Curso básico de análisis de datos con R

Alfredo Sánchez Alberca

2021-12-29

Contents

Prefacio

Propósito de este manual

Este manual proporciona una introducción amigable al lenguaje de programación R para aquellas personas interesadas en utilizar este lenguaje para el análisis de datos.

El manual empieza con los conceptos básicos del lenguaje de programación R pero enseguida aborda su uso para la visualización y el análisis estadístico de datos, haciendo un recorrido por los test estadísticos más comunes.

Lo más interesante de este manual es la multitud de ejemplos que ilustran el uso de las técnicas estadísticas presentadas, así como los problemas propuestos.

El manual no aborda los fundamentos matemáticos de los análisis estadísticos presentados, aunque si explica brevemente cuándo deben usarse y cuándo no, así como las interpretaciones de los resultados obtenidos en los ejemplos. Si alguien está interesado en profundizar en los detalles matemáticos, puede visitar esta página.

No es un curso de programación en R, sino de uso de sus funciones predefinidas y de los paquetes más habituales para el análisis de datos.

Para cualquier comentario o sugerencia sobre este manual escriba al autor (as alber@ceu.es).

Licencia

Chapter 1

Introducción a R

La gran potencia de cómputo alcanzada por los ordenadores ha convertido a los mismos en poderosas herramientas al servicio de todas aquellas disciplinas que, como la Estadística, requieren manejar un gran volumen de datos. Actualmente, prácticamente nadie se plantea hacer un estudio estadístico serio sin la ayuda de un buen programa de análisis de datos.

R es un potente lenguaje de programación que incluye multitud de funciones para la representación y el análisis de datos. Fue desarrollado por Robert Gentleman y Ross Ihaka en la Universidad de Auckland en Nueva Zelanda, aunque actualmente es mantenido por una enorme comunidad científica en todo el mundo.

Figure 1.1: Logotipo de R

Las ventajas de R frente a otros programas habituales de análisis de datos, como pueden ser SPSS, SAS o Matlab, son múltiples:

- Es software libre y por tanto gratuito. Puede descargarse desde la web http://www.r-project.org/.
- Es multiplataforma. Existen versiones para Windows, Macintosh, Linux y otras plataformas.

- Está avalado y en constante desarrollo por una amplia comunidad científica distribuida por todo el mundo que lo utiliza como estándar para el análisis de datos.
- Cuenta con multitud de paquetes para todo tipo de análisis estadísticos y representaciones gráficas, desde los más habituales, hasta los más novedosos y sofisticados que no incluyen otros programas. Los paquetes están organizados y documentados en un repositorio CRAN (Comprehensive R Archive Network) desde donde pueden descargarse libremente.
- Es programable, lo que permite que el usuario pueda crear fácilmente sus propias funciones o paquetes para análisis de datos específicos. Existen multitud de libros, manuales y tutoriales libres que permiten su aprendizaje e ilustran el análisis estadístico de datos en distintas disciplinas científicas como las Matemáticas, la Física, la Biología, la Psicología, la Medicina, etc.

1.1 Entornos de desarrollo

Por defecto el entorno de trabajo de R es en línea de comandos, lo que significa que los cálculos y los análisis se realizan mediante comandos o instrucciones que el usuario teclea en una ventana de texto. No obstante, existen distintas interfaces gráficas de usuario que facilitan su uso, sobre todo para usuarios noveles. Algunas de ellas, como las que se enumeran a continuación, son completos entornos de desarrollo que facilitan la gestión de cualquier proyecto:

- RStudio. Probablemente el entorno de desarrollo más extendido para programar con R ya que incorpora multitud de utilidades para facilitar la programación con R.
- RKWard. Es otra otro de los entornos de desarrollo más completos que además incluye a posibilidad de añadir nuevos menús y cuadros de diálogo personalizados.
- Visual Studio Code. Es un entorno de desarrollo de propósito general ampliamente extendido. Aunque no es un entorno de desarrollo específico para R, incluye una extensión con utilidades que facilitan mucho el desarrollo con R.

Chapter 2

Tipos de datos simples

En R existen distintos tipos de datos predefinidos simples:

- `numeric`: Es el tipo de los números. Secuencia de dígitos (pueden incluir el - para negativos y el punto como separador de decimales) que representan números. Por ejemplo, `1`, `-2.0`, `3.1415` o `4.5e3`.
  Por defecto, cualquier número que se teclee tomará este tipo.
- `double`: Es el tipo de los números reales. Secuencia de dígitos que incluyen decimales separados por punto. Por ejemplo `3.1415` o `-2.0`. Son una subclase del tipo de datos numérico.
- `integer`: Es el tipo de los números enteros. Secuencia de dígitos sin separador de decimales que representan un número entero. Por ejemplo `1` o `-2`. Son una subclase del tipo de datos numérico.
- `character`: Es cualquier cadena de caracteres alfanuméricos. Secuencia de caracteres alfanuméricos que representan texto. Se escriben entre comillas simples o dobles. Por ejemplo `"Hola"` o `'Hola'`.
- `logical`: Es el tipo de los booleanos. Puede tomar cualquiera de los dos valores lógicos `TRUE` (verdadero) o `FALSE` (falso). También se pueden abreviar como `T` o `F`.
- `NA`: Se utiliza para representar datos desconocidos o perdidos. Aunque en realidad es un dato lógico, puede considerarse con un tipo de dato especial.
- `NULL`: Se utiliza para representar la ausencia de datos. La principal diferencia con `NA` es que `NULL` aparece cuando se intenta acceder a un dato que no existe, mientras que `NA` se utiliza para representar explícitamente datos perdidos en un estudio.
- `factor`: Es un tipo de dato que solo puede tomar valores de un conjunto predefinido conocido como *niveles* del factor. Los factores se suelen utilizar

para representar datos cualitativos o categóricos. Para definir un factor se utiliza la siguiente función:

  - `factor(x, levels=niveles)`: Crea un dato de tipo factor con el valor `x`. Los niveles del factor pueden indicarse mediante el parámetro `levels`, pasándole un vector con los valores posibles.

