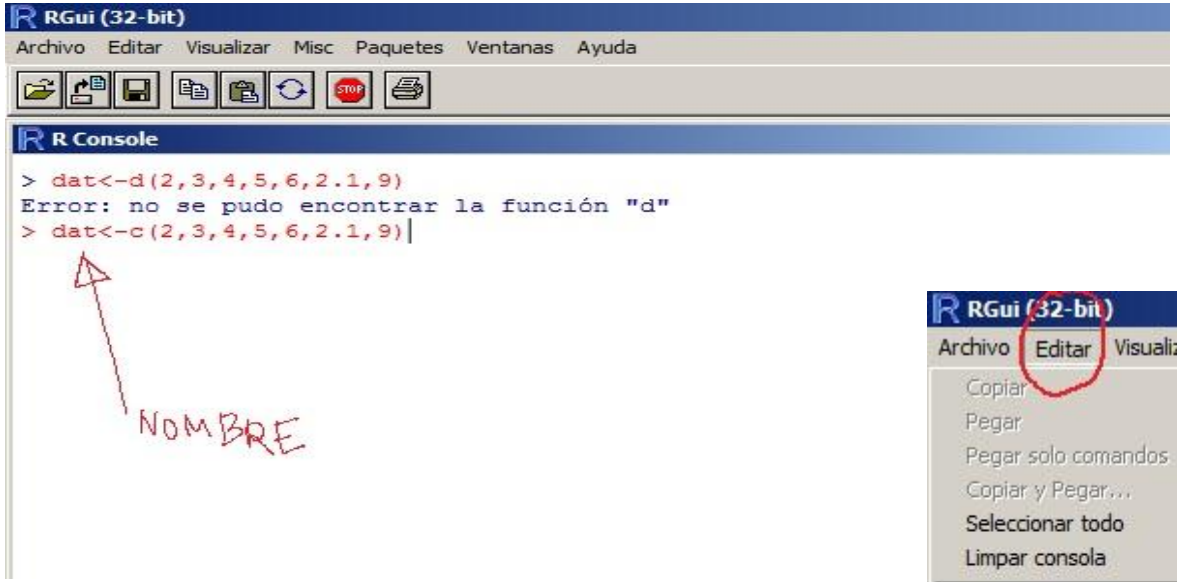
R como calculadora , funciones

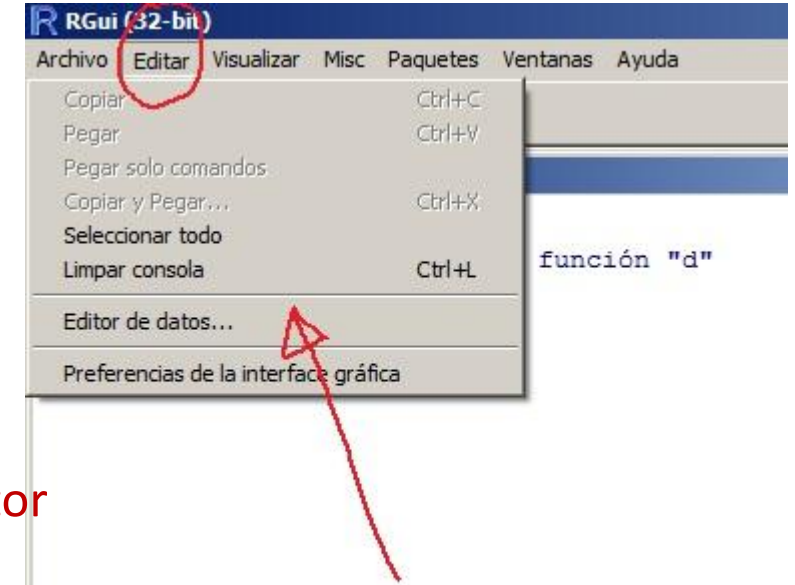
`> exp(3)` `[1] 20.08554`	El número e^3
`> sqrt(2)` `[1] 1.414214`	La raíz cuadrada de 2
`> log(3)` `[1] 1.098612`	El logaritmo neperiano de 3
`> log(3,10)` `[1] 0.4771213`	El logaritmo de 3 en base 10
`> abs(-3.4)` `[1] 3.4`	Valor absoluto de un número
`> pi` `[1] 3.141593`	El número pi

Más información con las funciones →

```
help("Math")
help("Arithmetic")
help("Trig")
help("Log")
help("Special")
```

