Subconjuntos de diferentes estructuras de datos

Esta sección está basada en Wickham (2014, Subsetting) disponible en línea.

Aprender a extraer subconjuntos de los datos es importante y permite realizar operaciones complejas con los mismos. De los conceptos importantes que se deben aprender son

- Los operadores para extraer subconjuntos (subsetting operators)
- Los 6 tipos de extracciones de subconjuntos
- Las diferencias a la hora de extraer subconjuntos de las diferentes estructuras de datos (factores, listas, matrices, dataframes)
- El uso de la extracción de subconjuntos junto a asignar variables.

Cuando tenemos que extraer pedazos de los datos (o analizar solamente parte de éstos), necesitamos complementar `str()` con `[[`, es decir, la estructura nos dirá cómo utilizar el operador subconjunto de manera que de hecho extraigamos lo que queremos.

Operadores para extraer subconjuntos

Dependiendo la estructura de datos que tenemos, será la forma en la que extraemos elementos de ella. Hay dos operadores de subconjunto: `[[` y `$`. `[[` se parece a `[` pero regresa un solo valor y te permite sacar pedazos de una lista. `$` es un atajo útil para `[[`.

Vectores atómicos

¿De qué formas puedo extraer elementos de un vector? Hay varias maneras **sin importar** la *clase* del vector.

- **Enteros positivos** regresan los elementos en las posiciones especificadas en el orden que especificamos.

```
x <- c(5.6, 7.8, 4.5, 3.3)

x[c(3, 1)]
```

```
## [1] 4.5 5.6
```

```
## Si duplicamos posiciones, nos regresa resultados duplicados
x[c(1, 1, 1)]
```

```
## [1] 5.6 5.6 5.6
```

```
## Si usamos valores reales, se coerciona a entero
x[c(1.1, 2.4)]
```

```
## [1] 5.6 7.8
```

```
x[order(x)]
```

```
## [1] 3.3 4.5 5.6 7.8
```

```
x[order(x, decreasing = T)]
```

```
## [1] 7.8 5.6 4.5 3.3
```

- **Enteros negativos** omiten los valores en las posiciones que se especifican.

```
x
```

```
## [1] 5.6 7.8 4.5 3.3
```

```
x[-c(3, 1)]
```

```
## [1] 7.8 3.3
```

Mezclar no funciona.

```
x[c(-3, 1)]
```

- **Vectores lógicos** selecciona los elementos cuyo valor correspondiente es `TRUE`. Esta es una de los tipos más útiles.

```
x[c(TRUE, TRUE, FALSE, FALSE)]
```

```
## [1] 5.6 7.8
```

```
x[c(TRUE, FALSE)] # Autocompleta el vector lógico al tamaño de x
```

```
## [1] 5.6 4.5
```

```
x[c(TRUE, TRUE, NA, FALSE)]
```

```
## [1] 5.6 7.8  NA
```

- **Nada** si no especifico nada, me regresa el vector original

```
x[]
```

```
## [1] 5.6 7.8 4.5 3.3
```

- **Cero** el índice cero no aplica en R, te regresa el vector vacio

```
x[0]
```

```
## numeric(0)
```

- Si el vector tiene **nombres** también los puedo usar.

```
names(x) <- c("a", "ab", "b", "c")
x["ab"]
```

```
##  ab
## 7.8
```

```
x["d"]
```

```
## <NA>
##   NA
```

```
x[grep("a", names(x))]
```

```
##   a  ab
## 5.6 7.8
```

Las **listas** operan básicamente igual a vectores recordando que si usamos `[` regresa una lista y tanto `[[` y `$` extrae componentes de la lista.

Matrices y arreglos

Para estructuras de mayor dimensión se pueden extraer de tres maneras:

- Con vectores múltiples
- Con un solo vector
- Con una matriz

```r
m <- matrix(1:12, nrow = 3)
colnames(m) <- LETTERS[1:4]
m[1:2, ]
```

```
##      A B C  D
## [1,] 1 4 7 10
## [2,] 2 5 8 11
```

```r
m[c(T, F, F), c("B", "C")]
```

```
## B C
## 4 7
```

```r
m[1, 4]
```

```
##  D
## 10
```

Como ven, es solamente generalizar lo que se hace en vectores replicándolo al número de dimensiones que se tiene.

```r
m[c(T, F, F)]
```

```
## [1]  1  4  7 10
```

```r
class(m[c(T, F, F)])
```

```
## [1] "integer"
```

[simplifica al objeto. En matriz, me quita la dimensionalidad, en listas me da lo que esta dentro de esa celda.

