Programming

Javier Garcia-Bernardo (based on material by Gerko Vink)

Introduction to R and RStudio

Recap

Last week

- How to use `R`, `RStudio`, `R`-scripts and `R`-notebooks
- Data types (elements)
  - character, numeric, integer, logical, factor
- Data structures: composed of data types
  - vector, matrix, list, **data.frame**
- Subsetting data structures
- Reading files in different formats

This week

How to organize and automate your code:

First half

- Control-flow:
  - Choice: if-else statements
  - Loops: For loops
- Functions
- Environments

Second half

- Reading and writing files in serveral formats
- Principles of tidy data and short comparison of base R and the tidyverse
- Inferential statistics: A primer of linear regression
- Best practices in R

Control-flow

New controls and functions

- Choice:
  - We often want to run some code ***only if*** some ***condition*** is true.
  - `if(cond) {cons.expr} else {alt.expr}`
- Loops:
  - We often want to repeat the execution of a piece of code many times.
  - `for(var in seq) {expr}`

Loops in R often happen under the hood, using apply functions:

- `apply()`: apply a function to margins of a matrix
- `sapply()`: apply a function to elements of a list, `vector` or `matrix` return
- `lapply()`: apply a function to elements of a list, `list` return

Control-flow (I): Choice

If statement

Operation of an **if** statement:

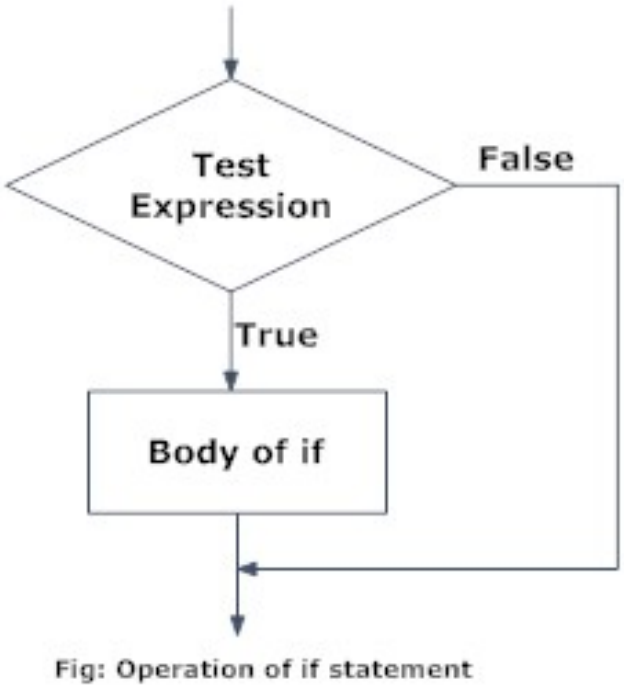


Fig: Operation of if statement

Figure 1: Source: datamentor.io

Code of an if statement:

```
value <- 3
if (value > 3) { #text expression
  print("Value greater than 3") #body of if
}
```