Introducir datos directamente una sola variable

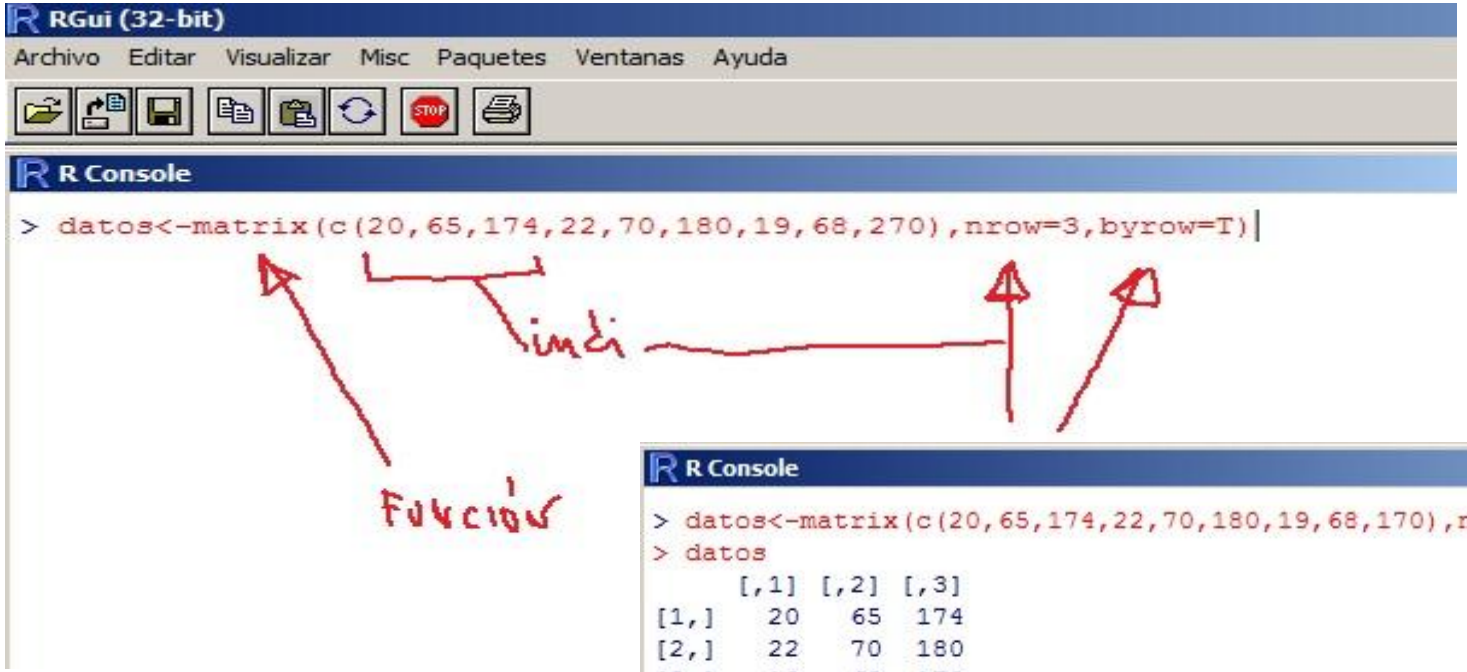




Los datos se pueden modificar con el editor

Comprobar cambios con la función **mean(nombre)**

Más de una variable , matrices



Función dimnames →

```
R Console
> datos<-matrix(c(20,65,174,22,70,180,19,68,170),nrow=3,byrow=T)
> datos
     [,1] [,2] [,3]
[1,]   20   65  174
[2,]   22   70  180
[3,]   19   68  170
> dimnames(datos)<-list(c("antón","Juan","Alex"), # hola estoy aqui
+  c("edad","peso","altura"))
> datos
      edad peso altura
antón   20   65    174
Juan    22   70    180
Alex    19   68    170
>
```

data	datos que forman la matriz
nrow	número de filas de la matriz
ncol	número de columnas de la matriz
byrow	Los datos se colocan por filas o por columnas según se van leyendo. Por defecto se colocan por columnas.