Dataframes

```r
df <- data.frame(x = 1:3, y = 3:1, z = letters[1:3])

df[c(1, 2), ]
```

```
##   x y z
## 1 1 3 a
## 2 2 2 b
```

```r
df[, c(1, 2)]
```

```
##   x y
## 1 1 3
## 2 2 2
## 3 3 1
```

```r
df[, c("z", "x")]
```

```
##   z x
## 1 a 1
## 2 b 2
## 3 c 3
```

```r
df[c("z", "x")]
```

```
##   z x
## 1 a 1
```

```
## 2 b 2
## 3 c 3
```

```
class(df[, c("z", "x")])
```

```
## [1] "data.frame"
```

```
class(df[c("z", "x")])
```

```
## [1] "data.frame"
```

```
str(df["x"])
```

```
## 'data.frame':    3 obs. of  1 variable:
##  $ x: int  1 2 3
```

```
str(df[, "x"])
```

```
##  int [1:3] 1 2 3
```

```
str(df$x)
```

```
##  int [1:3] 1 2 3
```

Ejercicios

1. Utiliza la base mtcars
2. Arregla los errores al extraer subconjuntos en dataframes
   ```
   mtcars[mtcars$cyl = 4, ]
   mtcars[-1:4, ]
   mtcars[mtcars$cyl <= 5]
   mtcars[mtcars$cyl == 4 | 6, ]
   ```
3. ¿Por qué al correr *x <- 1:5; x[NA]* obtengo valores perdidos?
4. Genera una matriz cuadrada tamaño 5 llamada m. ¿Qué te da correr m[upper.tri(m)]?
5. ¿Por qué al realizar mtcars[1:20] me da un error? ¿Por qué mtcars[1:2] no me lo da? ¿Por qué mtcars[1:20,] es distinto?
6. Haz una función que extraiga la diagonal de la matriz m que creaste antes. Debe dar el mismo resultado que ejecutar diag(m)
7. ¿Qué hace df[is.na(df)] <- 0?

Asignar a un subconjunto

Muchas veces lo que necesitamos es encontrar ciertos valores para poder reemplazarlos con algo más. Por ejemplo, muchas veces queremos imputar valores perdidos con cierto valor.

```
# Variables continuas
x <- c(1, 2, 3, NA, NaN, 7)
media <- mean(x, na.rm = T)
media
```

```
## [1] 3.25
```

```
x[is.na(x)] <- media
x
```

```
## [1] 1.00 2.00 3.00 3.25 3.25 7.00
```

```
# Variables discretas
x <- c(rep("azul", 3), "verde", NA, "verde", rep("rojo", 4))
x
```

```
##  [1] "azul"  "azul"  "azul"  "verde" NA      "verde" "rojo"  "rojo"
##  [9] "rojo"  "rojo"
```

```
moda <- names(table(x))[which(table(x) == max(table(x)))] # Engorroso, no?
x[is.na(x)] <- moda
x
```

```
##  [1] "azul"  "azul"  "azul"  "verde" "rojo"  "verde" "rojo"  "rojo"
##  [9] "rojo"  "rojo"
```

```
# Puedo reemplazar partes de un vector
x <- 1:5
x[c(1, 2)] <- 2:3
x
```

```
## [1] 2 3 3 4 5
```

```
# Las longitudes de las asignaciones tienen que ser iguales
x[-1] <- 4:1
x
```

```
## [1] 2 4 3 2 1
```

```
# No se revisan duplicados
x[c(1, 1)] <- 2:3
x
```

```
## [1] 3 4 3 2 1
```

```
# Puedo sustituir valores considerando toda la logica
x <- c(1:10)
x[x > 5] <- 0
x
```

```
##  [1] 1 2 3 4 5 0 0 0 0 0
```

Por último, es útil notar la utilidad de asignar utilizando la forma de asignar **nada** mencionada anteriormente.

```
class(mtcars)
```

```
## [1] "data.frame"
```

```
mtcars[] <- lapply(mtcars, as.integer)
class(mtcars)
```

```
## [1] "data.frame"
```

```
dim(mtcars)
```

```
## [1] 32 11
```

```
mtcars <- lapply(mtcars, as.integer)
class(mtcars)
```

```
## [1] "list"
```

```
dim(mtcars)
```

```
## NULL
```

Asignar utilizando el operador de suconjunto a nada nos permite preservar la estructura del objeto original así como su clase.

