# Chapter 6: Programming in R

Mark Andrews

## Contents

# Introduction

When being introduced initially, R is often described as a programming language. This statement is technically correct, and practically very important too, but it is somewhat misleading too and so ought to be put in context. While R does have all the major programming features that we expect in any programming language — functions, conditionals, iterations, etc — and these are the reasons for R's power and extensibility, whenever we use R, especially at the beginning, we are not necessarily programming in the sense as we would be if we were using C/C++, Java, etc. Certainly, whenever we use R, we are writing code whenever and we usually put this code in scripts, which can then be executed as a batch. However, writing loops, functions, custom R objects classes, etc., is often not done until users reach a certain level of proficiency and confidence. In fact one could use R productively without engaging in programming of this kind at all. This point is meant to reassure newcomers to R that learning to use R productively does not require the same considerable initial investment that might be required when learning general purpose programming languages. We can start with learning individual R commands, build up a repertoire of widely used commands, and eventually after a certain level of familiarity and confidence has been achieved, we can start learning how to program in R.

In this chapter, we aim to provide an introduction to the major programming features of R. We'll begin with *functions* both because they can be very simple to use and because of the major role functions play in programming in R generally. We will then consider *conditionals*, which allow us to execute different blocks of code depending on whether certain conditions are true. We will then consider *iterations*, also known as *loops*. This will lead on to *functionals*, which can often take the place of traditional loops in R. As part of our consideration of functionals, we will also consider the `purrr` package that is part of the tidyverse. The aim of `purrr` is to make functionals easier and more efficient to use. We will consider object oriented programming R, which is a major programming tool for developing more complex and sophisticated software. Finally, we will consider how to create and distribute R packages.

# Functions

Functions in R, just like in any other programming language, allow us to create custom commands to perform specific calculations or carry out specific tasks. Whenever we find ourselves repeatedly using identical or similar code statements, we can create a function to execute this code. In R, functions usually, but not necessarily, take some R objects as input and always returns a new object (though this could be the `NULL` object, which signifies the absence of a defined value). We can use functions to carry out any calculations or any procedure that we could perform using any other R code.

Let's start with a numerical calculation. Let's say we have a vector of probabilities, and we want to calculate the logarithm, to a specified base, of the odds of these probabilities. In other words, let's say we have a vector of probabilities

$$p_1, p_2 \dots p_n$$

and we want to calculate

$$\log_b \left( \frac{p_i}{1 - p_i} \right) \quad \text{for } i \in 1 \dots n.$$

This calculation is simple to perform using R. For example, let's say our probabilities are as follows.

```
p <- c(0.1, 0.25, 0.5, 0.75)
```

Then, the logarithm, to base 2, of the odds of the values in `p` are calculated as follows.

```
log(p/(1-p), base = 2)
#> [1] -3.169925 -1.584963  0.000000  1.584963
```

If, on the other hand, we wanted to logarithm to base 3 or base 5 of these odds, we could do the following.

```
log(p/(1-p), base = 3)
#> [1] -2 -1  0  1
log(p/(1-p), base = 5)
#> [1] -1.3652124 -0.6826062  0.0000000  0.6826062
```

Likewise, if we had the another vector of probabilities we could calculate the log to base 2 of their odds of these values as follows.

```
q <- c(0.33, 0.67, 0.99)
log(q/(1-q), base = 2)
#> [1] -1.021695  1.021695  6.629357
```

Clearly, in these examples, we are repeating the same or similar code statements. In general, we should avoid doing this. Not only is it tedious to repeatedly type the same code, but every time we do so, we introduce the possibility, however small, of a coding error. We can therefore create a function that takes any vector of probabilities and returns the logarithm of their odds to any desired base. We can define this function as follows:

```
log_odds <- function(p, b){
  log(p/(1-p), base = b)
}
```

To understand what is being defined here, first note that all functions in R are defined by the `function` keyword. The variable names within the brackets after the `function` statement are the function's so-called input *arguments*. In this case, we have two arguments, named `p` and `b`. If the function takes no input, then we'd simply write `function()` here. The code within the `{}` is known the function's *body*. This is otherwise normal R code, but as we will see, it is running in an informationally encapsulated environment. In this example, this code is operating on two variables, `p` and `b`, and these variables are what are we will supply as inputs whenever we call the function. The value that this code in the body calculates is then what the function returns. In order to use a function in R, we usually (though not necessarily, as we will see) assign it to some name using the usual assignment operator. In this case, we assign the function to the name `log_odds`. We can now *call* this function to the run the statements above as follows.

```
log_odds(p, 2)
#> [1] -3.169925 -1.584963  0.000000  1.584963
log_odds(p, 3)
#> [1] -2 -1  0  1
log_odds(p, 5)
#> [1] -1.3652124 -0.6826062  0.0000000  0.6826062
log_odds(q, 2)
#> [1] -1.021695  1.021695  6.629357
```

Note that although the body of the `log_odds` function lies within the `{}` curly braces, when a function's body contains just a single expression, we may omit the braces. For example, we could write `log_odds` as follows.

```
log_odds <- function(p, b) log(p/(1-p), base = b)
```

The curly braces can always be used, no matter how simple the function, but in particularly simple cases, it is not uncommon for them to be omitted.

## Input arguments

Notice that in the function definition, we stated that it takes two input arguments `p` and `b`, and that the code in the body explicitly operates on `p` and `b`. As the above examples make clear, the names we use for the input argument in the function *definition* are arbitrary and do not have to correspond to the names of the variables that we use when we *call* the function. By default, whatever variable is passed in first is what the function internally defines as `p` and the second variable is what is defines as `b`. For example, consider the

following code.

```
probs <- c(0.25, 0.75, 0.9)
log_base <- 2
log_odds(probs, log_base)
#> [1] -1.584963  1.584963  3.169925
```

The function `log_odds` takes the vector `probs` and the number `log_base` and internally refers to them as `p` and `b`, respectively, and then calculates and returns `log(p/(1-p), base = b)`. As we see in the following code, we may also explicitly indicate which variables are mapped to `p` and `b`.

```
log_odds(p=probs, b=log_base)
#> [1] -1.584963  1.584963  3.169925
```

When using explicit assignment like this, the order of the arguments no longer matters. Thus, we can write the above code as follows.

```
log_odds(b=log_base, p=probs)
#> [1] -1.584963  1.584963  3.169925
```

### Default values for arguments

In the function we defined above, we had to explicitly provide both the probabilities and the base of the logarithms as input arguments. In some cases, however, we may prefer to allow some input arguments to take default values. For example, in this case, we may prefer the base of the logarithms to be 10 by default. We define default values in the function definition, as in the following example.

```
log_odds2 <- function(p, b=10){
   log(p/(1-p), base = b)
}
```

We can use this function exactly as we did with the original version, i.e. by providing two input arguments explicitly, as in the following examples.

```
log_odds2(probs, log_base)
#> [1] -1.584963  1.584963  3.169925
log_odds2(b = 3, p = probs)
#> [1] -1  1  2
```

However, if we do not include the `b` argument explicitly, then it will default to `b=10`, as in the following example.

```
log_odds2(probs)
#> [1] -0.4771213  0.4771213  0.9542425
```

### Optional arguments

A function can also have an optional arguments indicated by `...` in the arguments list in the function definition. This is often to pass arguments to functions that are called within functions. Consider the following function that calculates a polynomial function.

```
f_poly <- function(x, y, s, t){
  x^s + y^t
}
```

Let's say that we want to create a function that returns the logarithm to some specified base of `f_poly` for any given values of `x`, `y`, `s`, and `t`. We could do the following.

```
log_f_poly <- function(b=2, x, y, s, t){
  log(f_poly(x, y, s, t), base = b)
}
```

However, an easier option would be the following.

```
log_f_poly <- function(..., b=2){
  log(f_poly(...), base = b)
}
```

This will capture all the arguments, other than `b`, in the call of `log_f_poly` and pass them to `f_poly`. We can see this in action in the following example.

```
x <- c(0.5, 1.0)
y <- c(1.0, 2.0)
s <- 2
t <- 3
log_f_poly(x, y, s, t, b=2)
#> [1] 0.3219281 3.1699250
```

Note that we may obtain the optional arguments as a list given by `...` by using `list(...)` in the code body, as in the following trivial example function, where we return the optional argument list.

```
f <- function(...){
  list(...)
}
f(x=1, y=2, z=3)
#> $x
#> [1] 1
#>
#> $y
#> [1] 2
#>
#> $z
#> [1] 3
```

By using `list(...)`, we can always extract and operate upon all the information provided by optional arguments.

### Missing arguments

Consider the following simple function.

```
add_xy <- function(x, y, z){
  x + y
}
```

The function definition states that there will be three input arguments, `x`, `y`, and `z`. However, in this case, it can be used without error if we only supply `x` and `y` because `z` is not used in the code body.

```
add_xy(5, 8)
#> [1] 13
```

In this case, we say that the `z`, which is explicitly stated as an input argument, is a missing argument. Clearly, missing arguments will not necessarily raise an error but we can always test whether any given argument is missing by using the `missing` function within the code body. In the following function, we test whether each of the three input arguments are missing or not.

```
is_missing_xyz <- function(x, y, z){
  c(missing(x),
    missing(y),
    missing(z))
}

is_missing_xyz(1, 2)
#> [1] FALSE FALSE  TRUE
is_missing_xyz(z = 1, y = 42)
#> [1]  TRUE FALSE FALSE
is_missing_xyz(5, 4, 3)
#> [1] FALSE FALSE FALSE
```

As we will see in examples later throughout this book, the ability to test for missing inputs provides additional flexibility in how we can use functions.