Función SEQ , genera números

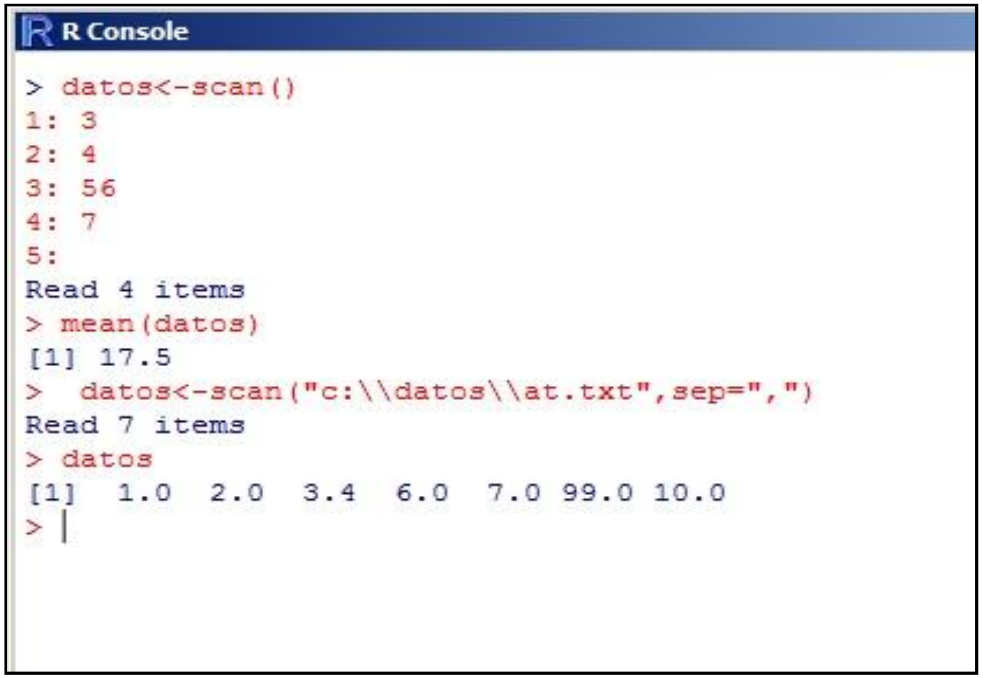
```
R Console
> seq(1,6)
[1] 1 2 3 4 5 6
> seq(1,6,by=0.5)
 [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
> seq(1,6,length=10)
 [1] 1.000000 1.555556 2.111111 2.666667 3.222222 3.777778 4.333333 4.888889
 [9] 5.444444 6.000000
>
```

Inserta datos desde la propia aplicación
Y los guarda en “datos”

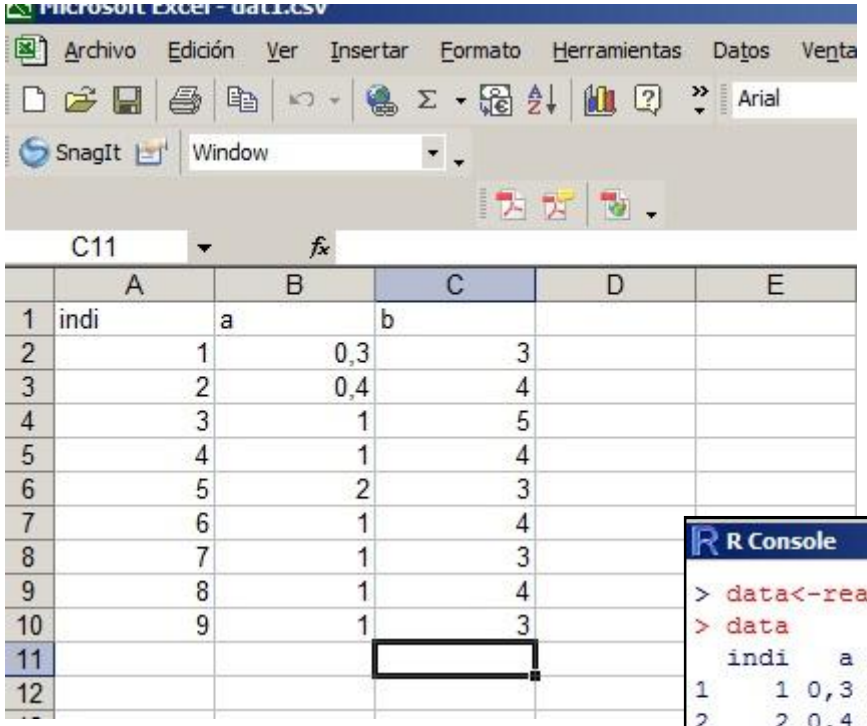
```
RGui (32-bit)
Archivo  Editar  Visualizar  Misc  Paquetes  Ventanas  Ayuda
R Console
> datos<-scan()
1: 3
2: 4
3: 56
4: 7
5:
Read 4 items
> mean(datos)
[1] 17.5
>
```