Para averiguar el tipo de un dato se puede utilizar la siguiente función:

- `class(x)`: Devuelve el tipo del dato `x`.

Ejemplo 2.1. A continuación se muestran los tipos de algunos datos.

```
class(3.1415)
#> [1] "numeric"
class(-1)
#> [1] "numeric"
class("Hola")
#> [1] "character"
class(TRUE)
#> [1] "logical"
class(NA)
#> [1] "logical"
class(NULL)
#> [1] "NULL"
class(factor('mujer', levels = c('hombre', 'mujer')))
#> [1] "factor"
```

También pueden utilizarse las siguientes funciones que devuelven un booleano:

- `is.numeric(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `numeric`.
- `is.double(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `double`.
- `is.integer(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `integer`.
- `is.character(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `character`.
- `is.logical(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `logical`.
- `is.na(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `NA`.
- `is.null(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `NULL`.
- `is.factor(x)`: Devuelve el booleano `TRUE` si `x` es del tipo `factor`.

2.1 Conversión de tipos

En muchas ocasiones es posible convertir un dato de un tipo a otro distinto. Para ello se usan las siguientes funciones:

- `as.numeric(x)`: Convierte el dato de `x` al tipo `numeric` siempre que sea posible o tenga sentido la conversión. Para convertir una cadena en un número, la cadena tiene que representar un número. El valor lógico `TRUE` se convierte en 1 y el `FALSE` en 0.

- `as.integer(x)`: Convierte el dato de `x` al tipo `integer` siempre que sea posible o tenga sentido la conversión. Para convertir una cadena en un número entero, la cadena tiene que representar un número entero. El valor lógico `TRUE` se convierte en 1 y el `FALSE` en 0.
- `as.character(x)`: Convierte el tipo de dato de `x` al tipo `character` simplemente añadiendo comillas.
- `as.logical(x)`: Convierte el tipo de dato de `x` al tipo lógico. Para datos numéricos, el 0 se convierte en `FALSE` y cualquier otro número en `TRUE`. Para cadenas se obtiene `NA` excepto para las cadenas `"TRUE"` y `"true"` que se convierten a `TRUE` y las cadenas `"FALSE"` y `"false"` que se convierten a `FALSE`.

El tipo `NA` no se puede convertir a ningún otro tipo pues representa la ausencia del dato. Lo mismo ocurre con `NULL`.

2.2 Operaciones con números

2.2.1 Operadores aritméticos

Los siguientes operadores permiten realizar las clásicas operaciones aritméticas entre datos numéricos:

- `x + y`: Devuelve la suma de `x` e `y`.
- `x - y`: Devuelve la resta de `x` e `y`.
- `x * y`: Devuelve el producto de `x` e `y`.
- `x / y`: Devuelve el cociente de `x` e `y`.

- `x %% y`: Devuelve el resto de la división entera de `x` e `y`.

- `x ^ y`: Devuelve la potencia `x` elevado a `y`.

Ejemplo 2.2. A continuación se muestran varios ejemplos de operaciones aritméticas.

```
2 + 3
#> [1] 5
5 * -2
#> [1] -10
5 / 2
#> [1] 2.5
5 %% 2
#> [1] 1
2 ^ 3
#> [1] 8
```

2.2.2 Operadores relacionales

Comparan dos números y devuelven un valor lógico.

- `x == y` : Devuelve `TRUE` si el número `x` es igual que el número `y`, y `FALSE` en caso contrario.
- `x > y` : Devuelve `TRUE` si el número `x` es mayor que el número `y`, y `FALSE` en caso contrario.
- `x < y` : Devuelve `TRUE` si el número `x` es menor que el número `y`, y `FALSE` en caso contrario.
- `x >= y` : Devuelve `TRUE` si el número `x` es mayor o igual que el número `y`, y `FALSE` en caso contrario.
- `x <= y` : Devuelve `TRUE` si el número `x` es menor o igual a que el número `y`, y `FALSE` en caso contrario.
- `x != y` : Devuelve `TRUE` si el número `x` es distinto del número `y`, y `FALSE` en caso contrario.

Ejemplo 2.3. A continuación se muestran varios ejemplos de operaciones relacionales.

```
3 == 3
#> [1] TRUE
3.1 <= 3
#> [1] FALSE
4 > 3
#> [1] TRUE
-1 != 1
#> [1] TRUE
5 %% 2
#> [1] 1
2 ^ 3
#> [1] 8
(2 + 3) ^ 2
#> [1] 25
```

2.3 Operaciones con cadenas

2.3.1 Funciones de cadenas

Existen muchas funciones para cadenas de texto pero las más comunes son:

- `nchar(c)`: Devuelve un número entero con el número de caracteres de la cadena.
- `paste(x, y, ..., sep=s)`: Concatena las cadenas `x`, `y`, etc. separándolas por la cadena `s`. Por defecto la cadena de separación es un espacio en blanco.
- `substr(c, start=i, stop=j)`: Devuelve la subcadena de la cadena `c` desde la posición `i` hasta la posición `j`. El primer carácter de una cadena

ocupa la posición 1.
- `tolower(c)`: Devuelve la cadena que resulta de convertir la cadena `c` a minúsculas.
- `toupper(c)`: Devuelve la cadena que resulta de convertir la cadena `c` a mayúsculas.