### Function return values

The functions we have defined thus far have had single expressions in their bodies. The values of these expressions are what are returned by the functions. Functions, however, can have arbitrarily many statements and expressions in their body. When there are multiple statements, the value of the last expression is the value that is returned. To illustrate this, we can make a multi-statement version of `log_odds2` as follows.

```
log_odds3 <- function(p, b=10){
  odds <- p / (1 - p)
  log(odds, base=b)
}
log_odds3(probs)
#> [1] -0.4771213  0.4771213  0.9542425
```

A variant of the function `log_odds3` is the following.

```
log_odds4 <- function(p, b=10){
  odds <- p / (1 - p)
  log_of_odds <- log(odds, base=b)
  log_of_odds
}
log_odds4(probs)
#> [1] -0.4771213  0.4771213  0.9542425
```

In both `log_odds3` and `log_odds4`, there are multiple code statements in their bodies. In both cases, the value returned is the value of the final expression in the body. It is also possible to have an explicit `return` statement in the function's code body.

```
log_odds5 <- function(p, b=10){
  odds <- p / (1 - p)
  log_of_odds <- log(odds, base=b)
  return(log_of_odds)
}
log_odds5(probs)
#> [1] -0.4771213  0.4771213  0.9542425
```

If a `return` statement is used, whenever it is reached, the function will immediately return its value and the remainder of the code in the body, if any, is not executed. In other words, the `return` statement allows us to *break out* of a function early, which is a useful feature that we will see below after we consider conditionals. It is conventional in R to only use explicit `return` statements for this purpose, and to normally just use the

final expression in the body as the function's returned value.

### Function scope and environment

Consider the following function.

```
assign_x <- function(){
  x <- 17
  x
}
```

This function will takes nothing as input, and in the body assigns the value of 17 to the variable `x` and then returns this value. We can use it as follows.

```
x <- 42
# returns 17
assign_x()
#> [1] 17
# but the value of x is still 42
x
#> [1] 42
```

We see that the original value of `x` remains unchanged despite the fact that we have the statement `x <- 17` within the function's body. This is because the variable the `x` within the function's body is a *local* variable, or rather it is a variable defined within a *local environment* belonging to the function. Variables in the local environment are not visible outside of that environment, which in all the examples thus far is the *global environment*, which is the top level environment in the R session. In this sense, any function is informationally encapsulated: the variables defined by normal assignment operations within the function's local environment do not exist outside of the function, nor do normal assignment operations within the function's local environment affect variables outside the function. Here is another example of this phenomenon.

```
assign_x2 <- function(x){
  x[2] <- 42
  x
}

x <- c(2, 4, 8)

# The `x` within `assign_x2` is changed
assign_x2(x)
#> [1]  2 42  8
# The `x` in the global environment is unchanged
x
#> [1] 2 4 8
```

Although variables in the local environment are not "visible" in the global environment, the opposite is not true. Variables in the global environment are visible in the local environment. Consider the following example.

```
increment_x <- function(){
  x + 1
}

x <- 42
increment_x()
#> [1] 43
```

The body of `increment_x` refers to a variable `x` that is not defined in the body, nor is it passed in as an input argument. When this happens, R looks for `x` outside the local environment. In the example above, it finds it in the global environment with the value of 42. It then increments that value by 1 and returns the result. Even in this case, however, the value of `x` in the global environment remains unchanged.

```
x <- 101
increment_x()
#> [1] 102
x
#> [1] 101
```

Thus far, we have mentioned a function's local environment and contrasted this with the global environment. However, functions may be nested. In that case, we have can multiple levels of environments. Consider the following example.

```
f <- function(){

  x <- 1

  g <- function(){
    x + 1
  }

  h <- function(){
    y + 2
  }

  c(g(), h())
}
```

In this example, the functions `g` and `h` have their own local environment, but these environments are within the local environment of the function `f`, which is within the global environment. In this case, we say that the environment of `f` is the *parent* environment of `g` and `h`, and the global environment is the parent environment of `f`. This nesting of environments determines how values of variables are looked up. For example, when `g` is called it looks for `x`, which does not exist in its local environment, so it looks for it in its parent environment, which is the local environment of `f`. When `h` is called, it looks for `y`, which exists neither within its own local environment, or within its parent environment, so it must look for it in the global environment, which is its grandparent environment (the parent environment of its parent environment). We can see this function in action in the following example.

```
y <- 42
f()
#> [1]  2 44
```

An important feature of function environments is that they are defined by where the function is defined not where it is called. This becomes important when a function is returned by another function. Consider the following example.

```
f <- function(){
  y <- 42

  function(x){
    x + y
  }

}
```

```
g <- f()
```

In this example, `g`, which is the output of the call of `f`, is a function whose parent environment is the environment of `f`. As such, we can do the following.

```
y <- 21
g(17)
#> [1] 59
```

Note that the result here is `17 + 42` and not `17 + 21`. The `g` function is defined as `function(x) x + y`, and so it must look to its parent's, or grandparent's and so on, environment to find the value of `y`. Although `y` takes the value of 21 in the global scope, it takes the value of 42 in the environment of `f`, which is `g`'s parent environment.

Finally, although we mentioned that normal assignments within a local environment do not affect values in the parent, or other ancestor, environments, the special assignment operator `<<-` can be used to assign value in the parent environment. As a simple example, consider the following.

```
f <- function(){
  x <<- 42
}

x <- 17
f()
x
#> [1] 42
```

In this case, we see that the assignment of the value of 42 to `y` has been applied to the parent environment of `f`, which is the global environment.

## Anonymous functions

In all the functions above, the functions where assigned to some name. This is not necessary. Consider the following example.

```
f <- function(x, g){
  g(sum(x))
}
```

Here, `f` takes an object `x` and a function `g` and calls `g(sum(x))` and returns the result. We can pass in any existing R function we wish as the value of `g`, as in the following examples.

```
x <- c(0.1, 1.1, 2.7)
f(x, log10)
#> [1] 0.5910646
f(x, sqrt)
#> [1] 1.974842
f(x, tanh)
#> [1] 0.9991809
```

Of course, we can also pass in any function we have defined ourselves.

```
square <- function(x) x^2
f(x, square)
#> [1] 15.21
```

However, we don't have to assign a name to our custom function and pass in that name. We can instead just pass in the unnamed, or *anonymous*, function itself as in the following examples.

```
f(x, function(x) x^2)
#> [1] 15.21
f(x, function(x) log(x^2))
#> [1] 2.721953
```

Anonymous functions are widely used in R, as we will see when we discuss *functionals* later in this chapter.

We can create a *self executing anonymous function* as in the following example.

```
y <- (function(x, y, z){x + y + z})(1, 2, 3)
y
#> [1] 6
```

Here, a function is created and called immediately and the result assigned to `y`. Given that a function like this can only be invoked once, it may seem pointless. However, it does allow us to write code in an informationally encapsulated environment, where we can possibly re-use variable names from the parent or global environment, and not interfere with variables or add clutter to those environments.

## Conditionals

Conditionals allows us to execute some code based on whether some condition is true or not. Consider the following simple example.

```
library(readr)
library(tibble)

# Make a data frame
data_df <- tibble(x = rnorm(10),
                  y = rnorm(10))


write_data <- TRUE

if (write_data) {
  write_csv(data_df, 'tmp_data.csv')
}
```

Here, we write `data_df` to a `.csv` file if and only if `write_data` is true. The conditional statement begins with the keyword `if` followed by a round bracketed expression that must contain an expression that evaluates to `TRUE` or `FALSE`. This is then followed by a code block delimited by `{` and `}`. Everything in this code block is executed if and only if the expression is true. Note that if the code block contains only a single expression, just as in the case of the code body of functions, the curly braces surrounding the code block in the conditional can be omitted, as in the following example.

```
if (write_data) write_csv(data_df, 'tmp_data.csv')
```

### *if ... else* statments

The conditional statements in the examples so far will execute some code if a condition is true, and do nothing otherwise. Sometimes, however, we want to execute one code block if the condition is true and execute an alternative code block if it is false. To do this, we use an *if ... else* statement, as in the following example.

```
use_new_data <- TRUE

if (use_new_data){
  data_df <- read_csv('data_new.csv')
```

```
} else {
  data_df <- read_csv('data_old.csv')
}
```

As we can see, if `use_new_data` is true, we read in the data from `data_new.csv`, and otherwise we read the data in from `data_old.csv`.

### Nesting *if* and *if ... else* statements

We may nest *if* and *if ... else* statements. In other words, we may evaluate one condition, and if it is true, we may evaluate another condition, and so on. In the following example, if `data_1.csv` exists, we will read in its data. If it does not exist, then we test if `data_2.csv` exists. If it does, we read in its data. If not, we read in the data from `data_3.csv`.

```
if (file.exists('data_1.csv')) {
  data_df <- read_csv('data_1.csv')
} else if (file.exists('data_2.csv')) {
  data_df <- read_csv('data_2.csv')
} else {
  data_df <- read_csv('data_3.csv')
}
```

It should be noted that this example may not seem like nesting of *if ... else* statements, but rather like a chaining of these statements. However, it is exactly equivalent to the following, clearly nested, *if ... else* statements.

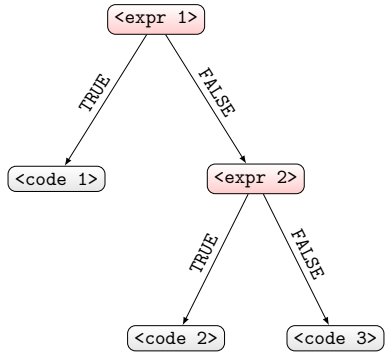
```
if (file.exists('data_1.csv')) {
  data_df <- read_csv('data_1.csv')
} else {
  if (file.exists('data_2.csv')) {
      data_df <- read_csv('data_2.csv')
  } else {
    data_df <- read_csv('data_3.csv')
  }
}
```

The only difference between these two versions is that the first version omits the (in this case, optional) `{}` after the first occurrence of `else`. As such, whether we see conditional statements like the first version as a chaining or a nesting is not meaningful.