Archivo de texto creado

Función y argumentos para utilizarlo



Datos desde excel o archivo csv



	A	B	C	D	E
1	indi	a	b		
2	1	0,3	3		
3	2	0,4	4		
4	3	1	5		
5	4	1	4		
6	5	2	3		
7	6	1	4		
8	7	1	3		
9	8	1	4		
10	9	1	3		
11					
12					

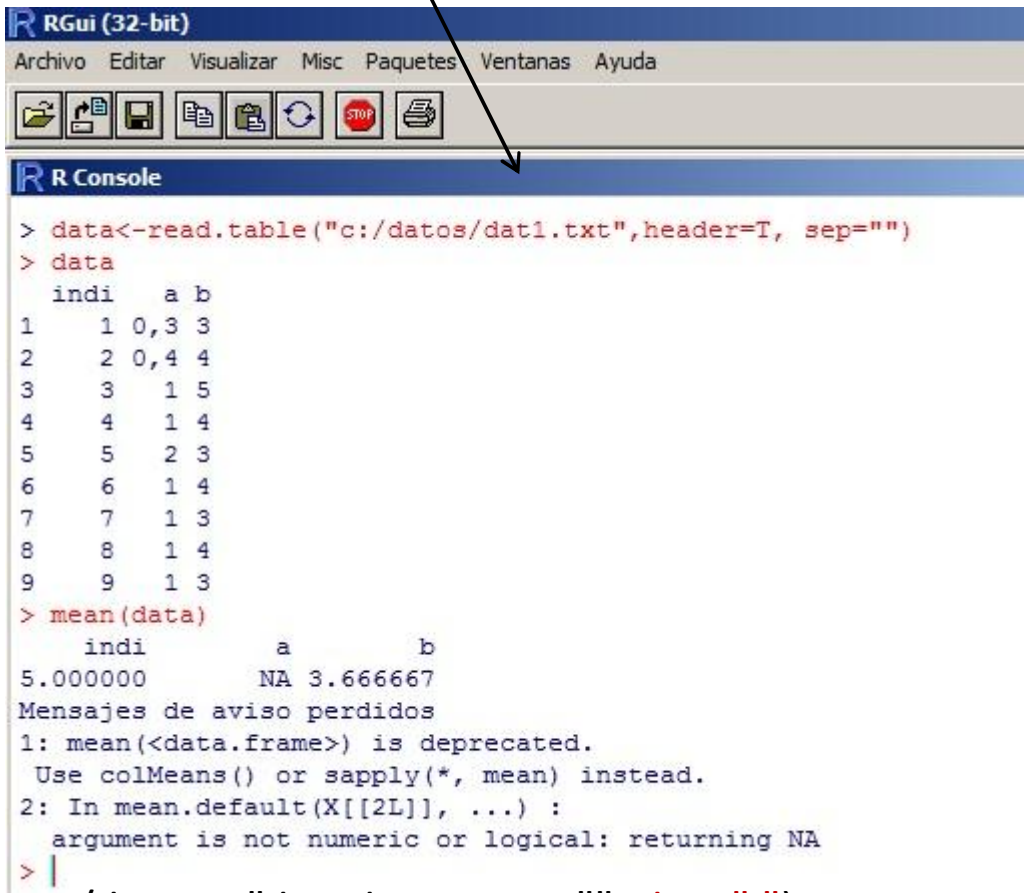
← Hoja de excel guadar como csv

Función read.csv
header=T primera fila nombres
sep=";" el separador es ;

R Console

```
> data<-read.csv("c:/datos/dat1.csv",header=T, sep=";")
> data
  indi   a b
1    1 0,3 3
2    2 0,4 4
3    3   1 5
4    4   1 4
5    5   2 3
6    6   1 4
7    7   1 3
8    8   1 4
9    9   1 3
> 
```

Insertar datos con txt



Si calculamos las media observamos que la de la variable "a" no puede calcularse, por el problema de la coma

solución : data<-read.table("c:/datos/dat1.txt",header=T, sep="", dec=",")