Ejemplo 2.4. A continuación se muestran varios ejemplos de operaciones con cadenas de texto.

```
nchar("Me gusta R")
#> [1] 10
paste("Me", "gusta", "R")
#> [1] "Me gusta R"
paste("Me", "gusta", "R", sep = "-")
#> [1] "Me-gusta-R"
paste("Me", "gusta", "R", sep = "")
#> [1] "MegustaR"
substr("Me gusta R", 4, 8)
#> [1] "gusta"
tolower("Me gusta R")
#> [1] "me gusta r"
toupper("Me gusta R")
#> [1] "ME GUSTA R"
```

2.3.2 Operaciones de comparación de cadenas

- `x == y` : Devuelve `TRUE` si la cadena `x` es igual que la cadena `y`, y `FALSE` en caso contrario.
- `x > y` : Devuelve `TRUE` si la cadena `x` sucede a la cadena `y`, y `FALSE` en caso contrario.
- `x < y` : Devuelve `TRUE` si la cadena `x` antecede a la cadena `y`, y `FALSE` en caso contrario.
- `x >= y` : Devuelve `TRUE` si la cadena `x` sucede o es igual a la cadena `y`, y `FALSE` en caso contrario.
- `x <= y` : Devuelve `TRUE` si la cadena `x` antecede o es igual a la cadena `y`, y `FALSE` en caso contrario.
- `x != y` : Devuelve `TRUE` si la cadena `x` es distinta de la cadena `y`, y `FALSE` en caso contrario.

Utilizan el orden alfabético, las minúsculas van antes que las mayúsculas, y los números antes que las letras.

Ejemplo 2.5. A continuación se muestran varios ejemplos de operaciones de comparación de cadenas.

```
"R" == "R"
#> [1] TRUE
"R" == "r"
```

```
#> [1] FALSE
"uno" < "dos"
#> [1] FALSE
"A" > "a"
#> [1] TRUE
"" < "R"
#> [1] TRUE
```

2.4 Operaciones con datos lógicos o booleanos

2.4.1 Operadores lógicos

A la hora de comparar valores lógicos R asocia a `TRUE` el valor 1 y a `FALSE` el valor 0.

- `x == y`: Devuelve `TRUE` si los booleanos `x` y `y` son iguales, y `FALSE` en caso contrario.
- `x < y`: Devuelve `TRUE` si el booleano `x` es menor que el booleano `y`, y `FALSE` en caso contrario.
- `x <= y`: Devuelve `TRUE` si el booleano `x` es menor o igual que el booleano `y`, y `FALSE` en caso contrario.
- `x > y`: Devuelve `TRUE` si el booleano `x` es mayor que el booleano `y`, y `FALSE` en caso contrario.
- `x >= y`: Devuelve `TRUE` si el booleano `x` es mayor o igual que el booleano `y`, y `FALSE` en caso contrario.
- `x != y`: Devuelve `TRUE` si el booleano `x` es distinto que el booleano `y`, y `FALSE` en caso contrario.
- Negación `!b`: Devuelve `TRUE` si el booleano `b` es `FALSE`, y `FALSE` si es `TRUE`.
- Conjunción `x & y`: Devuelve `TRUE` si los booleanos `x`, y `y` son `TRUE` y `FALSE` en caso contrario.
- Disyunción `x | y`: Devuelve `TRUE` si alguno de los booleanos `x` o `y` son `TRUE`, y `FALSE` en caso contrario.

Tabla de verdad

x	y	!x	x & y	x | y
FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	TRUE
TRUE	TRUE	FALSE	TRUE	TRUE

Ejemplo 2.6. A continuación se muestran varios ejemplos de operaciones lógicas con booleanos.

```
!TRUE
#> [1] FALSE
FALSE | TRUE
#> [1] TRUE
FALSE | FALSE
#> [1] FALSE
TRUE & FALSE
#> [1] FALSE
TRUE & TRUE
#> [1] TRUE
```

2.5 Variables

Una variable es un identificador ligado a algún valor.

Reglas para nombrarlas:

- Comienzan siempre por una letra o punto, seguida de otras letras, números, puntos o guiones bajos. Si empieza por punto no puede seguirle un número.
- No se pueden utilizarse palabras reservadas del lenguaje.

A diferencia de otros lenguajes de programación, las variables no tienen asociado un tipo y no es necesario declararlas antes de usarlas (tipado dinámico).

Para asignar un valor a una variable se utiliza el operador de asignación `<-`:

- `x <- y`: Asigna el valor `y` a la variable `x`.

Aunque es menos común también se puede utilizar el operador `=`.

Se puede crear una variable sin ningún valor asociado asignándole el valor `NULL`.

Una vez definida una variable, puede utilizarse en cualquier expresión donde tenga sentido el valor que tiene asociado.

Si una variable ya no va a usarse, es posible eliminarla y liberar el espacio que ocupan sus datos asociados con la siguiente función:

- `rm(x)`: Elimina la variable `x`.

Ejemplo 2.7. A continuación se muestran varios ejemplos de asignaciones de valores a variables.

```
x <- 3
x
#> [1] 3
y <- x + 2
y
#> [1] 5
# Valor no definido
```

```
x <- NULL
x
#> NULL
# Eliminar y
rm(y)
# A partir de aquí, una llamada a y produce un error.
```

2.5.1 Prioridad de los operadores

Al evaluar una expresiones R utiliza el siguiente orden de prioridad de evaluación:

1	Funciones predefinidas
2	Potencias (`^`)
3	Productos y cocientes (`*`, `/`, `%%`)
4	Sumas y restas (`+`, `-`)
5	Operadores relacionales (`==`, `>`, `<`, `>=`, `<=`, `!=`)
6	Negación (`!`)
7	Conjunción (`&`)
8	Disyunción (`|`)
9	Asignación (`<-`)

Se puede saltar el orden de evaluación utilizando paréntesis `( )`.

