Introducción a R

Felipe José Bravo Márquez

13 de noviembre de 2013

Motivación

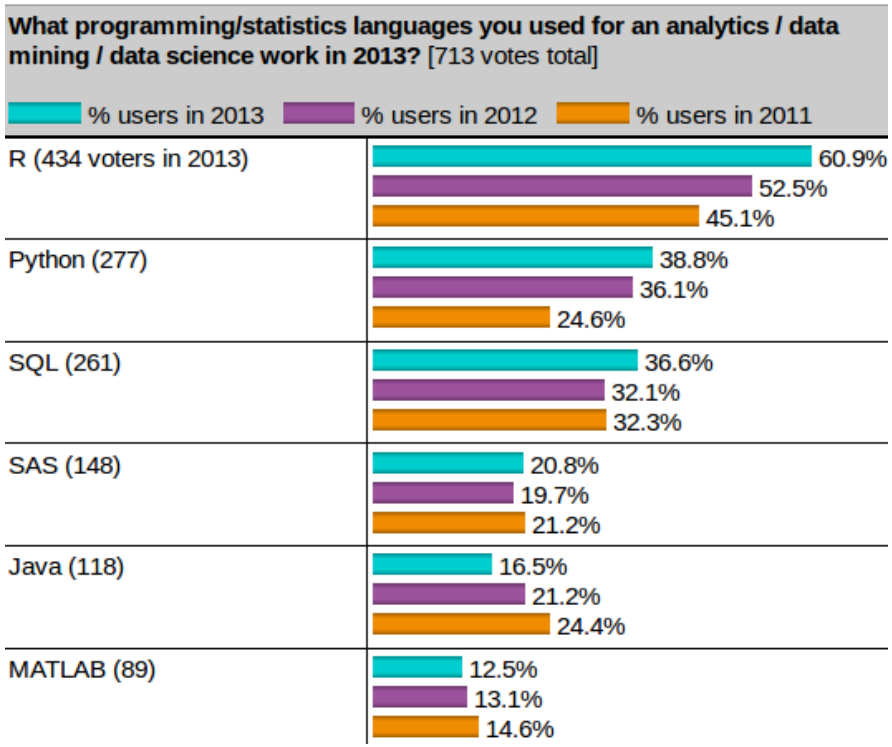
- Diaramente se almacenan masivamente grandes colecciones de datos.
- Ej: La Web, comercio electrónico, datos transaccionales.
- Los computadores se vuelven cada vez más baratos y con mayor poder de procesamiento.
- Analizar estos datos permite encontrar patrones ocultos.
- Un buen uso de los datos puede traer beneficios de negocio. Ej: segmentación de clientes, predicción de demanda.

El proyecto R para la estadística computacional

- R es un ambiente de programación estadístico totalmente gratuito: `http://www.r-project.org/`
- Permite manipular y almacenar datos de manera efectiva.
- Es un lenguaje de programación completo: variables, loop, condiciones, funciones.
- Provee muchas librerías para realizar distintos tipos de análisis sobre colecciones de datos, ej: visualización de datos, análisis de series temporales, análisis de grafos, análisis de texto.
- Las librerías junto a sus dependencias se encuentra ordenadas en un repositorio llamado **CRAN**: `http://cran.r-project.org/`