En el caso de listas, si combinamos un operador de subconjunto mas asignación a nulo, podemos remover objetos de ésta.

```r
x <- list(a = 1, b = 2)
x[[2]] <- NULL
str(x)
```

```
## List of 1
##  $ a: num 1
```

```r
x["b"] <- list(NULL)
str(x)
```

```
## List of 2
##  $ a: num 1
##  $ b: NULL
```

Operadores lógicos

Operador	Descripción
<	menor que
<=	menor o igual que
>	mayor que
==	exactamente igual que
!=	diferente de
!x	no x
x | y	x O y
x & y	x Y y
isTRUE(x)	checa si x es verdadero

Ejemplo: Supongamos que queremos saber qué elementos de x son menores que 5 ó mayores que 8.

```r
x <- c(1:10)
x[(x>8) | (x<5)]
```

```
## [1]  1  2  3  4  9 10
```

```r
# ¿Cómo funciona?
x
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```r
x > 8
```

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

```r
x < 5
```

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

```r
x > 8 | x < 5
```

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

```r
# x > 8 || x < 5
x[c(T,T,T,T,F,F,F,F,T,T)]
```

```
## [1]  1  2  3  4  9 10
```

`||` vs. `|` y `&&` vs. `&`
La diferencia entre & y && (o | y ||) es que el primero es vectorizado y el segundo no.

Ejercicio ¿Qué crees que pasa en las siguientes situaciones?

```
rm(list = ls())
TRUE || a
FALSE && a
TRUE && a
TRUE | a
FALSE & a
```

La forma larga (la versión doble) no parece ser muy útil. El propósito de ésta es que es más apropiado cuando se programa usando estructuras de control, por ejemplo, en **ifs***.

```
if( c(T, F) ) print("Hola")
```

```
## [1] "Hola"
```

Poner el && me garantiza que la condicional será evaluado sobre un único valor falso/verdadero.

```
(-2:2) >= 0
```

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

```
(-2:2) <= 0
```

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

```
((-2:2) >= 0) && ((-2:2) <= 0)
```

```
## [1] FALSE
```

Ejercicio Explora los siguientes comandos

```
impares <- 1:10 %% 2 == 1
mult.3 <- 1:10 %% 3 == 0

impares & mult.3
impares | mult.3

xor(impares, mult.3)
```

¿Por qué tanto detalle? Aplicaciones

Una de las formas más fáciles de frustrarse con R (y con cualquier otro lenguaje) es no saber decirle al lenguaje lo que se desea hacer. Entender cómo manipular las estructuras de datos y la lógica detrás de su comportamiento ahorra mucho sufrimiento y permite adaptarse ante cosas que necesitamos que aún no se encuentran implementadas por alguien más de una manera más sencilla.

Con saber de subconjuntos podemos realizar varias tareas indispensables.

Buscarv o buscarh

Excel es excelente haciendo estas tareas. Lo malo de excel es que no es **reproducible**. Es muy común que resulte imposible llegar de los datos originales al resultado final pues muchos pasos intermedios de limpieza no están documentados de forma alguna. Un *script* de limpieza nos permite no solamente ir del *raw* a la estructura de datos limpia y analizable sino que permite que alguien más verifique las operaciones que se están realizando, se identifiquen errores y que, cuando nos llega un nuevo mes, sea trivial incluir estos datos al resultado final.

```
rm(list = ls())
x <- c("m", "f", "u", "f", "f", "m", "m")
busca <- c(m = "Male", f = "Female", u = NA)
busca[x]
```

```
##        m        f        u        f        f        m        m
##   "Male" "Female"       NA "Female" "Female"   "Male"   "Male"
```

```
unname(busca[x])
```

```
## [1] "Male"   "Female" NA       "Female" "Female" "Male"   "Male"
```

```
c(m = "humano", f = "humano", u = "desconocido")[x]
```

```
##             m             f             u             f             f
##      "humano"      "humano" "desconocido"      "humano"      "humano"
##             m             m
##      "humano"      "humano"
```

Esto nos permite pegar un vector a una base de datos de acuerdo a una condición.

```
calificaciones <- c(10, 9, 5, 5, 6)
aprueba <- data.frame(
  calificacion = 10:1,
  descripcion = c(rep("excelente", 2), "bueno", rep("aceptable", 2), rep("no satisfactorio", 5)),
  aprobatorio = c(rep(T, 5), rep(F, 5))
)
id <- match(calificaciones, aprueba$calificacion)

aprueba[id, ]
```

```
##     calificacion      descripcion aprobatorio
## 1             10        excelente        TRUE
## 2              9        excelente        TRUE
## 6              5 no satisfactorio       FALSE
## 6.1            5 no satisfactorio       FALSE
## 5              6        aceptable        TRUE
```

Ejercicios

1. Realiza la misma operación con las calificaciones pero utilizando los nombres de las filas, es decir, los rownames(aprueba)
2. Carga la libreria ggplot2 y utiliza la base de datos diamonds
3. Utiliza el comando match para quedarte con las variables cut y x
4. Genera la variable categórica tal que, si el precio es mayor que 5,000 el valor de price.cat es cara, si es mayor que 2,000 es normal y barata en otro caso.