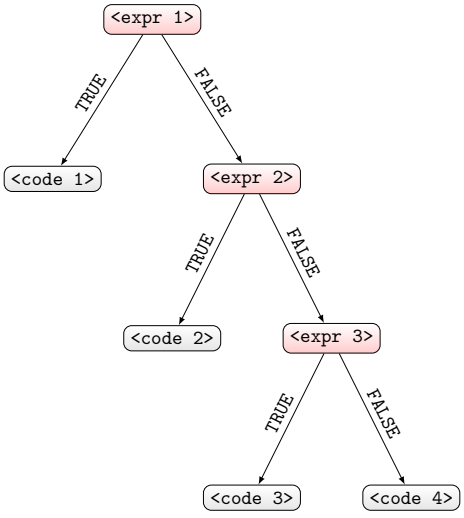
Ultimately, nested (or chained) *if* and *if ... else* statements allow us to evaluate any binary decision tree. See Figure 1 for some examples.

Nested *if* and *if ... else* statements can be used when we must choose between $K > 2$ different options. In the following example, we sample $n = 10$ random data points from one of 5 different distributions.

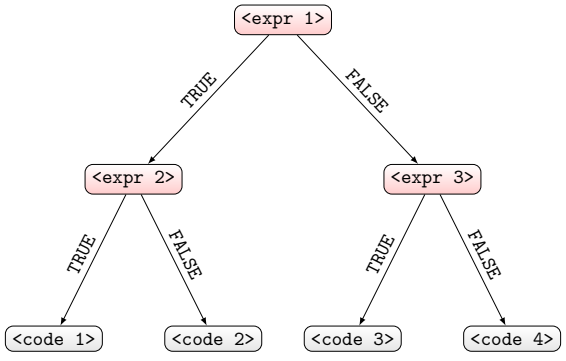
```
n <- 10
distribution <- 'student_t'

if (distribution == 'normal') {
  y <- rnorm(n, mean = 100, sd = 15)
} else if (distribution == 'log_normal') {
  y <- log(rnorm(n, mean = 100, sd = 15))
} else if (distribution == 'student_t') {
  y <- rt(n, df = 10)
} else if (distribution == 'chisq') {
  y <- rchisq(n, df = 3)
```




```
if (<expr 1>) {
  <code 1>
} else if (<expr 2>) {
  <code 2>
} else {
  <code 3>
}
```




```
if (<expr 1>) {
  <code 1>
} else if (<expr 2>) {
  <code 2>
} else if (<expr 3>) {
  <code 3>
} else {
  <code 4>
}
```




```
if (<expr 1>) {
  if (<expr 2>) {
    <code 1>
  } else {
    <code 2>
  }
 } else if (<expr 3>) {
    <code 3>
 } else {
    <code 4>
 }
```

Figure 1: Using nested *if* or *if ... else* statements, we can create any binary decision tree where each terminal node, or leaf, of the tree is a code block and every nonterminal node is an logical expression that evaluates to true or false.

```
} else if (distribution == 'uniform'){
  y <- runif(n, min = -10, 10)
}
```

### `switch` functions

While, as we've just seen, it is possible to evaluate a non-binary decisions using nested *if* and *if ... else* statements, sometimes it may be natural and simpler to use a *switch* function, which executes different expressions or code blocks depending on the value of a variable. As an example, we can re-implement the previous example of random number generation using a *switch* function.

```
distribution <- 'chisq'

y <- switch (distribution,
             normal = rnorm(n, mean = 100, sd = 15),
             log_normal = log(rnorm(n, mean = 100, sd = 15)),
             student_t = rt(n, df = 10),
             chisq = rchisq(n, df = 3),
             uniform = runif(n, min = -10, 10)
)
```

In this example, the `switch` function's first argument is the name of a variable that takes one of five different values, namely `normal`, `log_normal`, `student_t`, `chisq`, or `uniform`. Based on which value is matched, it executes the code that corresponds to that value. In this example, the value of `distribution` is `chisq` and so the `rchisq(n, df = 3)` code is executed.

The *switch* function can also be used by choosing the code based on an index rather than by matching a name. In the following example, we choose the second of the five listed options by setting `distribution <- 2`.

```
distribution <- 2

y <- switch (distribution,
             rnorm(n, mean = 100, sd = 15),
             log(rnorm(n, mean = 100, sd = 15)),
             rt(n, df = 10),
             rchisq(n, df = 3),
             runif(n, min = -10, 10)
)
```

In the above examples, the code that is executed is always a simple expression. However, it is possible to have an arbitrary code block instead, as in the following example.

```
distribution <- 'normal'

switch (distribution,
        'normal' = {
          mu <- runif(1, min=-10, max=10)
          sigma <- runif(1, min=0.01, max = 10)
          y <- rnorm(n, mean = mu, sd = sigma)
        },
        'student_t' = {
          mu <- runif(1, min=10, max=20)
          sigma <- runif(1, min=1.01, max = 3)
          y <- mu + rt(n, df=1) * sigma
        }
)
```

In this case, the entire code block corresponding to matched name of `distribution` is executed in the global namespace.

### `ifelse`, `if_else`, and `case_when`

If the code being executed by an *if ... else* statement is simple, such as a single expression, and optionally if we need to have the conditional vectorized, then we can use a `ifelse` function. In the following example, for each value of `reaction_time`, if it is less than 300, we return `'fast'` and otherwise, we return `'slow'`.

```
reaction_time <- c(1000, 300, 200, 250, 450, 300, 250, NA)
ifelse(reaction_time < 300, 'fast', 'slow')
#> [1] "slow" "slow" "fast" "fast" "slow" "slow" "fast" NA
```

As can be seen, we obtain a vector of the same length as `reaction` with values `'fast'` and `'slow'`.

The same functionality of `ifelse` can be obtained with the `if_else` function in `dplyr`.

```
library(dplyr)
if_else(reaction_time < 300, 'fast', 'slow')
#> [1] "slow" "slow" "fast" "fast" "slow" "slow" "fast" NA
```

The `if_else` is identical to `ifelse` but it requires that the two expressions that the expressions corresponding to the true and false values of the logical condition are of the same type. In addition, `if_else` provides us the option of replacing missing values with values of our choice.

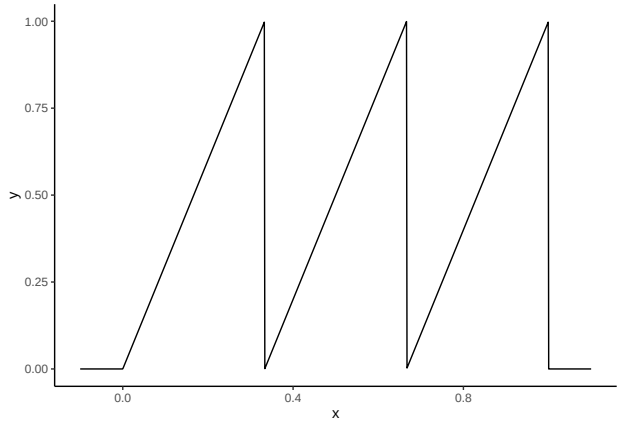



Figure 2: The ternary sawtooth map. This nonlinear function can be easily implemented using a case_when function.

The `dplyr` packages also provides the `case_when` function that can be seen as a vectorized version of nested *if ... else* statements. It is primarily intended to be used as part of `dplyr` pipelines, especially within `mutate` functions. As an example, we can use `case_when` to create a nonlinear function known as the ternary sawtooth map, which is depicted in Figure 2. First, consider how we could implement this function using nested *if ... else* statements.

```
sawtooth <- function(x){
  if (x < 0){
    0
  } else if (x < 1/3) {
    3 * x
  } else if (x < 2/3) {
    3 * x - 1
  } else if (x <= 1) {
```

```
    3 * x - 2
  } else {
    0
  }
}
```

Using `case_when`, this function could be reimplemented as follows.

```
sawtooth2 <- function(x){
  case_when(
    x < 0 ~ 0,
    x < 1/3 ~ 3 * x,
    x < 2/3 ~ 3 * x - 1,
    x <= 1 ~ 3 * x - 2,
    TRUE ~ 0
  )
}
```

Each line within `case_when` (or more precisely, each comma delimited argument), contains a *formula*, which is a statement with a `~`. To the left of `~` is a logical expression, and to the right is the code that is executed if the logical expression is true. Sometimes more than one logical expression within `case_when` is true. That is the case here. For example, if `x < 0` is true, then `x < 1/3` must be true too, and so on. The `case_when` function proceeds through the logical expressions in order and executes the code corresponding to the first if expression that is true and then stops. This means that the order of the expressions is vital to how `case_when` works. Note that the final statement in the example above is `TRUE ~ 0`. This plays the role equivalent to `else` in an `if ... else` statement. Because the expression to left of `~` obviously is always true, the code the right will be executed if and only if all other expressions are false.

The `case_when` function is more compact than nested *if ... else* statements. More usefully, it is vectorized. In other words, the `sawtooth` function, which uses nested *if ... else* statements, will not work properly if the input argument is a vector rather than a single value (which is, strictly speaking, a vector of length 1). By contrast, `sawtooth2`, which uses `case_when` will work will a vector input, as in the following example.

```
x <- c(0.1, 0.3, 0.5, 0.6, 0.7, 0.9)
sawtooth2(x)
#> [1] 0.3 0.9 0.5 0.8 0.1 0.7
```

# Iterations

In R, there are two types of iterations or *loops*, which we informally refer to as for loops and while loops.