If-else statements

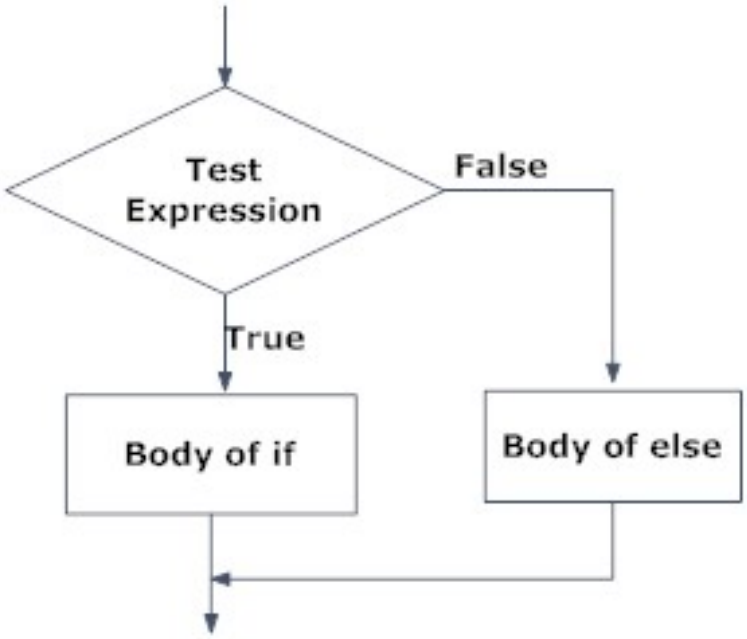


Fig: Operation of if...else statement

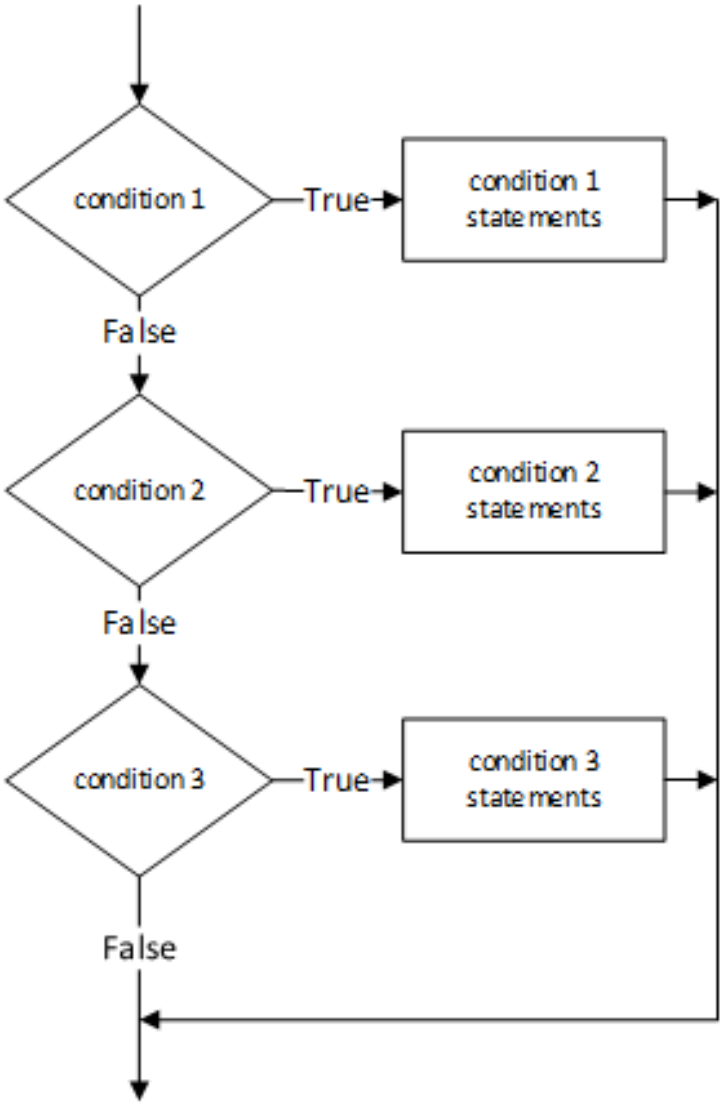
Operation of an if-else statement:

Code of an if-else statment:

```
value <- 3
if (value > 3) { #test expression
  print("Value greater than three") #body of if
} else {
    print("Value <= 3") #body of else
}
```

```
## [1] "Value <= 3"
```

If-else statements



Operation of an if-else if statement:

Code of an if-else if statment:

```
value <- 3
if (value > 3) { #condition 1
  print("Value greater than 3") #condition 1 statements
} else if (value > 1) { #condition 2
 print("Value greater than 1")  #condition 2 statements
} else if (value > 0) { #condition 3
 print("Value greater than 0")  #condition 3 statements
}
```

```
## [1] "Value greater than 1"
```

You can also add an else at the end.

Subsetting consists of if-else statements

Remember our example from last time

```
example_vector = c(1,2,3,4,5,6,7,8,9)
example_vector>3
```

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

```
example_vector[example_vector>3]
```

```
## [1] 4 5 6 7 8 9
```

The computer keeps the value of the elements of example_vector **if** the corresponding elements in the condition (`example_vector>3`) are TRUE.