Muestras aleatorias

Podemos utilizar índices enteros para generar muestras aleatorias de nuestras bases de datos o de nuestros vectores.

```
set.seed(102030)
aprueba[sample(nrow(aprueba)), ]
```

```
##    calificacion      descripcion aprobatorio
## 10            1 no satisfactorio       FALSE
```

```
## 2             9        excelente        TRUE
## 8             3 no satisfactorio       FALSE
## 7             4 no satisfactorio       FALSE
## 9             2 no satisfactorio       FALSE
## 3             8            bueno        TRUE
## 1            10        excelente        TRUE
## 6             5 no satisfactorio       FALSE
## 4             7        aceptable        TRUE
## 5             6        aceptable        TRUE
```

```
aprueba[sample(nrow(aprueba), replace = T, size = 5), ]
```

```
##     calificacion      descripcion aprobatorio
## 4              7        aceptable        TRUE
## 1             10        excelente        TRUE
## 4.1            7        aceptable        TRUE
## 7              4 no satisfactorio       FALSE
## 2              9        excelente        TRUE
```

Ejercicios

1. Utiliza la base de datos de iris y genera un conjunto de prueba y uno de entrenamiento correspondientes al 20 y 80 % de los datos, respectivamente.
2. Genera un vector x de tamaño 1000 con realizaciones de una normal media 10, varianza 3.
3. Crea 100 muestras bootstrap del vector x.
4. Calcula la media para cada una de tus muestras.
5. Grafica con la función hist() el vector de medias de tus muestras.
6. Genera un vector l de letras, tamaño 10 y ordénalo. (Usa letters y order).
7. Ordena la base cars de acuerdo a distancia, en forma descendiente (muestra la cola -usa tail- de la base ordenada).

Expande bases

Ahora, a veces tenemos tablas de resumen pero quisieramos extraer los datos originales. Combinamos rep con subconjuntos de enteros para expandir.

```
df <- data.frame(
  color = c("azul", "verde", "amarillo"),
  n = c(4, 3, 5)
)
df
```

```
##      color n
## 1     azul 4
## 2    verde 3
## 3 amarillo 5
```

```
df[rep(1:nrow(df), df$n), ]
```

```
##        color n
## 1       azul 4
## 1.1     azul 4
## 1.2     azul 4
## 1.3     azul 4
## 2      verde 3
## 2.1    verde 3
```

```
## 2.2    verde 3
## 3   amarillo 5
## 3.1 amarillo 5
## 3.2 amarillo 5
## 3.3 amarillo 5
## 3.4 amarillo 5
```

Otras

Ya estuvimos utilizando otras aplicaciones de estos comandos: ordenamientos, selección de filas o columnas según una condición lógica.

También utilizamos un comando muy útil llamado which.

```
set.seed(45)
x <- sample(letters, 10)
x <= "e"
```

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

```
which(x <= "e")
```

```
## [1]  7  9 10
```

Junto con which, puedes usar intersect y union.

```
pares <- 1:10 %% 2 == 0
m.5 <- 1:10 %% 5 == 0

c(1:10)[union(which(pares), which(m.5))]
c(1:10)[intersect(which(pares), which(m.5))]
c(1:10)[which(xor(pares, m.5))]
```

Split-apply-combine

Muchos problemas en el análisis de datos pueden ser resueltos aplicando la estrategia separa, aplica y combina (SAC) en donde divides un problema en pequeños pedazos manejables, operas en forma independiente cada uno de éstos y después combinas los resultados obtenidos (Wickham 2011).

Esta estrategia se utiliza en diversas etapas del análisis de datos, por ejemplo (Wickham 2011):

- **Preparación de datos.** Cuando se crean nuevas variables según grupos, cuando se realizan ordenamientos por grupos, cuando se estandariza o normaliza variables.
- **Estadística descriptiva.** Cuando se crean agregados por grupos como sus medias o medianas.
- **Modelado**. Cuando se calculan modelos separados para cada panel en un estudio de este tipo. Estos modelos pueden examinarse por separado o unificarse para construir modelos más sofisticados que los conjuguen.

Esta estrategia se utiliza en muchas herramientas: en las tablas dinámicas de `Microsoft Excel`, el operador `group by` de `SQL`, el argumento `by` disponible en algunos procedimientos de `SAS` (Wickham 2011).

El paradigma split-apply-combine se resume en la figura 1.