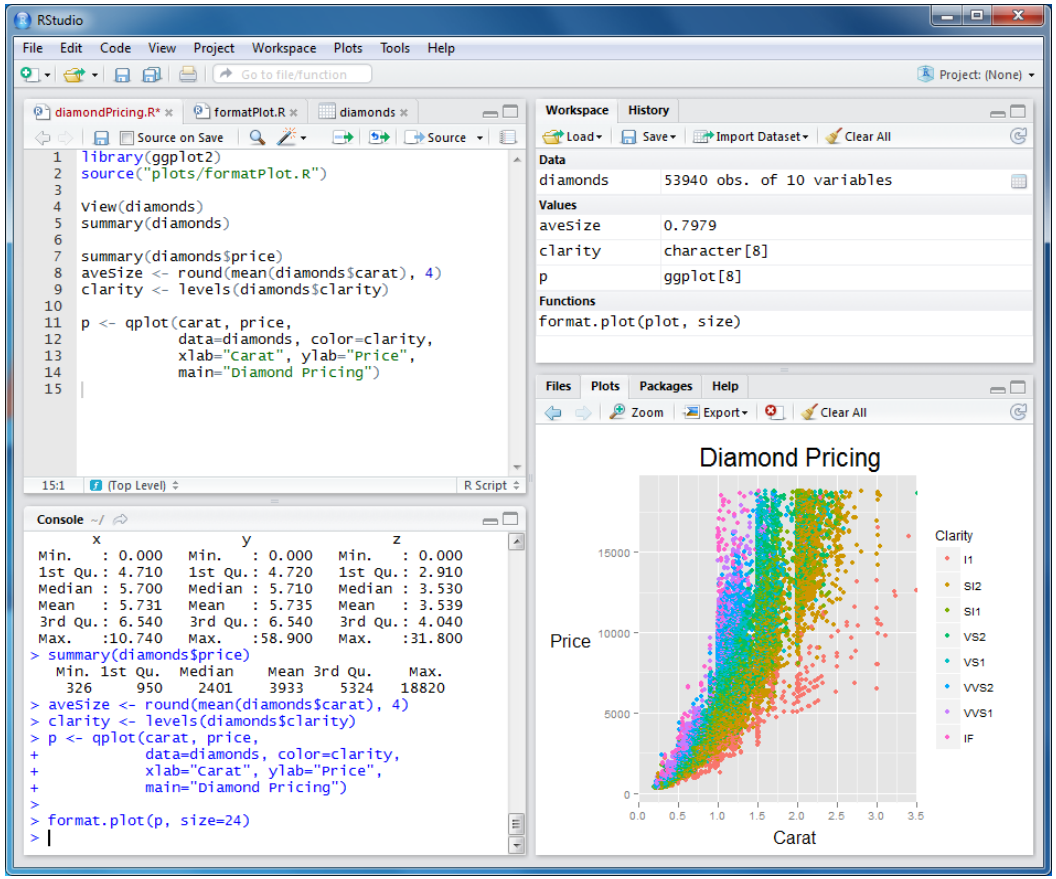
¿Por qué usar R?

- R es software libre a diferencia de Matlab, SPSS, STATA.
- Esta disponible para muchos sistemas operativos: Windows, MAC OS X, Linux.
- Según la última encuesta de **KDnuggets**, R es el lenguaje de programación preferido para realizar análisis de datos, minería de datos y ciencia de datos.
- `http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining-data-science.html`



RStudio

- R funciona a través de la línea de comandos.
- Para trabajar en un entorno más amigable usaremos RStudio.
- También es gratis y se puede descargar para distintos sistemas operativos en este link: `http://www.rstudio.com/ide/download/desktop`



R puede ser usado como una calculadora

```
> 4*5
[1] 20
> 2^3
[1] 8
> exp(-5)
[1] 0.006737947
> log(4)
[1] 1.386294
```

Declarando Variables

- Las variables se pueden asignar usando `<-` , `=` o la función `assign`

```
a<-1
b=3
assign("tres",3)
d<-a+b
ver<-T # equivalente a TRUE
pal<-"hola"
```

- Por convención usamos la primera forma (`<-`).
- Las variables pueden ser de clase **numeric**, **factor**, **character**, **logical**, entre otras.
- Para ver el tipo de una variable usamos el comando `class`.

```
  > class(a)
[1] "numeric"
> class(ver)
[1] "logical"
> class(pal)
[1] "character"
```

Funciones

- Las funciones se declaran como variables y se crean con la expresión **function**:

```
suma<-function(a=2,b=1){
  a+b;
}

fac<-function(n){
  ifelse(n==1,return(1),return(n*factorial(n-1)))
}
```

- Los parámetros de la función se pueden declaran con un valor específico para usarlos como valores predeterminados cuando no entregamos valores para esos parámetros:

```
> suma(3,4)
[1] 7
> suma()
[1] 3
```

- Las funciones son del tipo **function**:

```
> class(suma)
[1] "function"
```

Ayuda y el Workspace

- Para leer documentación sobre una función usamos **help** o **?**:

```
help(ls)
?ls
#Para un comando particular
help("for")
```