Control-flow (II): Loops

For loops

For loops are used when we want to perform some repetitive calculations.

```
# Let's print the numbers 1 to 6 one by one.
print(1)
## [1] 1
print(2)
## [1] 2
print(3)
## [1] 3
print(4)
## [1] 4
print(5)
## [1] 5
print(6)
## [1] 6
```

For-loops

For-loops allow us to automate this!

For each element of `1:6`, print the element:

```
for (i in 1:6){
  print(i)
}
```

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

For-loops

You can use any variable name, `i` is a convention for counting/index.

```
for (some_var_name in 1:6){
  print(some_var_name)
}
```

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

For-loops (visually)

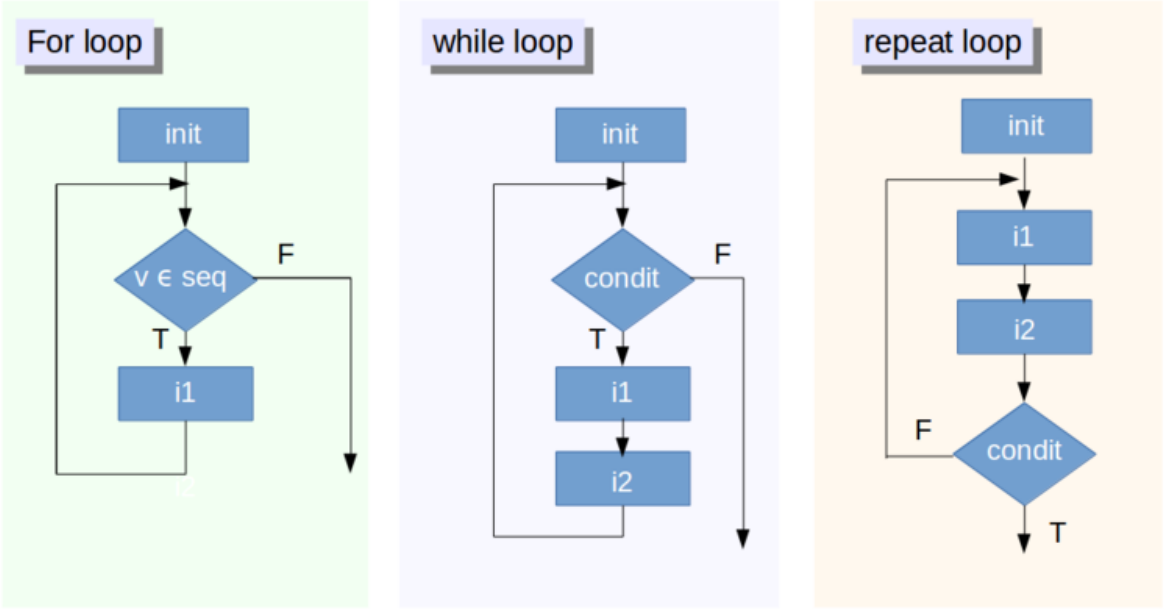


Figure 2: Source: datacamp.com

Subsetting consists of for-loops and if-else statements

```
example_vector = c(1,2,3,4,5,6,7,8,9)
example_vector>3
```

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

```
example_vector[example_vector>3]
```

```
## [1] 4 5 6 7 8 9
```

For each element in example_vector, keep the value ***if*** the corresponding element of the condition (`example_vector>3`) is TRUE

For-loops

Often you don't want to iterate over a range, but over an object

```
for (element in c("Amsterdam","Rotterdam","Eindhoven")){
  print(element)
}
```

```
## [1] "Amsterdam"
## [1] "Rotterdam"
## [1] "Eindhoven"
```

```
for (element in c("Amsterdam","Rotterdam","Eindhoven")){
  print(element)
  if (element == "Amsterdam"){
    print("Terrible football team.")
  } else {
    print("Not the prettiest city, but at least their football team is okay.")
  }
}
```

```
## [1] "Amsterdam"
## [1] "Terrible football team."
## [1] "Rotterdam"
## [1] "Not the prettiest city, but at least their football team is okay."
## [1] "Eindhoven"
## [1] "Not the prettiest city, but at least their football team is okay."
```