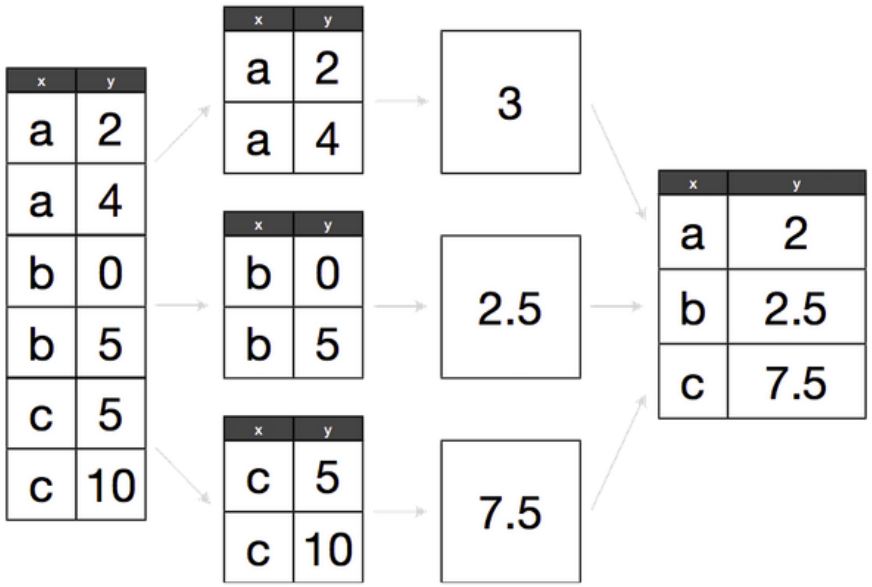


Figura 1: Ejemplificación del split-apply-combine Vaidyanathan (2014, Split-Apply-Combine).

Replicamos los vectores x y y de la figura en un *dataframe*:

```
letras <- c("a", "b", "c")
df <- data.frame(
  x = sort(letras[rep(seq(letras), 2)]),
  y = c(2, 4, 0, 5, 5, 10)
)
df
```

```
##   x  y
## 1 a  2
## 2 a  4
## 3 b  0
## 4 b  5
```

```
## 5 c  5
## 6 c 10
```

Queremos estimar la **media** de los valores en el vector y para cada tipo de letra en el vector x. Esto lo podemos hacer utilizando la estrategia **SAC**, como en la figura.

```
# Dividimos
for (l in unique(df$x) ){
  print(paste0("Grupo con letra: ", l))
  print(df[l == df$x, ])
}
```

```
## [1] "Grupo con letra: a"
##   x y
## 1 a 2
## 2 a 4
## [1] "Grupo con letra: b"
##   x y
## 3 b 0
## 4 b 5
## [1] "Grupo con letra: c"
##   x  y
## 5 c  5
## 6 c 10
```

```
# Aplicamos
for (l in unique(df$x) ){
  print(paste0("Media para valores de letra: ", l))
  print(mean(df[l == df$x, ]$y))
}
```

```
## [1] "Media para valores de letra: a"
## [1] 3
## [1] "Media para valores de letra: b"
## [1] 2.5
## [1] "Media para valores de letra: c"
## [1] 7.5
```

```
# Combinamos
medias <- list()
for (l in unique(df$x) ){
  medias[[l]] <- mean(df[l == df$x, ]$y)
}
as.data.frame(list(letras = names(medias), medias = unname(unlist(medias))))
```

```
##   letras medias
## 1      a    3.0
## 2      b    2.5
## 3      c    7.5
```

`R` tiene muchas funciones que facilitan realizar este tipo de operaciones. En particular, la familia `apply` fue pensada para realizarlas. Cada una de las funciones en esta familia recibe una estructura de datos en particular, aplica de determinada manera la función que se le especifica y combina los resultados de una forma específica.

apply

`apply` aplica una función a cada fila o columna en una matriz.

1. **Separa**: por columna o fila según se especifica en el parámetro `MARGIN` (1 para filas, 2 para columnas).
2. **Aplica**: la función que se especifica en el parámetro `FUN`.
3. **Combina**: regresa un vector con los resultados.

```
m <- matrix(c(1:5, 6:10), nrow = 5, ncol = 2)
# 1 is the row index 2 is the column index
m
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    7
## [3,]    3    8
## [4,]    4    9
## [5,]    5   10
```

```
apply(m, 1, sum)
```

```
## [1]  7  9 11 13 15
```

```
apply(m, 2, sum)
```

```
## [1] 15 40
```