- Todas las variables quedan en mi ambiente **workspace**. Para listarlos se usa el comando **objects** o **ls**. Para borrar una variable usamos **rm**:

```
objects()
ls()
rm(a)
#Para borrarlos todos
rm(list=ls())
```

- Puedo grabar todas mis variables de workspace en un archivo y así recuperar mi trabajo en una sesión futura:

```
save.image("myworkspace.RData")
#Luego lo cargamos
load("myworkspace.RData")
```

Vectores

- Para trabajar con colecciones de elementos declaramos **vectores** que se construyen con el comando **c**:

```
edades<-c(21,33,12,34,23,70,90,80,7,29,14,2,
          88,11,55,24,13,11,56,28,33)
```

- Para obtener el largo de un vector usamos el comando **length**, luego para obtener la suma de todos los elementos usamos **sum**:

```
> suma<-sum(edades)
> largo<-length(edades)
> suma
[1] 734
> largo
[1] 21
```

- Si operamos un vector por un escalar este valor se recicla para todos los elementos del vector:

```
> numeros<-c(1,2,3)
> numeros+3
[1] 4 5 6
> numeros*5
[1]  5 10 15
```

Vectores (2)

- Calcular la media y la varianza del vector edades usando los comandos **sum** y **length** en base a las siguientes ecuaciones:

$$\text{media}(\text{edades}) = \frac{\sum_{i=1}^{n} \text{edades}_i}{n} \quad (1)$$

$$\text{varianza}(\text{edades}) = \frac{\sum_{i=1}^{n} (\text{edades}_i - \text{media}(\text{edades}))^2}{n-1} \quad (2)$$

Vectores (4)

- Respuesta:

```
> media<-sum(edades)/length(edades)
> media
[1] 34.95238
> varianza<-sum((edades-media)^2)/(length(edades)-1)
> varianza
[1] 747.9476
```

- R dispone de funciones **mean** y **var**:

```
> mean(edades)
[1] 34.95238
> var(edades)
[1] 747.9476
```

Vectores (5)

- Cuando construimos vectores con elementos de distinto tipo, R los convierte todos a un tipo único:

```
> c("hola",2,T)
[1] "hola" "2"    "TRUE"
> c(TRUE,FALSE,500)
[1]   1   0 500
```

- Los elementos de un vector pueden se declarados con nombres para luego recuperarlos con el comando **names**:

```
> notas<-c(Juan=4.5,Luis=6.2,Romina=3.9,Felipe=2.8,Mariana=6.7)
> names(notas)
[1] "Juan"    "Luis"    "Romina"  "Felipe"  "Mariana"
```

- Podemos ordenar un vector usando el comando **sort**:

```
> names(sort(x=notas,decreasing=T))
[1] "Mariana" "Luis"    "Juan"    "Romina"  "Felipe"
```

Acceso Vectores

- R permite acceder a los elmentos de un vector por medio de índices numéricos `[i]`:

```
> notas[1] # primer elemento
Juan
 4.5
```

- El índice puede ser otro vector número para acceder a más de un elemento:

```
> notas[c(1,5)] # primer y quinto elemento
   Juan Mariana
    4.5     6.7
```

- Si queremos omitir algún elemento usamos índices negativos:

```
> notas[-2] # Todos menos el segundo
   Juan  Romina  Felipe Mariana
    4.5     3.9     2.8     6.7
```

- También se pueden acceder a los elementos por sus nombres:

```
> notas[c("Juan","Mariana")] # Sólo Juan y Mariana
   Juan Mariana
    4.5     6.7
```

Operando Vectores

- Vimos anteriormente que si opero un escalar por un vector, el escalar se aplica a todos los elementos del vector.
- Si tengo ahora dos vectores del mismo largo y los opero, la operación se hace elemento por elemento:

```
a<-c(1,2)
b<-c(3,4)
> a+b
[1] 4 6
> a*b
[1] 3 8
```

Operando Vectores (2)

- Si los vectores son de largo distinto, el más pequeño recicla sus elementos:

```
> d<-c(4,5,6,9)
> a+d
[1]  5  7  7 11
> c(a,a)+d
[1]  5  7  7 11
```