For-loops

Something a bit more useful

```
df <- data.frame("V1" = rnorm(5),
                 "V2" = rnorm(5, mean = 5, sd = 2),
                 "V3" = rnorm(5, mean = 6, sd = 1))

head(df)
```

```
##           V1       V2       V3
## 1  0.0513087 4.632519 6.554548
## 2 -0.7292805 4.850547 7.217293
## 3 -1.5589723 3.211326 5.692872
## 4 -0.5108245 0.999811 5.706375
## 5 -0.1332644 5.550657 6.076183
```

For-loops

Doing an operation on each column

```
for (col in names(df)) {
  print(col)
}
```

```
## [1] "V1"
## [1] "V2"
## [1] "V3"
```

```
for (col in names(df)) {
  print(col)
  print(mean(df[, col]))
}
```

```
## [1] "V1"
## [1] -0.5762066
## [1] "V2"
## [1] 3.848972
## [1] "V3"
## [1] 6.249454
```

For-loops

Doing an operation on each row

```
for (row in 1:nrow(df)) {
  row_values = df[row, ]
  print(row_values)
  print(sum(row_values>5))
}
```

```
##          V1       V2       V3
## 1 0.0513087 4.632519 6.554548
## [1] 1
##           V1       V2       V3
## 2 -0.7292805 4.850547 7.217293
## [1] 1
##          V1       V2       V3
## 3 -1.558972 3.211326 5.692872
## [1] 1
##           V1       V2       V3
## 4 -0.5108245 0.999811 5.706375
## [1] 1
##           V1       V2       V3
## 5 -0.1332644 5.550657 6.076183
## [1] 2
```

While loops

Do something forever until a condition is (not) met

```
i  = 0
while (i < 10) {
  i =  i + 1
  print(i)
}
```

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

More info on loops: https://www.datamentor.io/r-programming/break-next/

The apply() family

apply()

The apply family is a group of very useful functions that allow you to easily execute a function of your choice over a list of objects, such as a list, a data.frame, or matrix.

We will look at three examples:

- apply
- sapply
- lapply

apply()

apply is used for matrices/dataframes. It applies a function to each *row* or *column*. It returns a vector or a matrix.

```
head(df, 1)
##          V1       V2       V3
## 1 0.0513087 4.632519 6.554548
```

Apply it by row (MARGIN = 1):

```
apply(df, MARGIN = 1, mean)
## [1] 3.746125 3.779520 2.448409 2.065121 3.831192
```

Apply it by column (MARGIN = 2):

```
apply(df, MARGIN = 2, mean) #Identical to colMeans(df)
##         V1         V2         V3
## -0.5762066  3.8489719  6.2494542
```

sapply()

sapply() is used on list-objects. It returns a vector or a matrix.

```
my.list <- list(A = c(4, 2, 1), B = "Hello.", C = TRUE)
sapply(my.list, class)
```

```
##           A           B           C
##   "numeric" "character"   "logical"
```

```
my.list <- list(A = c(4, 2, 1), B = c("hello","Hello","Aa","aa"), C = c(FALSE,TRUE))
sapply(my.list, range)
```

```
##      A   B       C
## [1,] "1" "aa"    "0"
## [2,] "4" "Hello" "1"
```

Why is each element a character string?

sapply()

Any data.frame is also a list, where each column is one list-element.

This means we can use sapply on data frames as well, which is often useful.

```
sapply(df, mean)
```

```
##         V1         V2         V3
## -0.5762066  3.8489719  6.2494542
```

lapply()

`lapply()` is *exactly* the same as `sapply()`, but it returns a list instead of a vector.

```
lapply(df, class)
```

```
## $V1
## [1] "numeric"
##
## $V2
## [1] "numeric"
##
## $V3
## [1] "numeric"
```

Writing your own functions

What are functions?

Functions are reusable pieces of code that

1. take some standard input (e.g. a vector of numbers)
2. do some computation (e.g. calculate the mean)
3. return some standard output (e.g. one number with the mean)

We have been using a lot of functions: code of the form `something()` is usually a function.

```
mean(1:6)
```

```
## [1] 3.5
```

Our own function

We can make our own functions as follows:

```
squared <- function (x){
  x.square <- x * x
  return(x.square)
}

squared(4)
```

```
## [1] 16
```

`x`, the input, is called the (formal) *argument* of the function. `x.square` is called the *return value*.