Ejercicio

Haz una función que reciba un vector y devuelva la suma de la posición $v_i + v_{i+1}$. Para el n-esimo elemento, suma el primero. Aplica esa función a las columnas y filas de la matriz m.

```
# Respuesta
suma.rec <- function(v){
  resultado <- c()
  for (e in seq(length(v) - 1)){
    resultado[e] <- v[e] + v[e + 1]
  }
  resultado[length(v)] <- v[length(v)] + v[1]
  return(resultado)
}

m
apply(m, 1, suma.rec)
apply(m, 2, suma.rec)
```

lapply

`lapply` aplica una función a cada elemento en una lista. Como sabemos, un `data.frame` es únicamente un estilo particular de lista tal que todos sus elementos tienen el mismo tamaño. Por ende, también podemos utilizar `lapply` para iterar sobre las columnas de un `data.frame`.

```
lista <- list(a = 1:10, b = 2:20)
lapply(lista, mean)
```

```
## $a
## [1] 5.5
```

```
##
## $b
## [1] 11
```

```
df <- data.frame(a = 1:10, b = 11:20)
lapply(df, mean)
```

```
## $a
## [1] 5.5
##
## $b
## [1] 15.5
```

Ejercicio

El `summary` de un `data.frame` genera un resumen para los vectores que la conforman de acuerdo a la clase de la misma. Genera una función que regrese una tabla de frecuencias para factores y caracteres o una lista con media, desviación estándar para vectores numéricos o enteros. Aplícalo a la base diamonds usando lapply.

```
# Respuesta
mi.resumen <- function(vector){
  if( class(vector) == "factor" || class(vector) == "character"){
    table(vector)
  } else if ( class(vector) == "numeric" || class(vector) == "integer") {
    list(media = mean(vector), de = sqrt(var(vector)))
  }
}

lapply(names(diamonds), FUN = function(c) mi.resumen(diamonds[, c]))
```

sapply

`sapply` es otra versión de `lapply` que regresa una lista o un vector, dependiendo si se especifica el parámetro `simplify = T` y si la función aplicada regresa un único valor.

```
x <- sapply(lista, mean, simplify = F)
x
```

```
## $a
## [1] 5.5
##
## $b
## [1] 11
```

```
x <- sapply(lista, mean, simplify = T)
x
```

```
##    a    b
##  5.5 11.0
```

Ejercicio

Obtén un vector tipo caracter con los nombres de las clases de las columnas de iris.

```
# Respuesta
sapply(iris, class)
```

Ejercicio

Repite el ejercicio de la suma rara pero usa `sapply`.

Recuerda la instrucción: Haz una función que reciba un vector y devuelva la suma de la posición $v_i + v_{i+1}$. Para el n-esimo elemento, suma el primero. Utiliza `sapply` para realizar esta operacion.

```
# Respuesta
x <- 1:10
sapply(seq(x), FUN = function(i){
  if( i == length(x) ){
    x[1] + x[i]
  } else {
    x[i] + x[i + 1]
  }
})
```

mapply

`mapply` es como la versión multivariada de `sapply`. Le aplica una función a todos los elementos correspondientes de un argumento.

```
l1 <-list(a = c(1:5), b = c(6:10))
l2 <- list(c = c(11:15), d = c(16:20))
l1
```

```
## $a
## [1] 1 2 3 4 5
##
## $b
## [1]  6  7  8  9 10
```

```
l2
```

```
## $c
## [1] 11 12 13 14 15
##
## $d
## [1] 16 17 18 19 20
```

```
mapply(sum, l1$a, l1$b, l2$c, l2$d)
```

```
## [1] 34 38 42 46 50
```

```
l1[["a"]][1] + l1[["b"]][1] + l2[["c"]][1] + l2[["d"]][1]
```

```
## [1] 34
```

Ejercicio

Crea una matriz de 4 x 4 donde el primer renglón sea de unos, el segundo de dos, el tercero de 3 y el cuarto de 4. Usa `mapply` para hacerlo.

```
# Respuesta
m <- matrix(c(rep(1, 4), rep(2, 4), rep(3, 4), rep(4, 4)), nrow = 4, byrow = T)
m
```

```r
me <- t(mapply(rep, 1:4, 4))
me
```

tapply

`tapply` le aplica una función a subconjuntos de un vector.

```r
head(warpbreaks)
```

```
##   breaks wool tension
## 1     26    A       L
## 2     30    A       L
## 3     54    A       L
## 4     25    A       L
## 5     70    A       L
## 6     52    A       L
```

```r
with(warpbreaks, tapply(breaks, list(wool, tension), mean))
```

```
##          L        M        H
## A 44.55556 24.00000 24.55556
## B 28.22222 28.77778 18.77778
```

```r
tapply(warpbreaks$breaks,
       list(wool = warpbreaks$wool, tension = warpbreaks$tension),
       mean)
```

```
##     tension
## wool        L        M        H
##    A 44.55556 24.00000 24.55556
##    B 28.22222 28.77778 18.77778
```