- Si el largo del mayor no es múltiplo del largo del menor, recibimos una advertencia:

```
> c(1,2)+c(-9,2,3)
[1] -8  4  4
Warning message:
In c(1, 2) + c(-9, 2, 3) :
  longer object length is not a multiple of shorter object length
```

Comparando Vectores

- R soporta los operadores de comparación para variables numéricas:`>`, `<`, `==`, `<=`, `>=`, `!=` además de `&` `|` como los operadores **and** y **or** para variables lógicas:

```
> menores<-edades<18
> menores
 [1] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
[17]  TRUE  TRUE FALSE FALSE FALSE
```

- Si le damos a un vector un índice de variables lógicas recuperamos los valores donde el índice toma el valor verdadero:

```
> edades[menores]
[1] 12  7 14  2 11 13 11
```

- Ejercicio: calcular el promedio de edad de los elementos mayores de 18 años.

```
mean(edades[edades>=18])
```

Valores Nulos

- En R, los valores faltantes se escriben como `NA`. Es común que aparezcan cuando leemos datos de alguna base de datos. Algunas funciones no aceptan valores nulos por lo que hay que tenerlos en cuenta.

```
> missing_vector<-c(12,15,NA)
> missing_vector
[1] 12 15 NA
```

- Para chequear si una variable es nula usamos el comando `is.na`:

```
> missing_vector[!is.na(missing_vector)]
[1] 12 15
```

Secuencias

- Para crear un vector formado por una secuencia de números usamos el comando **seq**:

```
> pares<-seq(from=2,to=20,by=2)
> cuatro_mult<-seq(from=4,by=4,length=100)
> pares
 [1]  2  4  6  8 10 12 14 16 18 20
```

- También se pueden crear usando el operador (`:`):

```
> 1:10
 [1]  1  2  3  4  5  6  7  8  9 10
> seq(1,10,1)
 [1]  1  2  3  4  5  6  7  8  9 10
```

Repeticiones

- Para crear vectores que repitan un valor u otro vector varias veces usamos el comando **rep**. El primer valor es el objeto a repetir y el segundo es el número de repeticiones:

```
> rep(10,3)
[1] 10 10 10
> rep(c("hola","chao"),4)
[1] "hola" "chao" "hola" "chao" "hola" "chao" "hola" "chao"
```

- Problema: Crear una secuencia que repita 3 veces los 4 primeros múltiplos de 7.

```
> rep(seq(from=7,by=7,length=4),3)
 [1]  7 14 21 28  7 14 21 28  7 14 21 28
```

Generación de vectores aleatorios

- Para realizar experimentos o simular fenómenos de comportamiento conocido es muy útil generar vectores aleatorios.
- Si queremos números uniformemente distribuidos entre un máximo y un mínimo usamos **runif**:

```
> runif(n=5, min = 1, max = 10)
[1] 5.058862 1.737830 9.450956 9.149376 2.652774
```

- Si queremos números centrados en una media μ y con una desviación estándar σ, usamos una distribución normal con **rnorm** donde sabemos que el 68 % de las observaciones estarán alrededor $\mu \pm \sigma$, el 95 % en $\mu \pm 2\sigma$ y el 99,7 % en $\mu \pm 3\sigma$:

```
> rnorm(n=5, mean = 10, sd = 4)
[1] 12.081286  2.636001 16.001953  0.120463  6.211835
```

Generación de vectores aleatorios (2)

- Cuando queremos modelar un número de arribos por unidad de tiempo para simular modelos de colas, usamos la distribución de **Poisson** con **rpos**. El parámetro λ nos dice la cantidad promedio de llegadas en un período:

```
> rpois(n=10, lambda = 3)
 [1] 1 3 8 6 1 1 6 3 4 7
```

- Un experimento de distribución binomial se basa en tener *n* experimentos, donde en cada experimento realizamos *k* intentos de un fenómeno cuya probabilidad de acierto en cada intento es *p*. Con el comando **rbinom** podemos simular la cantidad de aciertos obtenidos en cada experimento.

```
> rbinom(n=10,size=2,prob=0.5)
 [1] 0 1 2 1 1 0 2 0 0 1
> rbinom(n=10,size=2,prob=0.7)
 [1] 1 2 2 1 0 1 2 2 2 2
> rbinom(n=10,size=2,prob=0.2)
 [1] 0 0 0 0 1 0 1 0 1 0
```