## `for` loops

In order to understand for loops, let us re-use the `sawtooth` function above. We mentioned that this function can not be applied to vectors but only to single values. Let's say we had the following vector of 1000 elements to which we wished to apply the `sawtooth` function.

```
N <- 1000
x <- seq(-0.1, 1.1, length.out = N)
```

In principle, we could apply `sawtooth` to each element of `x`, one element at a time, as follows.

```
# Create a vector of 0's of same length as x
# This can also be done with `y <- vector('double', N)`
y <- numeric(N)
```

```
y[1] <- sawtooth(x[1])
y[2] <- sawtooth(x[2])
y[3] <- sawtooth(x[3])
...
y[N] <- sawtooth(x[N]) # where N = 1000
```

It should be obvious that we want to avoid this at all costs. Instead, we can create a for loop as follows.

```
for (i in 1:N) {
  y[i] <- sawtooth(x[i])
}
```

Essentially, this loop repeatedly executes the statement `y[i] <- sawtooth(x[i])`. On the first iteration, `i` takes the value of `1`. On the second iteration, `i` takes the value of `2`, and so on, until the final iteration where `i` takes the values of `N`. In other words, for each value of `i` from `1` to `N`, we execute `y[i] <- sawtooth(x[i])`. This is exactly equivalent to doing the following.

```
y[1] <- sawtooth(x[1])
y[2] <- sawtooth(x[2])
y[3] <- sawtooth(x[3])
...
y[N] <- sawtooth(x[N]) # where N = 1000
```

The general form of a for loop is as follows.

```
for (<var> in <sequence>) {
  <code body>
}
```

The for loop iteration begins with the `for` keyword followed by a round bracketed expression of the form `(<var> in <sequence>)` where `<var>` is what we'll call the *loop variable* and `<sequence>` is (usually) a vector or list of items. After the round brackets is some code enclosed by `{}`. For each value in the `<sequence>`, the `<var>` is set to this value and the `<code body>` is executed. We could write this in pseudo-code as follows.

```
for each value in <sequence>
  set <var> equal to this value
  execute <code body>
```

In other words, the for loop executes the `<code body>` for each value in `<sequence>`, setting `<var>` to this value on each iteration.

Let's us look at some further examples.

**Example 1**: In the following, we take a list of people's names and print a greeting to them.

```
people <- c('bill', 'hillary', 'donald', 'george')
for (person in people){
  print(paste('Hello', person))
}
#> [1] "Hello bill"
#> [1] "Hello hillary"
#> [1] "Hello donald"
#> [1] "Hello george"
```

In this example, we execute `print(paste('Hello', person))` 4 times. On the first iteration, the value of `person` takes the value of `people[1]`, which is bill. On the second iteration, `person` takes the value of `people[2]`, which is hillary, and so on.

**Example 2**: The for loop in the previous example can also be implemented as follows.

```
for (i in seq_along(people)){
  print(paste('Hello', people[i]))
}
#> [1] "Hello bill"
#> [1] "Hello hillary"
#> [1] "Hello donald"
#> [1] "Hello george"
```

Here, `seq_along(people)` gives us the sequence of integers from 1 to `length(people)`. It is safer than doing, for example, `1:length(people)` because if `people` was in fact empty, `1:length(people)` would return 1, 0, while `seq_along(people)` would return an empty vector.

**Example 3**: Here, we sum all the elements in the vector `values`.

```
values <- c(51, 45, 53, 53, 46)
s <- 0
for (value in values){
  s <- s + value
}
s
#> [1] 248
```

**Example 4**: Here, we create a cumulative sum vector for `values`.

```
cumulative_values <- numeric(length(values))
for (i in seq_along(values)){
  if (i == 1){
    cumulative_values[i] <- values[i]
  }
  else {
    cumulative_values[i] <- cumulative_values[i - 1] + values[i]
  }
}
cumulative_values
#> [1]  51  96 149 202 248
```

**Example 5**: In this example, we implement the famous chaotic dynamical system described by May (1976). In this, the value of the system at time $t$ is

$$x_t = r x_{t-1}(1 - x_{t-1}),$$

where, in its chaotic regime, $r$ takes values approximately in the range $(3.5695, 4.0)$, and the initial value of the system is $x_1 \in (0, 1)$. In this example, we set $x_1 = 0.5$ and set $r = 3.75$, and iterate for 500 iterations.

```
N <- 500
r <- 3.75
x <- numeric(N)
x[1] <- 0.5
for (t in seq(2, N)){
  x[t] <- r * x[t-1] * (1 - x[t-1])
}
```

The plot of `x` for each value of `t` is shown in Figure 3.

*Example 6*: The following example illustrates how we can easily automate repetitive tasks, and save ourselves from considerable amount of time spent on dull manuals. Let's say we have a large set of `.csv` files that we need to read in as data frames and then concatenate them into one large data frame.

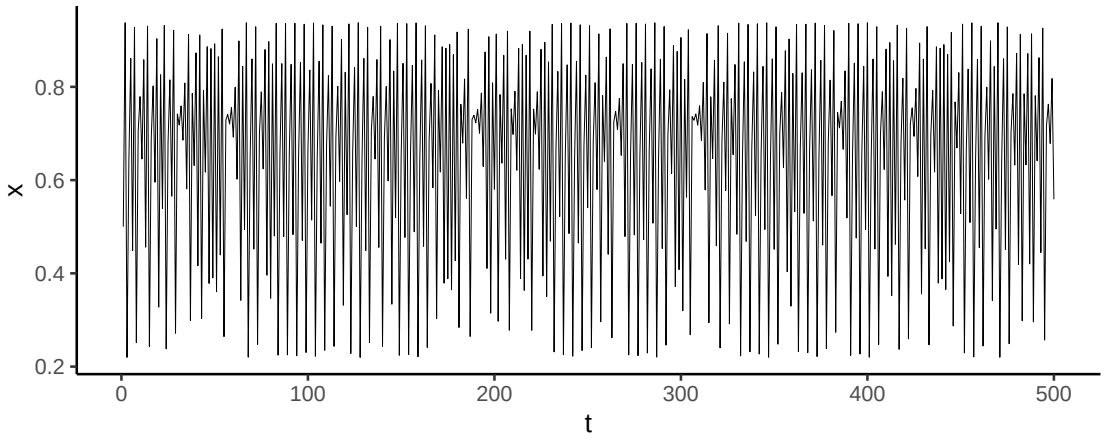



Figure 3: The time series of the chaotic logistic map. This time series can be obtained using a simple *for loop*.

```
file_names <- c('data_1.csv',
                'data_2.csv',
                'data_3.csv')
df_list <- vector('list', length(file_names))
for (i in seq_along(file_names)){
  df_list[[i]] <- read_csv(file_names[i])
}
data_df <- bind_rows(df_list)
```

In this example, there were just 3 files to read in and concatenate, but had there been hundreds or thousands of code, the required code would have been identical and the total running time might only be seconds or minutes.

### `while` loops

Unlike for loops, which iterate through each value in a sequence of values, while loops continuously execute the code body while a condition remain true. The basic The general form of a while loop is as follows.

```
while (<condition>) {
  <code body>
}
```

Here, `<condition>` is a logical expression that evaluates to `TRUE` or `FALSE`. The while loop will continue to execute `<code body>` as long as `<condition>` remains `TRUE`.

As a very simple example of a while loop, let's say we wish to find the largest value of $k$ such that $2^k \leq 10^6$.

```
k <- 0
while (2^(k+1) <= 10^6) {
  k <- k + 1
}
k
#> [1] 19
```

Here, we see that we keep incrementing `k` by 1 until `2^(k+1)` is greater than $10^6$. At that point, the condition `2^(k+1) <= 10^6` is false, and the *while loop* terminates. The final value of `k` is the highest of value of $k$ such that $2^k \leq 10^6$.

While loops are very powerful. Anything that can be implemented by a for loop can also be implemented by a while loop, but the converse is not true. An an example of a while loop implementing a for loop let's reimplement the summation for loop above.

```
values <- c(51, 45, 53, 53, 46)
i <- 1
s <- 0
while (i <= length(values)) {
  s <- s + values[i]
  i <- i + 1
}
s
#> [1] 248
```

A crucial feature of this while loop is the loop *counter* `i`, which is initialized to 1 and incremented by 1 on each iteration. This is a common feature of while loops.

Another common pattern in while loops is a conditional with a `break` statement. Whenever `break` occurs, the while loop terminates. Consider the following example where we again find the largest value of $k$ such that $2^k \leq 10^6$.

```
k <- 0
while (TRUE){

  if (2^(k + 1) > 10^6){
    break
  }

  k <- k + 1
}

k
#> [1] 19
```

Notice that in this case the while loop's condition is always `TRUE`, and the if statement with the break statement is doing the crucial work of terminating the loop.

An equivalent to a `while (TRUE)` is the `repeat` statement.

```
k <- 0
repeat {

  if (2^(k + 1) > 10^6){
    break
  }
```

```
  k <- k + 1
}
k
#> [1] 19
```

An example of a simple but nontrivial algorithm involving a while loop is where we sample from a discrete probability distribution. Consider the following probability distribution for a discrete random variable with 6 possible values.

```
p <- c(0.1, 0.2, 0.25, 0.15, 0.1, 0.2)
```

This tells us that the probability that the variable takes the value of 1 is `r p[1]`, that it takes the value of 2 is 0.2, and so on. If we would like to sample from probability distribution, we can do so using the following algorithm involving a while loop. First, we calculate the cumulative sum of `p`.

```
f <- cumsum(p)
f
#> [1] 0.10 0.30 0.55 0.70 0.80 1.00
```

Then, we sample uniformly at random from the interval $(0, 1)$, which we can do in R with `runif()`.

```
r <- runif(1)
```

Then we begin to step through each value of `f`, beginning with the first value, and test if `r` is less than or equal to this value. If it is, we stop. Otherwise, we move on to the next value in `f`.

```
k <- 1
while (TRUE) {

  if (r <= f[k]) {
    break
  }

  k <- k + 1

}
```

This sampler could also be implemented as follows.

```
k <- 1
while (r > f[k]) {
  k <- k + 1
}
```

# Functionals

Functionals are functions that take a function as input and return a vector. They play an important role in programming in R, often taking the place of for loops. There are many functionals in the base R language. Here, we will look at the most useful or widely used ones.