Ejemplo 2.8. A continuación se muestran varios ejemplos de evaluación de expresiones.

```
4 + 8 / 2 ^ 2
#> [1] 6
4 + (8 / 2) ^ 2
#> [1] 20
(4 + 8) / 2 ^ 2
#> [1] 3
(4 + 8 / 2) ^ 2
#> [1] 64
x <- 2
y <- 3
z <- ! x + 1 > y & y * 2  < x ^ 3
z
#> [1] TRUE
```

Ejercicio 2.1. Se dispone de los siguientes datos de una persona:

Variable	Valor
edad	20
estatura	165
peso	60
sexo	mujer

1. Declarar las variables anteriores y asignarles los valores correspondientes.
2. Definir la variable numérica imc con el índice de masa corporal aplicando la siguiente fórmula a las variables anteriores:

$$\text{imc} = \frac{\text{peso (kg)}}{\text{estatura (m)}^2}$$

3. Definir la variable booleana obesa con el valor correspondiente a la siguiente condición: ser mujer y no tener una edad superior a 60 y tener un índice de masa corporal mayor o igual que 30. ¿Es esta persona obesa?

Solución. Solución el ejercicio.

```
# Declaración de variables
edad <- 20
estatura <- 165
peso <- 60
sexo <- "mujer"
# Cálculo del índice de masa corporal
imc <- peso / (estatura / 100) ^ 2
imc
#> [1] 22.03857
# Cálculo de la obesidad
obesa <- sexo == "mujer" & ! edad > 60 & imc >= 30
obesa
#> [1] FALSE
```

2.6

Chapter 3

Tipos de datos estructurados

Los tipos estructurados de datos, a diferencia de los simples, son colecciones de datos con una determinada estructura. En R existen varios tipos tipos estructurados de datos que pueden clasificarse de acuerdo a su dimensión y a si son homogéneos (todos sus elementos son del mismo tipo) o heterogéneos.

Dimensiones	Homogéneos	Heterogéneos
1	Vector	Lista
2	Matriz	Data frame
n	Array	

Para averiguar la estructura de un dato estructurado se puede utilizar la función siguiente:

- `str(x)`: Devuelve una cadena de texto con la estructura de `x` en un formato amigable para los humanos.

3.1 Vectores

El vector es el tipo de dato estructurado más básicos en R. Un vector es una colección ordenada de elementos del mismo tipo.

3.1.1 Creación de vectores

Para construir un vector se utiliza la función de combinación `c()`:

- `c(x1, x2, ...)`: Devuelve el vector formado por los elementos `x1`, `x2`, etc.

También es posible utilizar el operador `:` para generar un vector de números enteros consecutivos:

- `x:y`: Devuelve el vector de números enteros consecutivos desde `x` hasta `y`.

Ejemplo 3.1. A continuación se muestran varios ejemplos de construcción de vectores.

```
c(1, 2, 3)
#> [1] 1 2 3
c("uno", "dos", "tres")
#> [1] "uno"  "dos"  "tres"
# Vector vacío
c()
#> NULL
# Vector con elementos perdidos
c(1, NA, 3)
#> [1]  1 NA  3
# Vector de números enteros consecutivos del 2 al 6
2:6
#> [1] 2 3 4 5 6
```

3.1.1.1 Vectores con nombres

Es posible asignar un nombre a cada elemento de un vector. Los nombres son etiquetas de texto que se asocian a cada elemento. Para asociar un nombre a un elemento se utiliza la sintaxis `nombre = valor`, donde `nombre` es una cadena de caracteres y `valor` es el elemento del vector.

Ejemplo 3.2. A continuación se muestra un ejemplo de creación de un vector con nombres.

```
c("Matemáticas" = 8.2, "Física" = 6.5, "Economía" = 4.5)
#> Matemáticas      Física    Economía
#>         8.2         6.5         4.5
```

Para acceder a los nombres de un vector se utiliza la siguiente función:

- `names(x)`: Devuelve un vector de cadenas de caracteres con los nombres de los elementos del vector `x`.

Ejemplo 3.3. A continuación se muestra un ejemplo de acceso a los nombres de un vector.

```
notas <- c("Matemáticas" = 8.2, "Física" = 6.5, "Economía" = 4.5)
names(notas)
#> [1] "Matemáticas" "Física"      "Economía"
```

3.1.2 Tamaño de un vector

El número de elementos de un vector es su *tamaño* y puede averiguarse con la siguiente función.

- `lenght(x)`: Devuelve el número de elementos del vector `x`.

Ejemplo 3.4. A continuación se muestran varios ejemplos de la obtención del tamaño de un vector.

```
length(c(1, 2, 3))
#> [1] 3
length(c())
#> [1] 0
```

3.1.3 Coerción de elementos

Puesto que los elementos de un vector tienen que ser del mismo tipo, cuando se crea un vector con datos de distintos tipos, la función `c()` los convertirá al mismo tipo, lo que se conoce como *coerción* de tipos. La coerción se produce de los tipos menos flexibles a los más flexibles: `logical` < `integer` < `double` < `character`.

Ejemplo 3.5. A continuación se muestran varios ejemplos de coerciones.

```
c(1, 2.5)
#> [1] 1.0 2.5
c(FALSE, TRUE, 2)
#> [1] 0 1 2
c(FALSE, TRUE, 2, "tres")
#> [1] "FALSE" "TRUE"  "2"     "tres"
```

3.1.4 Acceso a los elementos de un vector

Para acceder a los elementos de un vector se utiliza un índice. Como veremos a continuación, este índice puede ser entero, lógico o de cadena de caracteres y se indica siempre entre corchetes `[ ]` a continuación del vector.

3.1.4.1 Acceso mediante un índice entero

Los elementos de un vector están ordenados y el acceso más simple a ellos es mediante su número de orden, es decir, indicando entre corchetes el entero que corresponde a su número de orden. Se puede acceder simultáneamente a varios elementos mediante un vector con sus números de orden. Por otro lado, también es posible usar enteros negativos y en tal caso se obtendrán todos los elementos del vector excepto los que ocupan las posiciones correspondientes al valor absoluto de los índices negativos. Esta es la forma más habitual de eliminar elementos de un vector.