Ejercicio

Utiliza la función `tapply` y la base de datos `diamonds` (que está dentro del paquete `ggplot2`) para obtener las medias de la variable `carat` para los grupos formados por la variable categórica `cut` y la variable categórica `color`.

```r
# Respuesta
library(ggplot2)
with(diamonds,
     tapply(carat,
            list(cut, color),
            mean))
```

by

`by` le aplica una función a subconjuntos de un `data.frame`. Se divide un data.frame según los valores de de uno o más factores. Se aplica la función FUN a cada subconjunto.

```r
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
```

```
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
by(data = iris[, 1:2], INDICES = iris[, "Species"], FUN = summary)
```

```
## iris[, "Species"]: setosa
##   Sepal.Length    Sepal.Width
##  Min.   :4.300   Min.   :2.300
##  1st Qu.:4.800   1st Qu.:3.200
##  Median :5.000   Median :3.400
##  Mean   :5.006   Mean   :3.428
##  3rd Qu.:5.200   3rd Qu.:3.675
##  Max.   :5.800   Max.   :4.400
## ------------------------------------------------------------
## iris[, "Species"]: versicolor
##   Sepal.Length    Sepal.Width
##  Min.   :4.900   Min.   :2.000
##  1st Qu.:5.600   1st Qu.:2.525
##  Median :5.900   Median :2.800
##  Mean   :5.936   Mean   :2.770
##  3rd Qu.:6.300   3rd Qu.:3.000
##  Max.   :7.000   Max.   :3.400
## ------------------------------------------------------------
## iris[, "Species"]: virginica
##   Sepal.Length    Sepal.Width
##  Min.   :4.900   Min.   :2.200
##  1st Qu.:6.225   1st Qu.:2.800
##  Median :6.500   Median :3.000
##  Mean   :6.588   Mean   :2.974
##  3rd Qu.:6.900   3rd Qu.:3.175
##  Max.   :7.900   Max.   :3.800
```

Puedo calcular, por ejemplo, la suma de los valores del largo y ancho de los sépalos en la base de datos iris según la especie.

```
res <- by(iris[, c("Sepal.Length", "Sepal.Width")], iris[, "Species"], sum)
```

Posteriormente, se pueden combinar los elementos.

```
as.data.frame(list(
  "species" = names(res),
  "suma" = sapply(seq(length(res)), FUN = function(i) res[[i]])
  ))
```

```
##      species  suma
## 1     setosa 421.7
## 2 versicolor 435.3
## 3  virginica 478.1
```

Ejercicio

Vuelve a utilizar la base de diamonds para calcular el promedio de carat según cut y color.

```
# Respuesta
library(ggplot2)
head(diamonds)
```

```
res <- by(diamonds[, c("carat")],
          list(cut = as.factor(diamonds$cut), color = as.factor(diamonds$color))
          , mean, simplify = T)
```

replicate

`replicate` es una función muy útil sobretodo en el contexto de simulación.