## `lapply`

One of the most widely used functionals in R is the `lapply` function, which takes two required arguments, a vector or list and a function, and then applies the function to each element in the vector or list and returns a new list. As an example, instead of using a for loop, we could use `lapply` to apply the `sawtooth` function used above to each element of a vector `x`. Here's the original for loop.

```
N <- 1000
x <- seq(-0.1, 1.1, length.out = N)
y <- numeric(N)
for (i in 1:N) {
  y[i] <- sawtooth(x[i])
}
```

We can replace the preassignment of `y`, i.e. `y <- numeric(N)`, and the for loop entirely using `lapply` as follows.

```
y <- lapply(x, sawtooth)
```

The `lapply` function returns a list with as many elements as there are elements in `x`.

```
length(x)
#> [1] 1000
length(y)
#> [1] 1000
class(y)
#> [1] "list"
head(y, 3) # first three elements of y
#> [[1]]
#> [1] 0
#>
#> [[2]]
#> [1] 0
#>
#> [[3]]
#> [1] 0
```

Should we prefer that `y` be a vector rather than a list, we can `unlist` it.

```
y <- unlist(y)
head(y, 3)
#> [1] 0 0 0
```

There are other options too for returning a vector from a functional, as we will see shortly.

As another application of `lapply`, consider the for loop example above where we read in data frames from multiple files using `read_csv` and stored them in a list that was then concatenated into `data_df`.

```
file_names <- c('data_1.csv',
                'data_2.csv',
                'data_3.csv')
df_list <- vector('list', length(file_names))
for (i in seq_along(file_names)){
  df_list[[i]] <- read_csv(file_names[i])
}
data_df <- bind_rows(df_list)
```

The preassignment to `df_list` and the the for loop can be replaced with the use of `lapply`.

```
df_list <- lapply(file_names, read_csv)
```

The general form of the `lapply` function is as follows.

```
returned_list <- lapply(<vector or list>, <function>)
```

Sometimes, we may need to use a function that takes arguments in `lapply`. As an example, let us imagine

that we wish to calculate the trimmed mean of each vector, and also ignoring missing values, in a list of vectors. The trimmed mean, which we will also consider in later chapters of this book, is where we compute the mean after a certain proportion of the high and low elements have been trimmed. The trimmed mean of a vector, where we remove 5% of values on the upper and lower extremes of the vector, and also ignoring missing values, is as follows.

```
x <- c(10, 20, NA, 125, 35, 15)
mean(x, trim = 0.05, na.rm = T)
#> [1] 41
```

To use the `trim = 0.05` and `na.rm = T` arguments when we use `lapply`, we supply them as optional arguments after the function name as follows.

```
data_vectors <- list(x = c(NA, 10, 11, 12, 1001, -20),
                     y = c(5, 10, 7, 2500, 6),
                     z = c(2, 4, 1000, 8, 5)
)
lapply(data_vectors, mean, trim = 0.20, na.rm = T) %>%
  unlist()
#>         x         y         z
#> 11.000000  7.666667  5.666667
```

Note that because data frames (and tibbles) are essentially lists of vectors, `lapply` can be easily used to apply a function to all columns of a data frame.

```
data_df <- tibble(x = rnorm(10),
                  y = rnorm(10),
                  z = rnorm(10))
lapply(data_df, mean) %>%
  unlist()
#>           x           y           z
#>  0.24504010 -0.43967492 -0.05399682
```

The above function accomplishes the same thing as the following `summarise_all` function, except that `summarize_all` returns a tibble.

```
summarise_all(data_df, mean)
#> # A tibble: 1 x 3
#>       x      y       z
#>   <dbl>  <dbl>   <dbl>
#> 1 0.245 -0.440 -0.0540
```

### sapply and vapply

As we've seen, `lapply` always returns a list. Sometimes, this returned list is a list of single values or list of vectors of the same length and type. In these cases, it would be preferable to convert these lists to vectors or matrices. We saw this in the case of using `sawtooth` with `lapply` above, where we manually converted the returned list into a vector using `unlist`. Variants of `lapply`, `sapply` and `vapply`, can facilitate doing these conversions. The `sapply` function works like `lapply` but will attempt to simplify the list as a vector or a matrix if possible. In the following example, we use `sapply` to apply `sawtooth` to each element of `x`.

```
N <- 1000
x <- seq(-0.1, 1.1, length.out = N)
y <- sapply(x, sawtooth)
head(y, 5)
#> [1] 0 0 0 0 0
```

Here, because the list that would have been returned by `lapply`, had we used it here, is a list of length `N` of

single numeric values (or numeric vectors of length 1), `sapply` can produce and return a numeric vector of length `N`.

If the results of `lapply` is a list of length `N` of numeric vectors of length 3, for example, then this list could be simplified to a matrix. In the following example, the list returned by `lapply` is a list like this.

```
data_df <- tibble(x = rnorm(100),
                  y = rnorm(100),
                  z = rnorm(100))
lapply(data_df, quantile)
#> $x
#>          0%         25%         50%         75%        100%
#> -2.31932737 -0.70960382 -0.05431694  0.72976278  2.13348636
#>
#> $y
#>          0%         25%         50%         75%        100%
#> -2.82300012 -0.48106048  0.04061442  0.61636766  2.13286973
#>
#> $z
#>         0%        25%        50%        75%       100%
#> -2.4662529 -0.6690109 -0.1111398  0.5306826  2.1873352
```

If we use `sapply` here instead, the result is a matrix.

```
sapply(data_df, quantile)
#>                 x           y          z
#> 0%   -2.31932737 -2.82300012 -2.4662529
#> 25%  -0.70960382 -0.48106048 -0.6690109
#> 50%  -0.05431694  0.04061442 -0.1111398
#> 75%   0.72976278  0.61636766  0.5306826
#> 100%  2.13348636  2.13286973  2.1873352
```

In cases where the list returned by `lapply` has elements of different lengths, `sapply` can not simplify it.

```
X <- list(x = rnorm(2),
          y = rnorm(3),
          z = rnorm(4)
)
sapply(X, function(x) x^2)
#> $x
#> [1] 0.0503656 0.1913169
#>
#> $y
#> [1] 1.7629691 0.2194715 1.7954867
#>
#> $z
#> [1] 2.7736662 0.2904671 0.9987699 0.1166150
```

The `vapply` function is a safer version of `sapply` because it specifies the nature of the returned values of each application of the function. For example, we know the default returned value of `quantile` will be a numeric vector of length 5. We can specify this as the `FUN.VALUE` argument to `vapply`.

```
vapply(data_df, quantile, FUN.VALUE=numeric(5))
#>                 x           y          z
#> 0%   -2.31932737 -2.82300012 -2.4662529
#> 25%  -0.70960382 -0.48106048 -0.6690109
#> 50%  -0.05431694  0.04061442 -0.1111398
```

```
#> 75%   0.72976278  0.61636766  0.5306826
#> 100%  2.13348636  2.13286973  2.1873352
```

### mapply and Map

The functionals `lapply`, `sapply`, `vapply` take a function and apply it to each element in a vector or list. In other words, each element in the vector or list is supplied as the argument to the function. While we may, as we say, have other arguments to the function set to fixed values for each function application, we can not use `lapply`, `sapply` or `vapply` to apply functions to two or more lists or vectors at the same time. Consider the following function.

```
power <- function(x, k) x^k
```

If we had a vector of `x` values, and set `k` to e.g. 3, we could do the following.

```
x <- c(2, 3, 4, 5)
sapply(x, power, k=5)
#> [1]   32  243 1024 3125
```

However, if we had the vector of `x` values and a vector of `k` values, we could do the following.

```
x <- c(2, 3, 4, 5)
k <- c(2, 3, 2, 2)
mapply(power, x = x, k = k)
#> [1]  4 27 16 25
```

As we can see, for each element `i` of `x` and `k`, we calculate `power(x[i], k[i])`. This `mapply` is therefore equivalent to the following for loop.

```
for (i in seq_along(x)){
  power(x[i], k[i])
}
```

With `mapply`, we can iterate over any number of lists of input arguments simultaneously. As an example, the random number generator `rnorm` function takes 3 arguments: `n`, `mean`, and `sd`. In the following, we apply `rnorm` to each value of of lists of these three arguments.

```
set.seed(101)
n <- c(2, 3, 5)
mu <- c(10, 100, 200)
sigma <- c(1, 10, 10)
mapply(rnorm, n = n, mean = mu, sd = sigma)
#> [[1]]
#> [1]  9.673964 10.552462
#>
#> [[2]]
#> [1]  93.25056 102.14359 103.10769
#>
#> [[3]]
#> [1] 211.7397 206.1879 198.8727 209.1703 197.7674
```

As we can see, we effectively execute `rnorm(n=3, mean=10, sd=1)`, `rnorm(n=5, mean=5, sd = 10)`, and so on.