En R los índices enteros para acceder a los elementos de un vector comienzan en 1, a diferencia de otros lenguajes de programación que empiezan en 0.

Ejemplo 3.6. A continuación se muestran varios ejemplos de acceso a los elementos de un vector mediante índices enteros.

```
x <- c(2,4,6,8,10)
# Acceso al elemento que está en la tercera posición
x[3]
#> [1] 6
# Acceso a los elementos de las posiciones 2 y 4
x[c(2, 4)]
#> [1] 4 8
# Acceso a todos los elementos excepto el primero y el quinto
x[c(-1, -5)]
#> [1] 4 6 8
```

3.1.4.2 Acceso mediante un índice lógico

Cuando se utiliza un índice lógico, se obtienen los elementos correspondientes a las posiciones donde está el valor booleano `TRUE`.

Ejemplo 3.7. A continuación se muestran varios ejemplos de acceso a los elementos de un vector mediante índices lógicos.

```
x <- c(2,4,6,8,10)
# Acceso al elemento que está en la tercera posición
x[c(F,F,T,F,F)]
#> [1] 6
# Acceso a los elementos de las posiciones 2 y 4
x[c(F,T,F,T,F)]
#> [1] 4 8
```

Esta forma de acceder es útil cuando se genera el vector de índices mediante operadores relacionales. Cuando se aplica un operador relacional a un vector se obtiene otro vector lógico que resulta de aplicar el operador relacional a cada uno de los elementos del vector. De esta manera se puede realizar filtros para obtener los elementos de un vector que cumplen una determinada condición.

Ejemplo 3.8. A continuación se muestran varios ejemplos de acceso a los elementos de un vector mediante condiciones.

```
x <- 1:6
x %% 2 == 0
#> [1] FALSE  TRUE FALSE  TRUE FALSE  TRUE
# Filtrado de los valores pares
x[x %% 2 == 0]
#> [1] 2 4 6
# Filtrado de los valores pares menores que 5
```

```
x[x %% 2 == 0 & x < 5]
#> [1] 2 4
```

3.1.4.3 Acceso mediante un índice de cadena

Si los elementos de un vector tienen nombre, es posible acceder a ellos usando sus nombres como índices.

Ejemplo 3.9. A continuación se muestran varios ejemplos de acceso a los elementos de un vector mediante índices de cadena.

```
notas <- c("Matemáticas" = 8.2, "Física" = 6.5, "Economía" = 4.5)
notas["Física"]
#> Física
#>    6.5
notas[c("Matemáticas", "Economía")]
#> Matemáticas    Economía
#>         8.2         4.5
```

3.1.5 Pertenencia a un vector

Para comprobar si un valor en particular es un elemento de un vector se puede utilizar el operador `%in%`:

- `x %in% y`: Devuelve el booleano `TRUE` si `x` es un elemento del vector `y`, y `FALSE` en caso contrario.

Ejemplo 3.10. A continuación se muestran varios ejemplos de pertenencia de elementos a un vector.

```
x <- 1:3
2 %in% x
#> [1] TRUE
4 %in% x
#> [1] FALSE
```

3.1.6 Modificación de los elementos de un vector

Para modificar uno o varios elementos de un vector basta con acceder a esos elementos y usar el operador de asignación para asignar nuevos valores.

Ejemplo 3.11. A continuación se muestran varios ejemplos de modificación de los elementos de un vector.

```
x <- c(1, 2, 3)
x[2] <- 0
x
#> [1] 1 0 3
```

```
x[c(1, 3)] <- 1
x
#> [1] 1 0 1
```

3.1.7 Añadir elementos a un vector

Para añadir nuevos elementos a un vector pueden usarse las siguientes funciones:

- `c(x, y)`: Devuelve el vector que resulta de añadir al vector `x` los elementos del vector `y`.
- `append(x, y, pos)`: Devuelve el vector que resulta de añadir al vector `x` los elementos del vector `y`, a continuación de la posición `pos`. El parámetro `pos` es opcional y si no se indica, los elementos de `y` se añaden al final de los de `x`.

Ejemplo 3.12. A continuación se muestran varios ejemplos de añadir nuevos elementos a un vector.

```
x <- c(1, 2, 3)
y <- c(x, c(4, 5))
y
#> [1] 1 2 3 4 5
y <- append(x, c(4, 5), 2)
y
#> [1] 1 2 4 5 3
```

3.1.8 Eliminación de un vector

Para eliminar los elementos de un vector basta con asignar `NULL` a la variable que lo contiene, pero si se quiere liberar la memoria que ocupa la variable se utiliza la función `rm()`.

3.2 Listas

Las listas son colecciones ordenadas de elementos de que pueden ser de distintos tipos. Los elementos de una lista también pueden ser de tipos estructurados (vectores o listas), lo que las convierte en el tipo de dato más versátil de R. Como veremos más adelante, otras estructuras de datos como los *data frames* o los propios modelos estadísticos se construyen usando listas.

3.2.1 Creación de listas

Para construir una lista se utiliza la función `list()`:

- `list(x1, x2, ...)`: Devuelve la lista con los elementos `x1`, `x2`, etc.