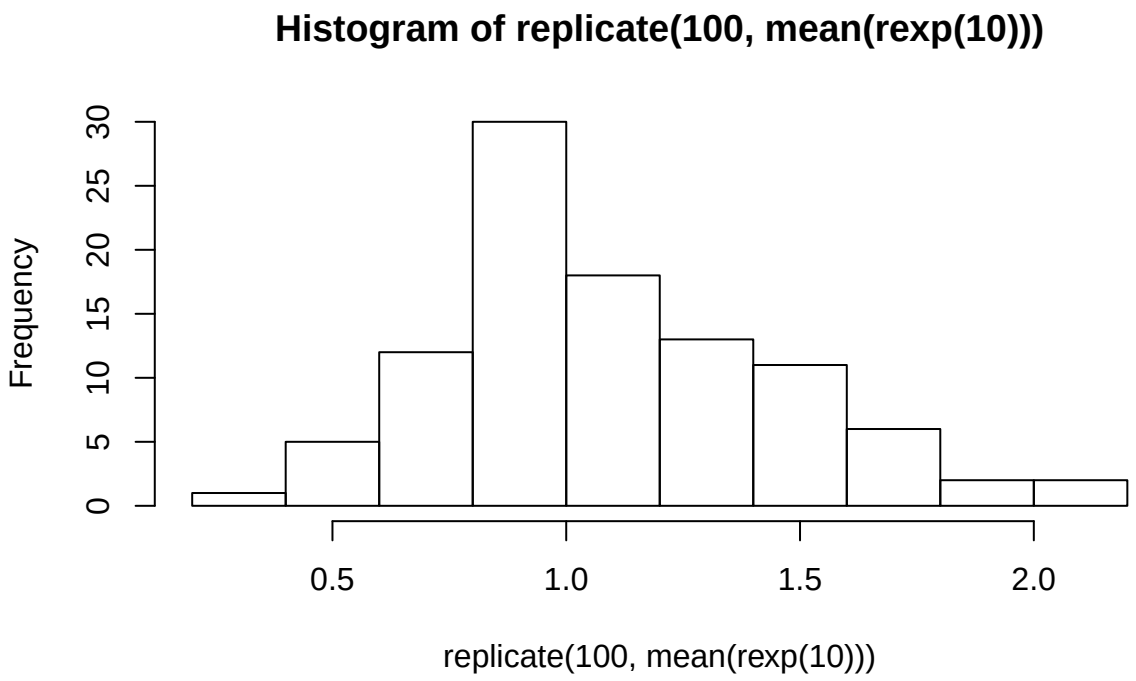
```
replicate(5, rnorm(6), simplify = F)
```

```
## [[1]]
## [1] -0.3347941 -0.5013782 -0.1745357  1.8090374 -0.2301050 -1.1304182
##
## [[2]]
## [1]  0.21598889  1.23223729  1.60935871  0.40155063 -0.27298403 -0.03615234
##
## [[3]]
## [1] -0.1503112  3.7688104 -1.6524960 -1.1351451  0.2276702 -0.1833185
##
## [[4]]
## [1] -0.41351862 -0.43759528 -0.02618435 -0.85983418  0.16654458  1.47549073
##
## [[5]]
## [1]  0.1954229  0.1594218 -0.7201933 -0.9355025  0.2854323 -0.7392351
```

```
replicate(6, rnorm(4), simplify = T)
```

```
##           [,1]       [,2]      [,3]       [,4]        [,5]       [,6]
## [1,]  0.429149  0.2459699 0.4778276 -0.3504094  0.20234597  2.1874601
## [2,]  2.733984 -0.7459894 0.7311985  1.1475974  0.11149061 -0.0709549
## [3,] -1.333403 -1.4841380 0.1721051  1.3500840 -1.52799033  1.2929966
## [4,]  1.860095  0.2220485 1.1866920  1.1161494  0.05750731  0.3768750
```

```
hist(replicate(100, mean(rexp(10))))
```



Ejercicio

Replica el ejercicio de muestras bootstrap utilizando la función `replicate`.

Recordando las instrucciones:

1. Genera un vector `x` de tamaño 1000 con realizaciones de una normal media 10, varianza 3.
2. Crea 100 muestras bootstrap del vector x.
3. Calcula la *media* para cada una de tus muestras.
4. Grafica con la función `hist()` el vector de medias de tus muestras.

```
# Respuesta
# 1
x <- rnorm(1000, mean = 10, sd = sqrt(3))
hist( # 4
  replicate(100, # 2
            mean(sample(x, size = 1000, replace = T))), # 3
  main = "Muestras bootstrap",
  xlab = "media",
  ylab = "frecuencia"
)
```

¿Puede ser más fácil?

La familia `apply` viene con R básico. Sin embargo, hay 3 implementaciones excelentes del paradigma split-apply-combine: `plyr`, `dplyr` y `data.table`.

Si la familia apply es poderosa, se queda corta comparada con estos tres. `plyr` es la primera versión de SAC de Wickham (Wickham 2011). Posteriormente, mejoró muchas de las funciones en `dplyr` (Wickham y Francois s.f.) sobretodo entorno a velocidad y facilidad de uso. `plyr` no termina de ser relevante pues varias de sus funciones aun no están en `dplyr` pero está por ser sustituido.

`data.table` (Dowle y col. 2015), es una implementación con una tradición muy diferente y tiene también funciones muy poderosas aunque con una sintaxis muy distinta a `dplyr`. Es absurdamente eficiente y tiene

múltiples aplicaciones.

Muchas de las funciones en `dplyr` también están implementadas en `data.table`. Ambos paquetes son útiles pero priorizan distintos elementos. En el capítulo siguiente se repasarán los verbos en `dplyr`.

```
sessionInfo()
```

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 15.04
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=es_MX.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=es_MX.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=es_MX.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=es_MX.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  base
##
## other attached packages:
## [1] xtable_1.8-2  rformat_0.1   rmarkdown_1.3
##
## loaded via a namespace (and not attached):
##  [1] backports_1.0.5 magrittr_1.5    rprojroot_1.2   tools_3.3.1
##  [5] htmltools_0.3.5 yaml_2.1.14     Rcpp_0.12.9     stringi_1.1.2
##  [9] knitr_1.15.1    methods_3.3.1   stringr_1.1.0   digest_0.6.11
## [13] evaluate_0.10
```