The `Map` function works just like `mapply`, with minor differences, such as not ever simplifying the results.

```
set.seed(101)
Map(rnorm, n = n, mean = mu, sd = sigma)
```

```
#> [[1]]
#> [1]  9.673964 10.552462
#>
#> [[2]]
#> [1]  93.25056 102.14359 103.10769
#>
#> [[3]]
#> [1] 211.7397 206.1879 198.8727 209.1703 197.7674
```

As we can see, `Map` and `mapply` are identical in their usage and in what they do. However, by default, `mapply` will attempt to simply its output like `sapply`, if possible. We saw this with the use of the `power` function with `mapply` above. However, if we replace `mapply` with `Map`, no simplification is applied, and we obtain a list as output.

```
x <- c(2, 3, 4, 5)
k <- c(2, 3, 2, 2)
Map(power, x = x, k = k)
#> [[1]]
#> [1] 4
#>
#> [[2]]
#> [1] 27
#>
#> [[3]]
#> [1] 16
#>
#> [[4]]
#> [1] 25
```

### Filter, Find, and Position

The `Filter` functional takes a *predicate*, which is a function that returns a logical value, and a vector or list and returns those elements of the list for which the predicate is true. As an example, here we have a data frame with three variables.

```
data_df <- tibble(x = rnorm(3),
                  y = rnorm(3),
                  z = c('a', 'b', 'c')
)
```

We can select the numeric vectors of `data_df` as follows.

```
Filter(is.numeric, data_df)
#> # A tibble: 3 x 2
#>       x      y
#>   <dbl>  <dbl>
#> 1  0.526 -1.47
#> 2 -0.795 -0.237
#> 3  1.43  -0.193
```

In this example, we are doing what could otherwise be accomplished with `dplyr`'s `select_if` function.

```
select_if(data_df, is.numeric)
#> # A tibble: 3 x 2
#>       x      y
#>   <dbl>  <dbl>
```

```
#> 1  0.526 -1.47
#> 2 -0.795 -0.237
#> 3  1.43  -0.193
```

Unlike `select_if`, `Filter` can be applied to data structure other than data frames. For example, in the following example, we select all elements of a random sample of 20 integers that are multiples of 3.

```
s <- sample.int(100, size=10, replace=T)
s
#>  [1] 31 79 51 14 67 42 50 43 14 25
Filter(function (x) x %% 3 == 0, s)
#> [1] 51 42
```

The `Find` function searches through a vector or list to find the first element for which a predicate is true.

```
Find(function (x) x %% 3 == 0, s)
#> [1] 51
```

If the predicate is not true of any element, then `NULL` is returned.

```
Find(function (x) x < 0, s)
#> NULL
```

The `Position` function is like `Find`, but returns the position of the first element for which the predicate is true.

```
Position(function (x) x %% 3 == 0, s)
#> [1] 3
```

## Functionals with purrr

The `purrr` package in the `tidyverse` provides functionals like those just covered, but which have consistent with one another in terms of how they are used, and also with how other `tidyverse` functions are used. In addition, `purrr` provides additional functional tools beyond those in base R.

### map

One of the main tools in `purrr` is `map` and its variants. It is very similar to `lapply`. It takes a list (or vector) and a function, and applies the function to each element of the list. To re-use an example from above, in the following, we apply `read_csv` to each file name in a list of file names, and collect the data frames that are produced in a list `data_df_list`.

```
library(purrr)
file_list <- c('data_1.csv',
               'data_2.csv',
               'data_3.csv')
data_df_list <- map(file_list, read_csv)
```

As with `lapply`, we can supply arguments for the function being applied as optional arguments. For example, if we wanted to read in the data files as in the previous example, but just read no more than 100 rows of data, supply the `n_max` argument to `read_csv` as follows.

```
data_df_list <- map(file_list, read_csv, n_max=100)
```

The `map` function have `_if` and `_at` variants, which function similarly to the `_if` and `_at` variants of the `dplyr` verbs we met earlier. For example, if we want to apply `read_csv` only if the name of the data file is not `data_2.csv`, we can use `map_if` as follows.

```
data_df_list <- map_if(file_list,
                       function(x) x != 'data_2.csv',
                       read_csv)
```

We could accomplish the same thing by using `map_at`, which can take positional arguments, or negative positions, to include or exclude elements of a list. For example, if we use `-2` as the second argument to `map_at` it will skip the second item in `file_list`.

```
data_df_list <- map_at(file_list,
                       -2,
                       read_csv)
```

When the list that is returned by `map` or its variants can be simplified to a vector, we can use the `map_dbl`, `map_int`, `map_lgl`, `map_chr` variants of `map` to simplify the list to a vector of doubles, integers, Booleans, or characters, respectively.

```
data_df <- read_csv('data_1.csv')

map_int(data_df, length)
#>    x    y    z
#> 1000 1000 1000
map_lgl(data_df, is_integer)
#>     x     y     z
#> FALSE FALSE FALSE
map_dbl(data_df, mean)
#>              x              y              z
#> -0.0259532006  0.0643665838 -0.0009872201
map_chr(data_df, class)
#>         x         y         z
#> "numeric" "numeric" "numeric"
```

There is also a `map_df` that can be used when the list contains data frames that can be concatenated together. For example, to read in all three data frames and concatenate them we can do the following.

```
data_df_all <- map_df(file_list, read_csv)
```

What is accomplished with `map_df` here is exactly what would be obtained by a combination of `map` and `dplyr`'s `bind_rows`.

```
data_df_all_2 <- map(file_list, read_csv) %>%
  bind_rows()
all_equal(data_df_all, data_df_all_2)
#> [1] TRUE
```

The use of `map_df` in the above example is exactly equivalent to using `map_dfr`, which forces the creation of the data frame by row binding.

```
data_df_all_3 <- map_dfr(file_list, read_csv)
all_equal(data_df_all, data_df_all_3)
#> [1] TRUE
```

On the other hand, if we wanted to create a data frame by column binding, we could use `map_dfc`. It is possible to do this with the data frames in the previous examples because they all have equal numbers of rows.

```
data_df_all_4 <- map_dfc(file_list, read_csv)
```

The resulting data frame `data_df_all_4` is of the following dimensions.

```
dim(data_df_all_4)
#> [1] 1000    9
```

Its column names are as follows.

```
names(data_df_all_4)
#> [1] "x"  "y"  "z"  "x1" "y1" "z1" "x2" "y2" "z2"
```

### purrr style anonymous functions

We saw above that we can use anonymous functions in `lapply`, `sapply`, etc., functionals. This can be done in `purrr` functionals like the `map` family too, as we see in the following example.

```
map_dbl(data_df,
        function(x) mean(log(abs(x)))
)
#>          x          y          z
#> -0.6400092 -0.5868030 -0.5765645
```

However, `purrr` provides *syntactic sugar* to allow us to rewrite this as follows.

```
map_dbl(data_df,
        ~ mean(log(abs(.)))
)
#>          x          y          z
#> -0.6400092 -0.5868030 -0.5765645
```

In other words, in place of the `function(x)`, we have `~`, and in place of the anonymous function's input variable we have `.`.

### map2 and pmap

When we have two or more than two sets of input arguments, we can use `map2` and `pmap`, respectively. Both of these functions also have the `_lgl`, `_int`, `_dbl`, `_chr`, `_df`, `_dfr`, `_dfc` variants that we saw with `map`.

As an example of a `map2` function, we'll use the `power` function that two input arguments.

```
x <- c(2, 3, 4, 5)
k <- c(2, 3, 2, 2)
map2_dbl(x, k, power)
#> [1]  4 27 16 25
```

As an example of a `pmap` function, we can reimplement the `rnorm` based sampler that we originally wrote with `mapply`. The first argument to `pmap` is a list whose length is the number of arguments being passed to the function. If want to iterate over different values of the `n`, `mean` and `sd` arguments for `rnorm`, as we did in the example above, we'd set up a list like the following.

```
args <- list(n = c(2, 3, 5),
             mean = c(10, 100, 200),
             sd = c(1, 10, 10))
```

We then use `pmap` as follows.

```
pmap(args, rnorm)
#> [[1]]
#> [1] 8.734939 9.313147
#>
#> [[2]]
```

```
#> [1]  95.54338 112.24082 103.59814
#>
#> [[3]]
#> [1] 204.0077 201.1068 194.4416 217.8691 204.9785
```

### walk

The `walk` function in `purrr` is like `map` but is used with functions that are called just for their side effects. Put informally, a side effect of a function is an effect performed by a function other than its return value. A very common side effect of a function is writing something to a file. For example, while the `write_csv` function does return a value, it is simply the value of name of the file it is written. The real action done by `write_csv` is in its side effect. Like `map`, `walk` has has variants `walk2` and `pwalk` that are exactly analogous to `map2` and `pmap`.

As an example, let us perform the inverse of `map(file_list, read_csv)`, which took a list of `.csv` files, read in their data as data frames, and then stored them in the list `data_df_list`. The inverse should take each element in the list and write it back to the appropriately named file. For these, we can use `walk2`, which will iterate over the data frames in the list and the names in `file_list`.

```
walk2(data_df_list, file_list, write_csv)
```

them in he output of that function was a list of data frames.

### keep and discard

The `purrr` package also provides us with workalike to base R's `Filter`. For example, if we want to select out elements of a list or vector for which some predicate is true, we can use `keep`. To select those elements for which the predicate is not true, we can use `discard`. As an example, here we select out the even numbers in a seq of integers from 1 to 20.

```
keep(seq(20), ~ . %% 2 == 0)
#>  [1]  2  4  6  8 10 12 14 16 18 20
```

To select the odds numbers, which are obviously not the even numbers, we can use `discard` as follows.

```
discard(seq(20), ~ . %% 2 == 0)
#>  [1]  1  3  5  7  9 11 13 15 17 19
```

## Handling errors & raising errors

In any code, however well written, there is always a chance of an error occurring. An error indicates that something has gone wrong, or something unexpected has happened, and this has forced the program to stop. As an example, consider the following.

```
x <- c('1', '2', '3')
sum(x)
#> Error in sum(x): invalid 'type' (character) of argument
```

In this example, the vector `x` that we supplied to `sum` is a character vector and the sum of the elements of a character vector is not defined. As we see, the error message that was printed out tells us what went wrong.

Sometimes errors arise because of programming mistakes. For example, we could use the wrong variable, function, or file name in our code. Other times, we might make essentially typographic errors in our code, missing a comma or period where it is required, or using a comma with a period is required or vice versa. Problems like this are coding errors that we hope we will catch immediately as soon as we try to run the code. Other errors arise through no fault of our own, but rather because of some exceptional circumstances. As

an example, if we were processing large number of data files, like in some of the examples we've seen so far, some data files might be simply ill-formed or empty. What we'd like in situations like this is to essentially automatically deal with the problems. It might be possible to automatically correct the problems, or try some alternatives that might works. Failing this, we might like to just skip over the problems and move off. Many programming languages, R included, provide us tools to allow us to automatically handle exceptions or expected errors.