Variables Categóricas o Factores

- Además de las variables numéricas o lógicas, se puede trabajar con variables categóricas. Ej: color, sexo, clase social.
- Se crean con el comando **factor** y los posibles valores de la variable se guardan en el atributo **levels**.

```
> gente<-factor(c("Hombre","Mujer","Mujer","Mujer","Hombre"))
> gente
[1] Hombre Mujer  Mujer  Mujer  Hombre
Levels: Hombre Mujer
> class(gente)
[1] "factor"
> levels(gente)
[1] "Hombre" "Mujer"
#Puedo renombrar a los niveles
> levels(gente)<-c("Man","Woman")
> gente
[1] Man   Woman Woman Woman Man
Levels: Man Woman
```

Agregando variables por categorías con **tapply**

- Si tenemos un vector numérico y otro categórico del mismo largo podemos aplicar una función de agregación.
- Ejemplo: Creo una categoría para el vector edades de niveles *niño*, *adolescente*, *adulto*:

```
categ_edades<-ifelse(edades<12,"niño",
                     ifelse(edades<18,"adolescente","adulto"))
class(categ_edades)
[1] "character"
#Convierto a factor con as.factor
categ_edades<-as.factor(categ_edades)
```

- Ahora cuento la cantidad de personas por categoría, y calculo la media y la desviación estándar para cada grupo:

```
tapply(edades,categ_edades,length)
adolescente      adulto        niño
          3          14           4
> tapply(edades,categ_edades,mean)
adolescente      adulto        niño
   13.00000    47.42857     7.75000
> tapply(edades,categ_edades,sd)
adolescente      adulto        niño
   1.000000   25.294312    4.272002
```

Manejo de Strings

- Puedo imprimir un string usando el comando **cat**:

```
> saludo<-"Hola Mundo"
> cat(saludo)
Hola Mundo
```

- Para concatenar dos strings uso el comando **paste**:

```
> paste("Hola","Chao",sep="-")
[1] "Hola-Chao"
> paste("persona",1:4, sep="")
[1] "persona1" "persona2" "persona3" "persona4"
> paste(saludo,1:3, sep=" ")
[1] "Hola Mundo 1" "Hola Mundo 2" "Hola Mundo 3"
```

- Para extraer sub-cadenas usamos el comando **substr**:

```
> substr(saludo,1,4)
[1] "Hola"
```

- Existe un vector llamado **letters** que tiene todas las letras del abecedario, útil para nombrar variables:

```
> letters[1:4]
[1] "a" "b" "c" "d"
```

Matrices

- Las matrices son vectores de dos dimensiones. Por defecto se van llenando por columna:

```
> matriz_por_col<-matrix(data=1:12,nrow=3,ncol=4)
> matriz_por_col
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

- Para llenarlas por fila uso el parámetro **byrow**:

```
> matriz_por_fil<-matrix(data=1:12,nrow=4,ncol=3,byrow=T)
> matriz_por_fil
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
[4,]   10   11   12
```

- Accedemos a la dimensión de la matriz con el comando **dim**.

```
 > dim(matriz_por_fil)
[1] 4 3
```

Matrices (2)

- Para acceder a los elementos de una matriz tengo que especificar las filas *i* y las columnas *j* `[i,j]`. Si dejo alguno de los dos valores vacío se recuperan todos las filas o columnas:

```
 > matriz_por_fil[2,] #Segunda fila, todas las columnas
[1] 4 5 6
> matriz_por_fil[2,1] # Segunda fila, primera columna
[1] 4
> matriz_por_fil[-1,-2] # Descarto fila 1 y columna 2
     [,1] [,2]
[1,]    4    6
[2,]    7    9
[3,]   10   12
```

- Para acceder a los nombres de las filas o columnas usamos **rownames** y **colnames** de forma análoga a como usamos **names** para los vectores.