Ejemplo 3.13. A continuación se muestran varios ejemplos de creación de listas.

```
list(1, "dos", TRUE)
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] "dos"
#>
#> [[3]]
#> [1] TRUE
# Lista con vectores y listas
x <- list(1, c("dos", "tres"), list(4, "cinco"))
x
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] "dos"  "tres"
#>
#> [[3]]
#> [[3]][[1]]
#> [1] 4
#>
#> [[3]][[2]]
#> [1] "cinco"
str(x)
#> List of 3
#>  $ : num 1
#>  $ : chr [1:2] "dos" "tres"
#>  $ :List of 2
#>   ..$ : num 4
#>   ..$ : chr "cinco"
# Lista vacía
list()
#> list()
```

3.2.1.1 Listas con nombres

Ejemplo 3.14. A continuación se muestra un ejemplo de creación de una lista con nombres.

```
list("nombre" = "María", "edad" = 21, "dirección" = list("calle" = "Delicias", "número" = 24, "mu
#> $nombre
#> [1] "María"
#>
```

```
#> $edad
#> [1] 21
#>
#> $dirección
#> $dirección$calle
#> [1] "Delicias"
#>
#> $dirección$número
#> [1] 24
#>
#> $dirección$municipio
#> [1] "Madrid"
```

Para obtener los nombres de una lista se utiliza la siguiente función:

- `names(x)`: Devuelve un vector de cadenas de caracteres con los nombres de los elementos de la lista `x`.

Ejemplo 3.15. A continuación se muestra un ejemplo de acceso a los nombres de una lista.

```
persona <- list("nombre" = "María", "edad" = 21, "dirección" = list("calle" = "Delicias
names(persona)
#> [1] "nombre"    "edad"      "dirección"
```

3.2.2 Tamaño de una lista

El número de elementos de una lista es su *tamaño* y puede averiguarse con la siguiente función:

- `lenght(x)`: Devuelve el número de elementos de la lista `x`.

Ejemplo 3.16. A continuación se muestran varios ejemplos la obtención del tamaño de una lista.

```
length(list(1, "dos", TRUE))
#> [1] 3
length(list(1, c("dos", "tres"), list(4, "cinco")))
#> [1] 3
length(list())
#> [1] 0
```

3.2.3 Acceso a los elementos de una lista

Se accede a los elementos de una lista de forma similar a los vectores, mediante índices enteros, lógicos o de cadena, entre corchetes `[ ]`.

3.2.3.1 Acceso mediante un índice entero

Al igual que los vectores, los elementos de una lista están ordenados y se puede utilizar un índice entero para acceder a los elementos que ocupan una determinada posición.

Ejemplo 3.17. A continuación se muestran varios ejemplos de acceso a los elementos de una lista mediante índices enteros.

```
x <- list(1, "dos", TRUE, 4.5)
# Acceso al elemento que está en la segunda posición
x[2]
#> [[1]]
#> [1] "dos"
# Acceso a los elementos de las posiciones 1 y 3
x[c(1, 3)]
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] TRUE
# Acceso a todos los elementos excepto el primero y el cuarto
x[c(-1, -4)]
#> [[1]]
#> [1] "dos"
#>
#> [[2]]
#> [1] TRUE
```

3.2.3.2 Acceso mediante un índice lógico

Cuando se utiliza un índice lógico, se obtienen los elementos correspondientes a las posiciones donde está el valor booleano `TRUE`.

Ejemplo 3.18. A continuación se muestran varios ejemplos de acceso a los elementos de una lista mediante índices lógicos.

```
x <- list(1, "dos", TRUE, 4.5)
x[c(T,F,F,T)]
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] 4.5
x < 2
#> Warning: NAs introducidos por coerción
#> [1]  TRUE    NA  TRUE FALSE
# Filtrado de valores menores que 2
```

```
x[x < 2]
#> Warning: NAs introducidos por coerción
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> NULL
#>
#> [[3]]
#> [1] TRUE
```

Obsérvese que para los elementos que no tiene sentido la comparación se obtiene `NA`, y que el acceso mediante este índice devuelve `NULL`.

3.2.3.3 Acceso mediante un índice de cadena

Si los elementos de una lista tienen nombre, se puede acceder a ellos utilizando sus nombres como índices. La única diferencia con el acceso mediante cadenas de vectores es que se obtiene siempre una lista, incluso cuando sólo se quiere acceder a un elemento. Para obtener un elemento, y no una lista con ese único elemento, se utilizan dobles corchetes `[[ ]]`.

Ejemplo 3.19. A continuación se muestran varios ejemplos de acceso a los elementos de una lista mediante índices de cadena.

```
persona <- list("nombre" = "María", "edad" = 21, "dirección" = list("calle" = "Delicias
persona[c("edad", "nombre")]
#> $edad
#> [1] 21
#>
#> $nombre
#> [1] "María"
persona["nombre"]
#> $nombre
#> [1] "María"
typeof(persona["nombre"])
#> [1] "list"
# Acceso a un único elemento
persona[["nombre"]]
#> [1] "María"
# Acceso a una lista anidada
persona[["dirección"]][["municipio"]]
#> [1] "Madrid"
```

Una alternativa a los dobles corchetes es el operador de acceso a listas `$`. Este operador además permite utilizar coincidencias parciales en los nombres de los elementos para acceder a ellos.

Ejemplo 3.20. A continuación se muestran varios ejemplos de acceso a los elementos de una lista mediante el operador `$`.

```
persona <- list("nombre" = "María", "edad" = 21, "dirección" = list("calle" = "Delicias", "número
# Acceso a un único elemento
persona$nombre
#> [1] "María"
# Acceso mediante coincidencia parcial
persona$nom
#> [1] "María"
# Acceso a una lista anidada
persona$dirección$municipio
#> [1] "Madrid"
```

3.2.4 Modificación de los elementos de una lista

Para modificar uno o varios elementos de una lista basta con acceder a esos elementos y reasignarles valores con el operador de asignación.

Ejemplo 3.21. A continuación se muestran varios ejemplos de modificación de los elementos de una lista.

```
persona <- list("nombre" = "María", "edad" = 21)
persona$edad <- 22
persona
#> $nombre
#> [1] "María"
#>
#> $edad
#> [1] 22
```

3.2.5 Añadir elementos a una lista

La forma más sencilla de añadir un elemento con nombre a una lista es indicando el nombre con el operador `$` y asignándole un valor con el operador de asignación `<-`:

- `x$nombre <- y`: Añade el elemento `y` a la lista `x` con el nombre `nombre`.