### try

One simple tool for exception handling is `try`. This allows us to try some code, but it is does not work, we skip over it with exiting the program. To explore `try`, let us create a small problem to deal with. We'll create a list of vectors and function that we'd like apply to each vector in the list.

```
vctr_list <- list(x = c(1, 2, 3),
                  y = c(4, 5, 6),
                  z = c('1', '2', '3')
)

log_abs_sum <- function(x){
  log(abs(sum(x)))
}
```

The function `log_abs_sum` will not work if its input vector is a character vector. As such, the following code, which uses `purrr`'s `map` functional, will fail.

```
map(vctr_list, log_abs_sum)
#> Error in sum(x): invalid 'type' (character) of argument
```

We can rewrite `log_abs_sum` using a `try` statement as follows.

```
log_abs_sum_i <- function(x){
  try(
    log(abs(sum(x)))
  )
}
```

Now, the `map` functional will work.

```
map(vctr_list, log_abs_sum_i)
#> Error in sum(x) : invalid 'type' (character) of argument
#> $x
#> [1] 1.791759
#>
#> $y
#> [1] 2.70805
#>
#> $z
#> [1] "Error in sum(x) : invalid 'type' (character) of argument\n"
#> attr(,"class")
#> [1] "try-error"
#> attr(,"condition")
#> <simpleError in sum(x): invalid 'type' (character) of argument>
```

We see, of course, that errors are still raised and that resulting list produced by the `map` contains information related to the errors that occurred with `z`. We could suppress the error messages entirely by setting `silent = TRUE` inside the `try` statement.

### tryCatch

More generally useful than `try` is `tryCatch` that calls another function to handle an error if it arises.

```
log_abs_sum_ii <- function(x){
  tryCatch(
    log(abs(sum(x))),
    error = function(e){
      NA
    }
  )
}
```

In this version of `log_abs_sum`, whenever any error occurs the function corresponding to `error` inside `tryCatch` is called. This function takes one argument, which we name `e` here. This is the error condition object and contains all the information relating to the error that was raised. In the example above, whenever an error arises, we call a function that simply returns `NA`. When used with `map`, the result is as follows.

```
map(vctr_list, log_abs_sum_ii)
#> $x
#> [1] 1.791759
#>
#> $y
#> [1] 2.70805
#>
#> $z
#> [1] NA
```

Should we wish to do so, we may filter out, using `purrr`'s `discard`, the elements that lead to errors.

```
map(vctr_list, log_abs_sum_ii) %>%
  discard(is.na)
#> $x
#> [1] 1.791759
#>
#> $y
#> [1] 2.70805
```

On the other hand, we could also use `map_dbl` to return a vector with `NA` values.

```
map_dbl(vctr_list, log_abs_sum_ii)
#>        x        y        z
#> 1.791759 2.708050       NA
```

### purrr's safely and possibly

The `purrr` package provides us another means to handle errors by wrapping functions with `safely`. The wrapped functions produces a list with two elements, one being the result of the function, which will be `NULL` if an error occurred, and the other element being the error, which is `NULL` if no error occurred. As an example, let's wrap `log_abs_sum` with `safely`.

```
log_abs_sum_iii <- safely(log_abs_sum)
```

If we use `log_abs_sum_iii` on an appropriate input, we get the following output.

```
log_abs_sum_iii(c(1, 2, 3))
#> $result
#> [1] 1.791759
#>
```

```
#> $error
#> NULL
```

By contrast, if we use it on an inappropriate input, we get the output like the following.

```
log_abs_sum_iii(c('1', '2', '3'))
#> $result
#> NULL
#>
#> $error
#> <simpleError in sum(x): invalid 'type' (character) of argument>
```

If we use a `safely` wrapped function with `map`, we'd get a list of lists. Each element of the list is a list with a `result` and `error` element.

```
map(vctr_list, log_abs_sum_iii)
#> $x
#> $x$result
#> [1] 1.791759
#>
#> $x$error
#> NULL
#>
#>
#> $y
#> $y$result
#> [1] 2.70805
#>
#> $y$error
#> NULL
#>
#>
#> $z
#> $z$result
#> NULL
#>
#> $z$error
#> <simpleError in sum(x): invalid 'type' (character) of argument>
```

We can use `purrr`'s `transpose` function to turn this list "inside out". In this case, it would return a list of two lists, one with the results for each element of `vctr_list`, and one with the errors.

```
map(vctr_list, log_abs_sum_iii) %>%
  transpose()
#> $result
#> $result$x
#> [1] 1.791759
#>
#> $result$y
#> [1] 2.70805
#>
#> $result$z
#> NULL
#>
#>
#> $error
```

```
#> $error$x
#> NULL
#>
#> $error$y
#> NULL
#>
#> $error$z
#> <simpleError in sum(x): invalid 'type' (character) of argument>
```

With `magrittr`'s `extract2`, we may extract the `result` element and discard the `NULL` results.

```
library(magrittr)
map(vctr_list, log_abs_sum_iii) %>%
  transpose() %>%
  extract2('result') %>%
  discard(is.null)
#> $x
#> [1] 1.791759
#>
#> $y
#> [1] 2.70805
```

A related option is to use `purrr`'s `possibly` function where we supply a default value in case of `error`. In the following example, we produce `NA` whenever an error occurs.

```
log_abs_sum_iv <- possibly(log_abs_sum, NA)
```

This essentially allows us to accomplish what `log_abs_sum_ii` using `tryCatch` did.

```
map_dbl(vctr_list, log_abs_sum_iv)
#>        x        y        z
#> 1.791759 2.708050       NA
```

## Raising errors

In the above examples, the errors were raised by the `sum` function when it was called with a non numerical input vector. We can create our own functions that raise errors easily using the `stop` or `stopifnot` functions. As an example, if want to create a function `f` that takes exactly one argument and for it to be a scalar number, i.e. a numerical vector of length one, we can check these conditions are met and raise a custom error message otherwise.

```
f <- function(x){
  if (!is.numeric(x)){
    stop('The input must be numerical.')
  }
  if (length(x) > 1){
    stop('The input must be a scalar.')
  }

  x
}
```

This will raise different errors with different inappropriate input arguments.

```
f(101)  # this one works
#> [1] 101
f('42') # not numerical
```

```
#> Error in f("42"): The input must be numerical.
f(c(1, 2, 3)) # vector needs to be length 1
#> Error in f(c(1, 2, 3)): The input must be a scalar.
```

We can also use the shorter `stopifnot` function. With this, we list all the conditions that we require to be true. If any are false, an error is raised.

```
f <- function(x){
  stopifnot(is.numeric(x),
            length(x) == 1)
  x
}

f(101)
#> [1] 101
f('42')
#> Error in f("42"): is.numeric(x) is not TRUE
f(c(1, 2, 3))
#> Error in f(c(1, 2, 3)): length(x) == 1 is not TRUE
```

As a more realistic example involving `tryCatch`, let us consider a situation where we want to read in and process data frames from many `.csv` files and then concatenate them

## Object oriented programming

Object oriented programming (OOP) is *programming paradigm*. It is used extensively in some of the world's most widely used general purpose programming languages such as Java, C++, and Python. Although there are different approaches to OOP, the most widely recognized approach, which is sometimes referred to as *classical* OOP, or *encapsulated* OOP, and which is the approach used in Java, C++, and Python, and many other major programming languages, is based on the creation of *classes* that have internal variables and functions, known as *methods*, and operate on these variables. All objects are instances of these classes. As an example, if we were to create an object oriented statistical regression program, we would create a regression model class whose attributes might include the set of predictor variables, the outcome variables, the parameters of the statistical model, etc. Its methods might include functions to infer the parameters' values from data, to do model testing and evaluations, to do prediction etc. For any given regression analysis, we'd begin by initiating an object that is an instance of the regression class. This object would have its own internal state, which would include the specific values of its predictor and outcome variables, etc. We would then execute this object's methods to do things like infer its regression coefficients, make predictions, etc.

In contrast to in Java, C++, and Python, OOP in R is, to be blunt, a bit of a mess. For starters, there are at least 4 styles of OOP in R: S3, S4, RC, and R6. S3 (based on OOP in the S language, version 3) is the most widely used, but is quite basic, relatively informal, and is likely to be very unfamiliar to those used to classical OOP. S4 (based on OOP in the S language, version 4) is a more strict and formal version of S3. Both S3 and S4 are based on the concept of *generic functions*, as opposed to classical OOP. RC (*Reference classes*) is an approach to OOP in R that was released with R version 2.12 (in 2010). It is strict and formal and designed to be a classical OOP approach. R6 is a OOP R package, rather than a built in feature of the R language. It is intended to be a replacement to RC that is simpler, faster, and more lightweight, but like RC, it is also a classical OOP approach. R6 plays a major role in the Shiny web application framework. There is no unanimous advice about which R approach to OOP to follow. Programmers who are used to classical OOP from other languages will naturally gravitate towards to RC or R6. However, the general advice of some in the R community, see Wickham (2019), is to not use this as the default approach to OOP in R as it may lead to very non-idiomatic R code. Instead, S3 is recommended as the default because it is so extensively used throughout R. Others recommend S4 as the default.

In this chapter, we provide just brief introduction to OOP in R. For a more in depth treatment, we recommend

Wickham (2019). Here, we will provide an introduction to just S3 classes and R6 classes. We chose to concentrate on these two because S3 is definitely the most widely used OOP approach, and knowledge of it is essential to understand much of the R code base. Having an understanding of S3 classes can help with the understanding of S4 classes, as the latter can be seen as a more formalized version of the former. We look at R6 because it is both simple and classical (in the sense of classical OOP mentioned above). Understanding R6 allows one to pursue this familiar approach to OOP in R, and RC based OOP can be seen as based on similar principles to R6.