```
> rownames(matriz_por_fil)<-paste("r",1:4,sep="")
> colnames(matriz_por_fil)<-paste("c",1:3,sep="")
> matriz_por_fil["r2","c3"]
[1] 6
```

Matrices (3)

- Puedo agregarle nuevas filas o nuevas columnas a una matriz usando **rbind** y **cbind** respectivamente:

```
  > rbind(matriz_por_fil,r5=1:3)
   c1 c2 c3
r1  1  2  3
r2  4  5  6
r3  7  8  9
r4 10 11 12
r5  1  2  3
> cbind(matriz_por_fil,c4=4:1)
   c1 c2 c3 c4
r1  1  2  3  4
r2  4  5  6  3
r3  7  8  9  2
r4 10 11 12  1
```

Matrices (4)

- Operaciones algebraicas como la multiplicación de matrices se hace con `%*%`:

```
>a<-matriz_por_col %*% matrix_por_fil
      c1  c2  c3
[1,] 166 188 210
[2,] 188 214 240
[3,] 210 240 270
```

- Si usamos solamente el operador `*`, la multiplicación se hace elemento por elemento (sólo para matrices de igual dimensión). Esto aplica también para la suma, la resta, la división y otro tipo de operadores.

Matrices (5)

- Podemos transponer una matriz con **t**:

```
> t(a)
   [,1] [,2] [,3]
c1  166  188  210
c2  188  214  240
c3  210  240  270
```

- Los valores y vectores propios se calculan con **eigen**:

```
> eigen(a)
$values
[1] 6.483342e+02 1.665808e+00 3.437970e-14

$vectors
           [,1]        [,2]       [,3]
[1,] -0.5045331 -0.76077568  0.4082483
[2,] -0.5745157 -0.05714052 -0.8164966
[3,] -0.6444983  0.64649464  0.4082483
```

Arreglos

- Los arreglos son como las matrices pero de más dimensiones:

```
> arreglo<-array(1:8, dim=c(2,2,2))
> arreglo
, , 1

     [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2

     [,1] [,2]
[1,]    5    7
[2,]    6    8

> arreglo[1,2,1]
[1] 3
```

Listas

- Las matrices me restringen a que todos los vectores sean del mismo largo y del mismo tipo.
- Las listas me permiten agrupar objetos de cualquier tipo y de cualquier largo:

```
milista<-list(hombre="Pepe",mujer="Juana",
              hijos=3,edades=c(4,8,12))
```

- Cuando accedo a sus elementos usando `[i]` recupero una sub-lista:

```
> milista[c(3,4)] # Sublista
$hijos
[1] 3
$edades
[1]  4  8 12
```

- Para acceder a una elemento particular tengo tres opciones:

```
milista[[1]]
milista[["hombre"]]
milista$hombre

[1] "Pepe"
```

Ejercicio Lista

- Crear una lista que tenga tres vectores de largo 100 generado por alguno de los mecanismos vistos para generar vectores aleatorios. Pueden variar las distribuciones o los parámetros. Asígnele nombres a cada uno de los vectores.

```
vectores<-list(normal=rnorm(n=100,mean=10,sd=5),
               poisson=rpois(n=100,lambda=10),
               uniforme=runif(n=100,min=5,max=15))
```

- Calcule la media y la desviación estándar de cada uno de los vectores de la lista.

```
medias<-vector()
desv<-vector()
for(i in 1:length(vectores)){
  medias[i]<-mean(vectores[[i]])
  desv[i]<-sd(vectores[[i]])
}
> medias
[1] 10.589222 10.390000  9.579866
> desv
[1] 5.155478 2.711349 2.905810
```

Cálculos agregados a Listas con **sapply** y **lapply**

- El ejercicio anterior se puede resolver de manera mucho más sencilla en R con unas funciones especiales para realizar agregación sobre listas.
- El comando **sapply** permite aplicar una función a cada elemento de una lista y devuelve los resultados en un vector. Luego **lapply** hace lo mismo pero retorna una lista:

```
 > sapply(vectores,mean)
   normal   poisson  uniforme
10.589222 10.390000  9.579866
> sapply(vectores,sd)
  normal  poisson uniforme
5.155478 2.711349 2.905810
```

- Ejercicio, programar una propia versión de **sapply**. Hint: En R una funciones puede recibir otra función como parámetro y aplicarla de manera genérica.