Datos en txt, con atributos

dat2.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
indi    dia     número
1       lunes   3
2       lunes   4
3       martes  5
4       miercoles       4
5       jueves  3
6       jueves  4
7       lunes   3
8       lunes   4
9       viernes 3
```

Utilización de variable de una base de datos $

Creación de tabla de contingencia

```
> data2<-read.table("c:/datos/dat2.txt",header=T, sep="", dec=",")
> data2
  indi       dia número
1    1     lunes      3
2    2     lunes      4
3    3    martes      5
4    4 miercoles      4
5    5    jueves      3
6    6    jueves      4
7    7     lunes      3
8    8     lunes      4
9    9   viernes      3
> mean(data2$número)
[1] 3.666667
> table(data2$dia,data2$número)

           3 4 5
  jueves   1 1 0
  lunes    2 2 0
  martes   0 0 1
  miercoles 0 1 0
  viernes  1 0 0
>
```

Funciones estadísticas básicas

cor	Correlación (admite uno o dos argumentos)
cumsum	Suma cumulativa de un vector
mean	Media aritmética
median	El percentil 0.5: la mediana
min	El mínimo de una serie de números
max	El máximo de una serie de números
prod	El producto de los elementos de un vector
quantile	Los percentiles de una distribución
range	Mínimo y máximo de un vector
sample	Muestreo aleatorio (y permutaciones)
sum	Suma aritmética
var	Varianza y covarianza
summary	Resumen de estadísticas

Ejemplo , correlación a,b
cuarties de a

R Console

```
> data<-read.table("c:/datos/dat1.txt",header=T, sep="", dec=",")
> data
  indi   a b
1    1 0.3 3
2    2 0.4 4
3    3 1.0 5
4    4 1.0 4
5    5 2.0 3
6    6 1.0 4
7    7 1.0 3
8    8 1.0 4
9    9 1.0 3
> cor(data$a,data$b)
[1] -0.147442
> quantile(data)
Error en `[.data.frame`(x, order(x, na.last = na.last, decreasing = d
  undefined columns selected
> quantile(data$a)
  0%  25%  50%  75% 100%
 0.3  1.0  1.0  1.0  2.0
>
```

Algunos ejemplos aplicación de funciones

R Console