El nuevo elemento se añade siempre al final de la lista.

Para añadir elementos sin nombre o en una posición determinada se puede utilizar la función `append()`:

- `append(x, y, pos)`: Devuelve la lista vector que resulta de añadir a `x` los elementos de la lista `y`, a continuación de la posición `pos`. El parámetro `pos` es opcional y si no se indica, los elementos de `y` se añaden al final de los de `x`.

Ejemplo 3.22. A continuación se muestran varios ejemplos de añadir nuevos elementos a una lista.

```
persona <- list("nombre" = "María", "edad" = 21)
persona$email <- "maria@ceu.es"
persona
#> $nombre
#> [1] "María"
#>
#> $edad
#> [1] 21
#>
#> $email
#> [1] "maria@ceu.es"
append(persona, list("sexo" = "Mujer"), 2)
#> $nombre
#> [1] "María"
#>
#> $edad
#> [1] 21
#>
#> $sexo
#> [1] "Mujer"
#>
#> $email
#> [1] "maria@ceu.es"
```

3.2.6 Conversión de una lista en un vector

Es posible convertir una lista en un vector con la siguiente función:

- `unlist(x)`: Devuelve el vector que resulta de aplanar recursivamente la lista `x` y convertir todos los elementos al mismo tipo mediante coerción de tipos.

Ejemplo 3.23. A continuación se muestran varios ejemplos de conversión de una lista en un vector.

```
persona <- list("nombre" = "María", "edad" = 21, "dirección" = list("calle" = "Delicias
unlist(persona)
#>              nombre                edad     dirección.calle
#>             "María"                "21"           "Delicias"
#>    dirección.número dirección.municipio
#>                "24"             "Madrid"
typeof(unlist(persona))
#> [1] "character"
```

3.3 Matrices

Una matriz es una estructura de datos bidimensional de elementos del mismo tipo organizados en filas y columnas. Una matriz es similar a un vector pero contiene una atributo adicional con sus dimensiones (número de filas y número de columnas).

3.3.1 Creación de matrices

Para crear una matriz se utiliza la siguiente función:

- `matrix(x, nrow = m, ncol = n)`: Devuelve la matriz con los elementos del vector `x` organizados en `n` filas y `m` columnas. Habitualmente basta con especificar el número de filas o el número de columnas.

Ejemplo 3.24. A continuación se muestran varios ejemplos de creación de matrices.

```
matrix(1:6, nrow = 2, ncol = 3)
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
matrix(1:6, nrow = 2)
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
matrix(1:6, ncol = 3)
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
# La matriz de 1 x 1
matrix()
#>      [,1]
#> [1,]   NA
```

Como se puede observar en el ejemplo anterior, los elementos se disponen por columnas, pero se pueden disponer los elementos por filas pasando el parámetro `byrow = TRUE` a la función `matrix`.

Ejemplo 3.25. A continuación se muestran varios ejemplos de creación de matrices.

```
matrix(1:6, nrow = 2)
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
matrix(1:6, nrow = 2, byrow = TRUE)
#>      [,1] [,2] [,3]
```

```
#> [1,]    1    2    3
#> [2,]    4    5    6
```

3.3.1.1 Matrices con nombres de filas y columnas

Es posible poner nombres a las filas y a las columnas de una matriz añadiendo el parámetro `dimnames` y pasándole una lista de dos vectores de cadenas con los nombres de las filas y las columnas respectivamente.

Ejemplo 3.26. A continuación se muestran varios ejemplos de creación de matrices con nombres de filas y columnas.

```
matrix(1:6, nrow = 2, ncol = 3, dimnames = list(c("fila1", "fila2"), c("columna1", "co
#>       columna1 columna2 columna3
#> fila1        1        3        5
#> fila2        2        4        6
```

Para obtener los nombres de las filas y las columnas de una matriz se utilizan las siguientes funciones:

- `rownames(x)`: Devuelve un vector de cadenas de caracteres con los nombres de las filas de la matriz `x`.
- `colnames(x)`: Devuelve un vector de cadenas de caracteres con los nombres de las columnas de la matriz `x`.

Ejemplo 3.27. A continuación se muestran varios ejemplos de creación de matrices con nombres de filas y columnas.

```
x <- matrix(1:6, nrow = 2, ncol = 3, dimnames = list(c("fila1", "fila2"), c("columna1"
rownames(x)
#> [1] "fila1" "fila2"
colnames(x)
#> [1] "columna1" "columna2" "columna3"
```

3.3.2 Tamaño y dimensiones de una matriz

Para obtener el número de elementos y las dimensiones de una matriz se pueden utilizar las siguientes funciones:

- `length(x)`: Devuelve un entero con el número de elementos de la matriz `x`.
- `nrow(x)`: Devuelve un entero con el número de filas de la matriz `x`.
- `ncol(x)`: Devuelve un entero con el número de columnas de la matriz `x`.
- `dim(x)`: Devuelve un vector de dos enteros con el número de filas y el número de columnas de la matriz `x`.

Ejemplo 3.28. A continuación se muestran varios ejemplos de acceso a las dimensiones de una matriz.

```
x <- matrix(1:6, nrow = 2)
length(x)
#> [1] 6
nrow(x)
#> [1] 2
ncol(x)
#> [1] 3
dim(x)
#> [1] 2 3
```

Usando esta última función se pueden modificar las dimensiones de una matriz asignando un vector de dos enteros con las nuevas dimensiones. Esto también permite crear una matriz a partir de un vector.