Our own function

If there is no `return()`, the last line is automatically returned, so we can also just write:

```
squared <- function(x){
  x * x
}

squared(-2)
```

```
## [1] 4
```

I do not recommend this, please always specify what you return unless you have a one-line function.

Our own function

We can also combine this with `apply()`

```
df
```

```
##           V1       V2       V3
## 1  0.0513087 4.632519 6.554548
## 2 -0.7292805 4.850547 7.217293
## 3 -1.5589723 3.211326 5.692872
## 4 -0.5108245 0.999811 5.706375
## 5 -0.1332644 5.550657 6.076183
```

```
sapply(df, squared)
```

```
##               V1        V2       V3
## [1,] 0.002632583 21.460230 42.96210
## [2,] 0.531850078 23.527803 52.08931
## [3,] 2.430394733 10.312617 32.40879
## [4,] 0.260941636  0.999622 32.56272
## [5,] 0.017759406 30.809792 36.92001
```

Default options in functions

- Default options for some arguments are provided in many functions.
- They allow us to provide an additional option, but if no choice is provided, we can choose for the user of the function.

```
is_contained <- function(str_1, str_2, print_input = TRUE){
  if (print_input){
    cat("Testing if", str_1, "contained in", str_2, "\n")
  }
  return(str_1 %in% str_2)
}
```

```
is_contained("R", "rstudio")
```

```
is_contained("R", "rstudio")
## Testing if R contained in rstudio
## [1] FALSE
is_contained("R", "rstudio", print_input = TRUE)
## Testing if R contained in rstudio
## [1] FALSE
is_contained("R", "rstudio", print_input = FALSE)
## [1] FALSE
```

Troubleshooting

- Your first self-written for-loop, or function, will probably not work.
- Don't panic! Just go line-by-line, keeping track of what is currently inside each variable.
- Stackoverflow is your friend.

Scoping rules in R

Global environemnt (workspace)

When you write the name of a variable, R needs to find the value.

In the interactive computation (outside of functions, e.g., your console), this happens in the following order:

- First, search the global environment (i.e., your workspace)
- If it cannot be found, search each of the loaded packages

```
search()
```

```
##  [1] ".GlobalEnv"        "package:forcats"   "package:stringr"
##  [4] "package:dplyr"     "package:purrr"     "package:readr"
##  [7] "package:tidyr"     "package:tibble"    "package:ggplot2"
## [10] "package:tidyverse" "package:stats"     "package:graphics"
## [13] "package:grDevices" "package:utils"     "package:datasets"
## [16] "package:methods"   "Autoloads"         "package:base"
```

The order of packages is important.

Scoping rules in R: Functions

Inside a function, this happens in the following order:

- First, search within the function.
- If it cannot be found, search in the global environment (i.e., your workspace)
- If it cannot be found, search each of the loaded packages

```
y <- 3
test_t <- function() {
  print(y)
}
test_t()
## [1] 3
```

```
y <- 3
test_t <- function() {
  y <- 2
  print(y)
}
test_t()
## [1] 2
```

Scoping rules in R: Functions

What happens inside a function, stays within a function (unless you specify it differently)

```
y <- 3
test_t <- function() {
  y <- 2
  print(y)
}
test_t()
```

```
## [1] 2
```

```
y
```

```
## [1] 3
```

Scoping rules in R: Packages

Packages are neatly contained/isolated, so they are not affected by your code.They do so through namespaces:

- Namespaces allow the package developer to hide functions and data.
- Objects in the global environment that match objects in the function's namespace are ignored when running functions from packages (prevent clashes)
- Functions are executed within the namespace of the package and have access to the global environment
- They provide a way to refer to an object, with the double colon `::`

```
dplyr::n_distinct(c(1,2,3,4,2))
```

```
## [1] 4
```

Scoping rules in R: Packages (good practices)

- Pass to the function (using arguments) ***everything*** that the function needs to use (i.e. don't define something outside the function that is being used for the function)

BAD

```
numbers <- c(1,2,3)
my_mean <- function() {
   return(mean(numbers))
}
```

GOOD

```
numbers <- c(1,2,3)
my_mean <- function(numbers) {
   return(mean(numbers))
}
```

Practical