```
> cumsum(data$a) #suma acumulativa
[1] 0.3 0.7 1.7 2.7 4.7 5.7 6.7 7.7 8.7
> quantile(data$a,probs=c(0.1,0.4,0.9))  #cuantiles ad-hoc
 10%  40%  90%
0.38 1.00 1.20
> sample(1:10,5,rep=T)  # muestreo con repetición
[1] 8 4 4 4 9
> sample(1:100,5,rep=T) # muestreo con repetición
[1] 84 71 46 99 45
> sample(1:100,5)# sin repetición
[1] 92 58 85 90 61
> # PROBABILIDADES
> rnorm(1)     # Generación de un dato de la normal estandar
[1] 0.8210304
> rnorm(5,mean=10,sd=3.4)    # Generación de un dato de una normal no estandar
[1] 15.33650 17.44111 13.21422 16.74950 10.33399
> dnorm(0)  # Evaluación de la función de densidad normal en el punto 0
[1] 0.3989423
> pnorm(0)  # Probabilidad acumulada bajo la normal en el punto 0
[1] 0.5
>
```

Distribuciones de probabilidad

binom	Binomial
cauchy	Cauchy
chisq	Chi cuadrado
beta	Beta
exp	Exponencial
gamma	Gamma
geom	Geométrica
hyper	Hipergeométrica
lnorm	Log-normal
logis	Logística
nbinom	Binomial negativa
nchisq	Chi cuadrado no central
norm	Normal
pois	Poisson
signrank	Distribución del test de Wilcoxon de rangos con signo
t	Student
unif	Uniforme
weibull	Weibull