Ejemplo 3.29. A continuación se muestran varios ejemplos de modificación de las dimensiones de una matriz.

```
x <- 1:6
dim(x) <- c(2, 3)
x
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
dim(x) <- c(3, 2)
x
#>      [,1] [,2]
#> [1,]    1    4
#> [2,]    2    5
#> [3,]    3    6
```

3.3.3 Acceso a los elementos de una matriz

Para acceder a los elementos de una matriz se utilizan dos índices (uno para las filas y otro para las columnas), separados por comas y entre corchetes `[]` a continuación de la matriz. Al igual que para los vectores, los índices pueden ser enteros, lógicos o de cadenas de caracteres.

3.3.3.1 Acceso mediante índices enteros

Para acceder a los elementos de una matriz mediante índices enteros se indica el número de fila y el número de columna del elemento entre corchetes:

- `x[i,j]`: Devuelve el elemento de la matriz `x` que está en la fila `i` y la columna `j`.

Se puede acceder a más de un elemento indicando un vector de enteros para las filas y otro para las columnas. De esta manera se obtiene una submatriz. Si no se indica la fila o la columna se obtienen todos los elementos de todas las filas o

columnas. Al igual que para vectores, se pueden utilizar enteros negativos para descartar filas o columnas

Ejemplo 3.30. A continuación se muestran varios ejemplos de acceso a los elementos de una matriz.

```
x <- matrix(1:9, nrow = 3)
x
#>      [,1] [,2] [,3]
#> [1,]    1    4    7
#> [2,]    2    5    8
#> [3,]    3    6    9
# Acceso al elemento de la segunda fila y tercera columna
x[2,3]
#> [1] 8
# Acceso a la submatriz de la primera y tercera filas, y tercera y segunda columnas
x[c(1, 3), c(3, 2)]
#>      [,1] [,2]
#> [1,]    7    4
#> [2,]    9    6
# Acceso a la primera fila
x[1, ]
#> [1] 1 4 7
# Acceso a la segunda columna
x[, 2]
#> [1] 4 5 6
# Acceso a la submatriz con todos los elementos salvo la tercera fila y la segunda col
x[-3, -2]
#>      [,1] [,2]
#> [1,]    1    7
#> [2,]    2    8
```

3.3.3.2 Acceso mediante índices lógicos

Cuando se utilizan índices lógicos, se obtienen los elementos correspondientes a las filas y columnas donde está el valor booleano `TRUE`.

Ejemplo 3.31. A continuación se muestran varios ejemplos de acceso a los elementos de una matriz.

```
x <- matrix(1:9, nrow = 3)
x
#>      [,1] [,2] [,3]
#> [1,]    1    4    7
#> [2,]    2    5    8
#> [3,]    3    6    9
# Acceso al elemento de la segunda fila y tercera columna
x[c(F, T, F), c(F, F, T)]
```

```
#> [1] 8
# Acceso a la submatriz de la primera y tercera filas, y segunda y tercera columnas
x[c(T, F, T), c(F, T, T)]
#>      [,1] [,2]
#> [1,]    4    7
#> [2,]    6    9
# Acceso a la primera fila
x[c(T, F, F), ]
#> [1] 1 4 7
# Acceso a la segunda columna
x[, c(F, T, F)]
#> [1] 4 5 6
```

3.3.3.3 Acceso mediante índices de cadena

Si las filas y las columnas de una matriz tienen nombre, es posible acceder a sus elementos usando los nombres de las filas y columnas como índices.

```
x <- matrix(1:9, nrow = 3, dimnames = list(c("f1", "f2", "f3"), c("c1", "c2", "c3")))
x
#>    c1 c2 c3
#> f1  1  4  7
#> f2  2  5  8
#> f3  3  6  9
# Acceso al elemento de la segunda fila y tercera columna
x["f2", "c3"]
#> [1] 8
# Acceso a la submatriz de la primera y tercera filas, y tercera y segunda columnas
x[c("f1", "f3"), c("c3", "c2")]
#>    c3 c2
#> f1  7  4
#> f3  9  6
```

Finalmente, es posible combinar distintos tipos de índices (enteros, lógicos o de cadena) para indicar las filas y las columnas a las que acceder.

3.4 Data frames

Chapter 4

Estructuras de control

4.1 Estructuras condicionales

4.1.1 La instrucción `if`

4.1.2 La función `ifelse()`

4.1.3 La función `switch()`

4.2 Bucles

4.2.1 Bucles iterativos (`for`)

4.2.2 Bucles condicionales

4.2.2.1 La instrucción `while`

4.2.2.2 La instrucción `repeat`

Chapter 5

Funciones

Chapter 6

Preprocesamiento de datos

Chapter 7

Gráficos y visualización de datos

7.1 Gramática de gráficos y el paquete `ggplot2`

7.2 Diagramas de puntos

7.3 Diagramas de barras

7.4 Diagramas de líneas

7.5 Histogramas

7.6 Diagramas de cajas

7.7 Diagramas de dispersión

7.8 Personalización de gráficos

7.8.1 Ejes

7.8.2 Leyendas

7.8.3 Facetas

Chapter 8

Estadística descriptiva

8.1 Tablas de frecuencias

8.1.1 Tablas de frecuencias de una variable

8.1.2 Tablas de frecuencias de dos variables (tablas de contingencia)

8.2 Estadísticos de tendencia central

8.2.1 Media

8.2.2 Mediana

8.2.3 Moda

8.3 Estadísticos de posición

8.3.1 Mínimo y máximo

8.3.2 Percentiles

8.4 Estadísticos de dispersión

8.4.1 Rango

8.4.2 Rango intercuartílico

8.4.3 Varianza y cuasivarianza

8.4.4 Desviación típica y cuasidesviación típica

8.4.5 Coeficiente de variación

8.5 Estadísticos de forma

8.5.1 Coeficiente de asimetría

8.5.2 Coeficiente de apuntamiento

8.6 Resúmenes descriptivos

Chapter 9

Estimación de parámetros y contrastes de hipótesis de variables cuantitativas

Chapter 10

Estimación de parámetros y contrastes de hipótesis de variables cualitativas

Chapter 11

Análisis de regresión simple