```
myapply<-function(lista,fun,...){
 resultado<-vector(length=length(lista))
 for(i in 1:length(lista)){
   resultado[i]<-fun(lista[[i]],...)
 }
 resultado
}
```

Data Frames

- El data.frame es el tipo de colección de datos más utilizada para trabajar con datasets en R.
- Un data.frame se compone de varios vectores, donde cada vector puede ser de distintos tipos, pero del mismo largo. Es equivalente a una tabla de una base de datos:

```
edades.frame<-data.frame(edad=edades,categoria=categ_edades)
```

```
> edades.frame
   edad   categoria
1    21       adulto
2    33       adulto
3    12 adolescente
```

- Las dimensiones de un data.frame se acceden de la misma manera que en una matriz:

```
> length(edades.frame)
[1] 2
> dim(edades.frame)
[1] 21  2
```

Data Frames (2)

- Puedo acceder a los elementos como si fuese una matriz o una lista:

```
> edades.frame[3,1] # La edad del tercer elemento
[1] 12
> edades.frame$edad[1:6] # La edad de los primeros 6 elementos
[1] 21 33 12 34 23 70
```

- También puede pasar cada variable del data.frame a mi workspace con el comando **attach** y así accederlas directamente:

```
attach(edades.frame)
> categoria[1:3]
[1] adulto      adulto      adolescente
Levels: adolescente adulto niño
```

- Puedo guardar un data.frame en un archivo csv (separado por comas u otra carácter) usando **write.table**:

```
write.table(x=edades.frame,file="edades.csv",sep=",",row.names=F)
```

- Pongo `row.names=F` para que no ponga los nombres de las columnas en el archivo.

Cargando Data Frames

- Puedo leer un data.frame desde archivos **csv** de manera nativa y desde otras fuentes (Excel, base de datos, etc.) usando librerías especiales:

```
my.frame<-read.table(file="edades.csv",header=T,sep=",")
```

- El parámetro `header` específica si quiero usar la primera fila para asignarle nombres a las columnas.
- Además R provee varias colecciones de datos para experimentar. Se pueden ver como el comando `data()`.
- Para ver todos los datasets disponibles de todas las librerías:

```
data(package = .packages(all.available = TRUE))
```

- Ahora podemos cargar un dataset, que se incluye como data.frame en mi workspace:

```
data(USArrests) # Arrestos en Estados Unidos por estado
```

Muestreo

- Cuando tenemos datasets muy grandes algunas técnicas estadísticas o de visualización pueden ser muy costosas computacionalmente.
- Se puede trabajar con una muestra aleatoria de los datos.
- La idea es que si la muestra es representativa, la propiedades observadas serán equivalentes a las de la población.
- En R se realiza el muestreo con el comando **sample**.
- Si la muestra es sin reemplazo, sacamos datos de manera aleatoria sin reponer el elemento. Entonces la muestra debe ser de menor tamaño que el dataset:

```
> sample(edades,size=4,replace=F)
[1] 80 88 12 23
```

Muestreo (2)

- Si la muestra es con reemplazo poddemos observar datos duplicados. De esta forma, la muestra puede ser incluso de mayor tamaño que la colección original:

```
sample(edades,size=100,replace=T)
```

- Cuando tenemos que los datos vienen etiquetados por alguna categoría y tomamos una muestra donde cada categoría tiene una participación proporcional a la de la colección original, tenemos un muestreo estratificado.
- Ejercicio: extraer una muestra aleatoria sin reemplazo que tenga 10 filas del data.frame **USArrests**.

```
USArrests[sample(1:(dim(USArrests)[1]),size=10,replace=F),]
```

Instalando librerías adicionales

- R tiene una comunidad muy activa que desarrolla muchas librerías para el análisis y la visualización de datos.
- Se pueden descargar librerías adicionales desde el repositorio CRAN directamente desde R.
- Las librerías se pueden instalar desde Rstudio o con el siguiente comando:

```
install.packages("rpart",dependencies=T)
```

- Luego para poder usarlas se cargan de la siguiente forma: `library(rpart)`.

Bilbiografía I

Venables, William N., David M. Smith, and R Development Core Team. *An introduction to R.*, 2002.