wilcox	Distribución de la suma de rangos de Wilcoxon

Trabajar con los prefijos : r= un dato,valor ; d= abcisa densidad; p=prob. acumulada; q=cuantiles

Trabajar con matrices

chol	Descomposición de Cholesky
crossprod	Producto cruzado: crossprod(x,y) es lo mismo que t(x) %*% y
diag	Crea una matriz diagonal o extrae la diagonal de una matriz
eigen	Valores propios
outer	Producto exterior de dos vectores
scale	Escala las columnas de una matriz
solve	Resuelve sistemas de ecuaciones lineales y calcula la inversa
svd	Descomposición en valores singulaers
qr	Descomposición QR
t	Traspuesta

Ejemplo

R Console

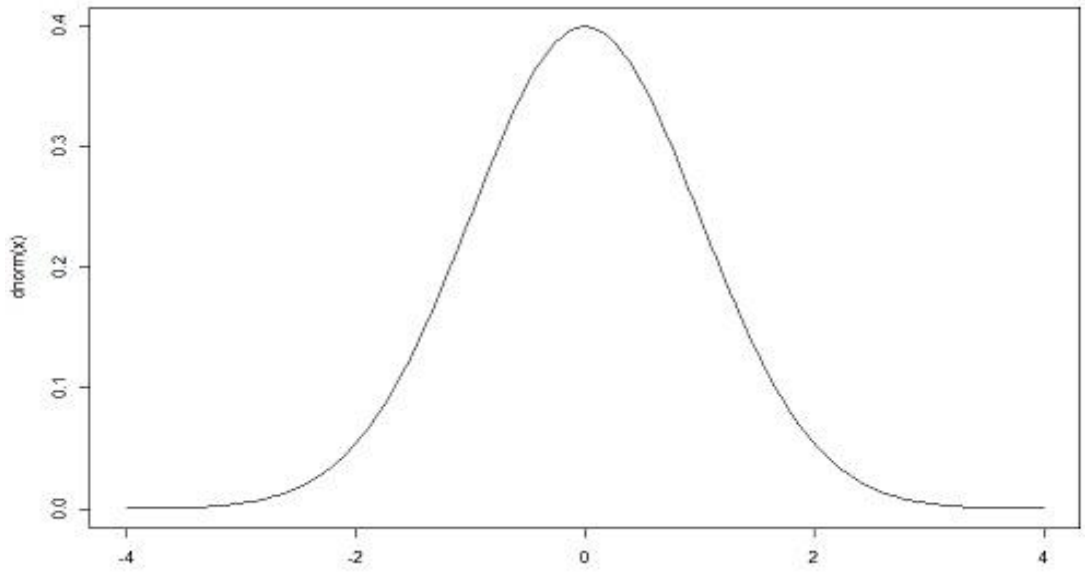
```
> A<-matrix(c(2,3,5,2),ncol=2,byrow=T)
> b<-c(8,9)
> solve(A,b)
[1] 1 2
> # ha resuelto el sistema 2x+3y=8 5x+2y=9
>
```

Gráficos **demo("graphics")**

Gráficos , distribución Normal

x<-seq(-4,4,length=200) generamos 200 valores entre -4,4

plot(x,dnorm(x),type="l") dibujamos



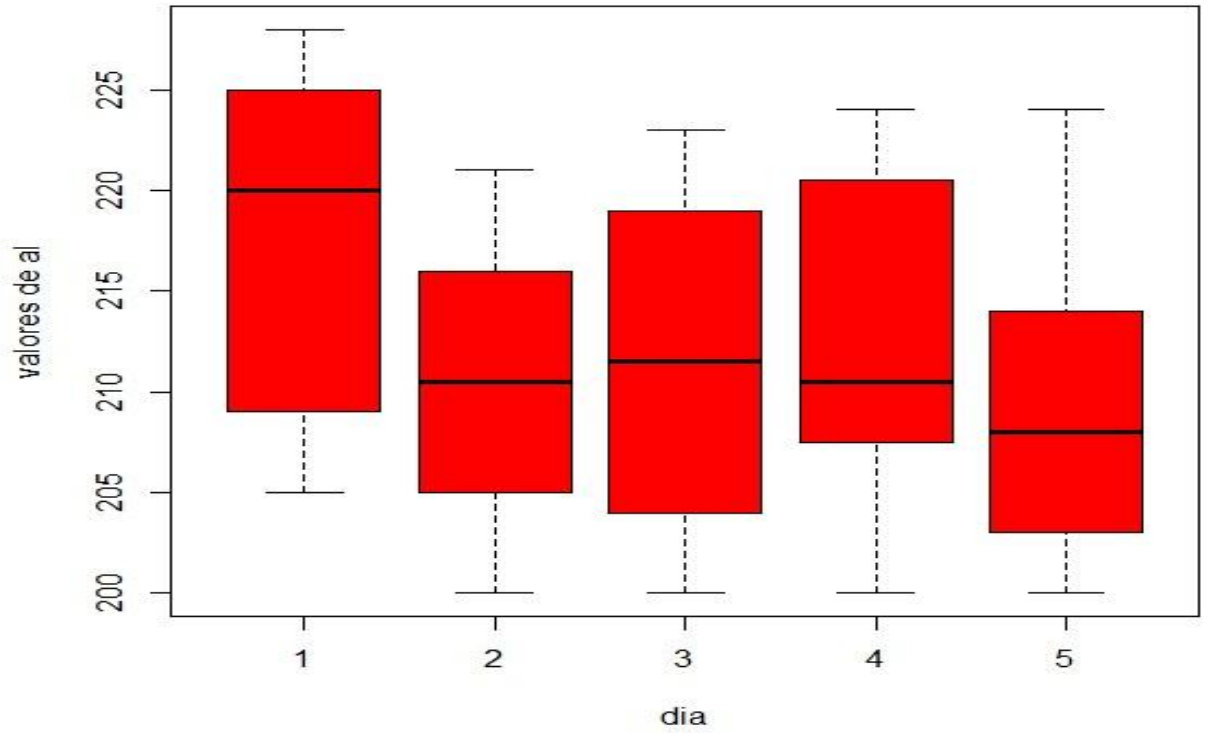
Ejemplo de gráfico en avo.csv

RGui (32-bit) - [R Console]

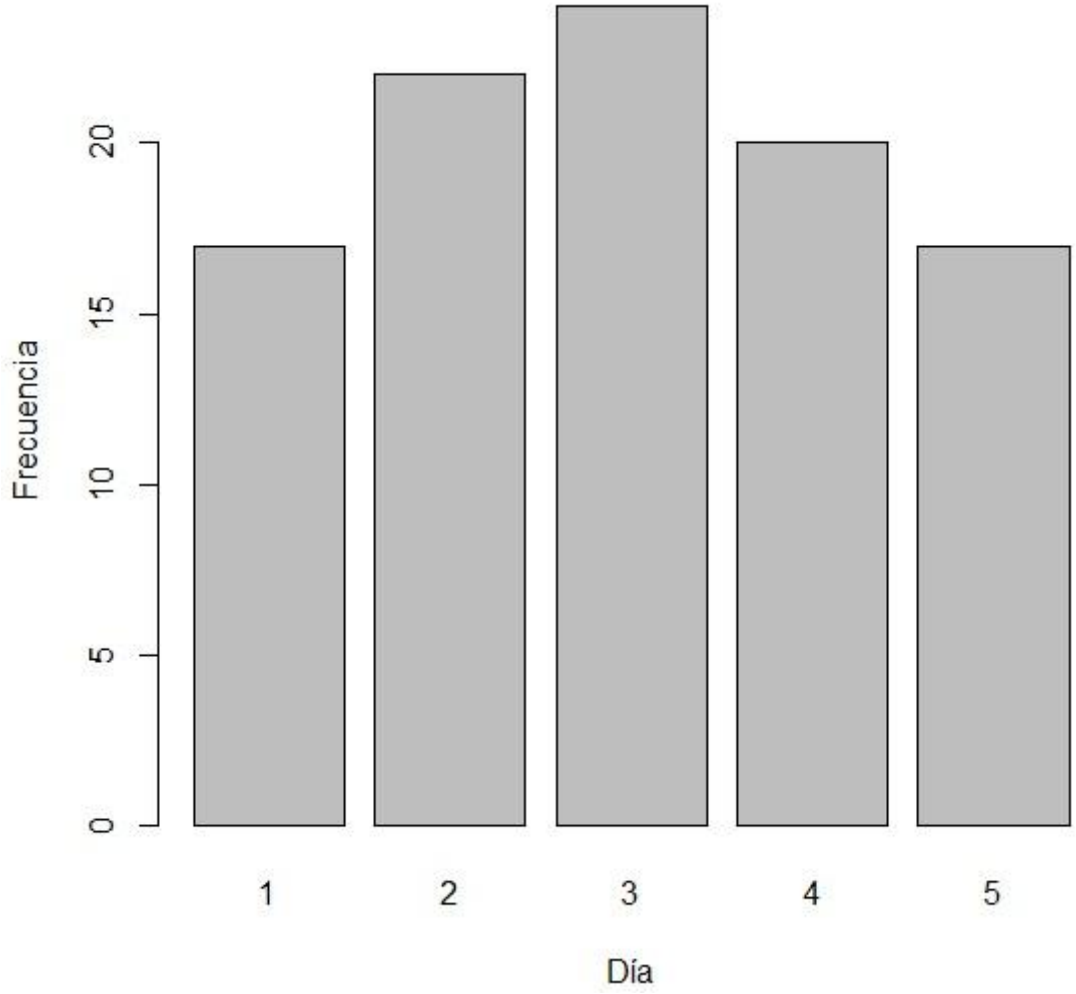
Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda

```
> data<-read.csv("c:/datos/avo.csv",header=T, sep=";")
> data
     al dia
1   221   1
2   228   1
3   209   1
4   209   1
5   220   1
6   228   1
7   205   1
8   228   1
```

boxplot(al~dia, ylab="valores de al", xlab="dia",col="red" ,data=data)



barplot(table(data$dia), xlab="Día", ylab="Frecuencia")



```
stripchart(data$al, method = "stack", pch = 2)
```

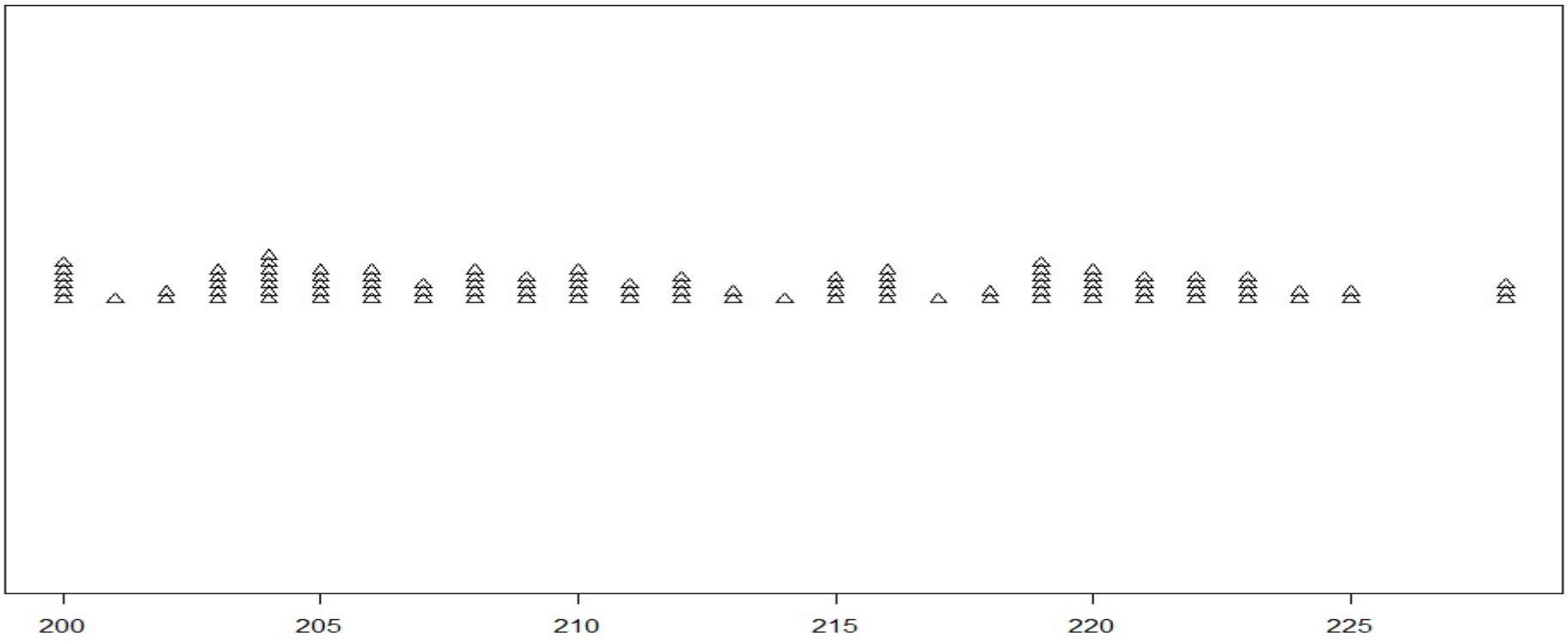


Diagrama de sectores

RGui (32-bit) - [R Console]

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda

```
> z <- c(15, 25, 36, 40)
> z.nombres <- c("ni~nos", "jovenes", "maduros", "ancianos")
> pie(z, labels = z.nombres)
>
```