## S3 classes and generic functions

The two key concepts to S3 OOP are *classes* and *generic functions*. S3 classes are particularly simple: they are simply labels that are attributes of objects. As an example, we could create an arbitrary list in R as follows.

```
my_list <- list(x = 1,
                y = 2,
                z = c('a', 'b', 'c'))
```

We can create a new class type `my_list_class` and make `my_list` an instance of `my_list_class` simply as follows.

```
class(my_list) <- 'my_list_class'
```

We could also accomplish the same thing using the `structure` function as follows.

```
my_list <- structure(list(x = 1,
                          y = 2,
                          z = c('a', 'b', 'c')),
                     class='my_list_class')
```

Just by printing `my_list`, we can see it has a `class` attribute whose value is `my_list_class`.

```
my_list
#> $x
#> [1] 1
#>
#> $y
#> [1] 2
#>
#> $z
#> [1] "a" "b" "c"
#>
#> attr(,"class")
#> [1] "my_list_class"
```

We can also see this using the `attr` function.

```
attr(my_list, 'class')
#> [1] "my_list_class"
```

Using some tools from the `sloop` package, we can verify the `my_list` is an instance of an S3 class of type `my_list_class`.

```
library(sloop)
otype(my_list)
#> [1] "S3"
s3_class(my_list)
#> [1] "my_list_class"
```

The next key concept of S3 OOP is that of *generic functions*. Generic functions are very widely used in R. In

brief, a generic function calls another function in its place based on the class type of its input argument. We've already met generic functions. For example, the widely used `summary` function in R is a generic function. When you call `summary(x)`, R looks up the class of `x` and then calls another function based on this class to do the summarization.

In order to explore generic functions, in the following code, we define three familiar objects that are all instances of S3 classes.

```
letters_var <- factor(sample(letters,
                     size = 100,
                     replace = T)
)

data_df_1 <- data.frame(x = rnorm(5),
                        y = rnorm(5))

data_df_2 <- tibble(x = rnorm(5),
                    y = rnorm(5))
```

We can view their class types using the `s3_class`.

```
s3_class(letters_var)
#> [1] "factor"
s3_class(data_df_1)
#> [1] "data.frame"
s3_class(data_df_2)
#> [1] "tbl_df"     "tbl"        "data.frame"
```

As we can see, `letters_var` is an instance of a `factor` class and `data_df_1` is an instance of a `data.frame` class. On the other hand, `data_df_2` is an instance of a `tbl_df` class, which is a sub-class of a `tbl`, which is sub-class of `data.frame`.

When we call `summary()` on these three objects, we get the following.

```
summary(letters_var)
#> a b c d e f g h i j k l m n o p q r s t u v w x y z
#> 1 3 4 1 3 3 5 5 3 5 3 5 3 7 5 3 5 3 3 3 5 8 4 1 7 2
summary(data_df_1)
#>        x                   y
#>  Min.   :-0.78273   Min.   :-0.9009
#>  1st Qu.:-0.77900   1st Qu.:-0.7685
#>  Median :-0.37480   Median :-0.6259
#>  Mean   :-0.43428   Mean   :-0.2662
#>  3rd Qu.:-0.31939   3rd Qu.: 0.3003
#>  Max.   : 0.08454   Max.   : 0.6637
summary(data_df_2)
#>        x                   y
#>  Min.   :-0.62795   Min.   :-0.46356
#>  1st Qu.:-0.48892   1st Qu.:-0.08399
#>  Median :-0.04692   Median : 0.16262
#>  Mean   :-0.17651   Mean   : 0.24257
#>  3rd Qu.: 0.07486   3rd Qu.: 0.30546
#>  Max.   : 0.20637   Max.   : 1.29231
```

What has happened here is that the `summary` function looks up the class of its input and tries to find an appropriate summary function. For example, the `letters_var` is an instance of a `factor` class, so `summary` looks for `summary.factor` and uses that. If, in fact, the `summary.factor` function did not exist, `summary`

would call the `summary.default` function. We can verify that this is happening behind the scenes by using the `sloop`'s `s3_dispatch` function.

```
s3_dispatch(summary(letters_var))
#> => summary.factor
#>  * summary.default
```

The `=>` indicates which function is called, which is `summary.factor`. The function below it, `summary.default`, is defined, which is indicated by `*`, but not called. Note that we can manually call `summary.factor(letters_var)` and this must return the same result as `summary(letters_var)`.

```
summary.factor(letters_var)
#> a b c d e f g h i j k l m n o p q r s t u v w x y z
#> 1 3 4 1 3 3 5 5 3 5 3 5 3 7 5 3 5 3 3 3 5 8 4 1 7 2
```

When we call `summary(data_df_1)`, the following happens.

```
s3_dispatch(summary(data_df_1))
#> => summary.data.frame
#>  * summary.default
```

We see here that `summary(data_df_1)` essentially calls `summary.data.frame(data_df_1)`, which we could have called manually as follows.

```
summary.data.frame(data_df_1)
#>        x                    y
#>  Min.   :-0.78273   Min.   :-0.9009
#>  1st Qu.:-0.77900   1st Qu.:-0.7685
#>  Median :-0.37480   Median :-0.6259
#>  Mean   :-0.43428   Mean   :-0.2662
#>  3rd Qu.:-0.31939   3rd Qu.: 0.3003
#>  Max.   : 0.08454   Max.   : 0.6637
```

A more interesting case is with `summary(data_df_2)`.

```
s3_dispatch(summary(data_df_2))
#>    summary.tbl_df
#>    summary.tbl
#> => summary.data.frame
#>  * summary.default
```

We see here that R looked first looked for `summary.tbl_df`, but this is not defined. It then looked for `summary.tbl`, but that too is not defined. It then looked for `summary.data.frame`, which does exist.

Given that we've "created" a class `my_list_class` of which `my_list` is an instance, we can now make a `summary.my_list_class` function that will be called when we call `summary(my_list)`. It is entirely up to us what `summary.my_list_class` does. Therefore, let us make it simply return the length of each element in the list.

```
summary.my_list_class <- function(the_list){
  lapply(the_list, length)
}
```

Now, when we call `summary(my_list)`, we get the following.

```
summary(my_list)
#> $x
#> [1] 1
#>
```

```
#> $y
#> [1] 1
#>
#> $z
#> [1] 3
```

We can verify what happened beneath the scenes with `s3_dispatch`.

```
s3_dispatch(summary(my_list))
#> => summary.my_list_class
#>  * summary.default
```

Thus far, we've created a trivial class and created a method for it using an existing generic function. Let us now create some less trivial, though still very simple, example classes and create some new generic functions. We will create classes that define two different probability distributions — a binomial distribution and exGaussian distribution — and we then create generic functions and class methods to calculate the mean and variance of each of these two distributions. It is not necessary to have a thorough understanding of these two distributions. In brief, however, the binomial distribution is a discrete probability distribution that tells us the number of times an event happened, e.g., a coin came up heads, out of a total number of independent trials where the probability of the event happening is a fixed constant on each trial. In the binomial distribution, there are two parameters: $n$ and $p$, where $n$ is the total number of trials and the $p$ is the probability of the event happening on any given trial. The mean of a binomial distribution is $np$ and the variance is $np(1-p)$. The exGaussian, by contrast, is a continuous distribution that is the sum of a normally distributed random variable and an exponentially distributed random variable. Intuitively, it can be viewed as simply a normal distribution but with a positive skew. It has three parameters, $\mu$, $\sigma$, and $\lambda$, which are the mean and standard deviation of the normal distribution, and the parameter of the exponential distribution respectively. The mean of the exGaussian is $\mu + 1/\lambda$, and its variance is $\sigma^2 + 1/\lambda^2$.

First, we will create two classes: `binom_dist` and `exgaussian_dist`. As we've seen, as class is just a label that is an attribute of an object, such as a list. We could proceed just as we did above, i.e. by using `structure(<some object>, class = "<class name>")`, but instead we'll create two *constructor* functions, roughly following the advice of Wickham (2019) Chapter 13.2, that create our objects as follows.

```
new_binom_dist <- function(n=1, p=0.5){
  structure(list(n = as.integer(n),
                 p = p),
            class='binom_dist')
}

new_exgaussian_dist <- function(mu = 0,
                                sigma = 1,
                                lambda = 1){
  structure(list(mu = mu,
                 sigma = sigma),
            class='exgaussian_dist')
}
```

Clearly, these constructors allow us to easily create new objects that are instances of the classes `binom_dist` and `exgaussian_dist`. Using constructors can also allow us to do some checking of inputs that the object requires, as we see in the following versions.

```
new_binom_dist <- function(n=1, p=0.5){
  stopifnot(is.numeric(n) &
            is.numeric(p) &
            n >= 0 &
            p >= 0 &
```

```
            p <= 1)
  structure(list(n = as.integer(n),
                 p = p),
            class='binom_dist')
}

new_exgaussian_dist <- function(mu = 0, sigma = 1, lambda = 1){
  stopifnot(is.numeric(mu) &
            is.numeric(sigma) &
            is.numeric(lambda) &
            sigma > 0 &
            lambda > 0)
  structure(list(mu = mu,
                 sigma = sigma,
                 lambda = lambda),
            class='exgaussian_dist')
}
```

With these constructors, we can create some probability distribution objects as follows.

```
binom_1 <- new_binom_dist(n = 250,
                          p = 0.45)
exgauss_1 <- new_exgaussian_dist(mu = 100,
                                 sigma = 15,
                                 lambda = 2)
```

Now, we will create two generic functions, one for returning the probability distribution's mean, and the other to return its variance. We'll call these `pd_mean` (for probability distribution mean) and `pd_var` (probability distribution variance). All that is required to create a generic function is to use the `UseMethod` function inside a function's definition as follows.

```
pd_mean <- function(x){
  UseMethod('pd_mean')
}

pd_var <- function(x){
  UseMethod('pd_var')
}
```

Just as was the case with `summary`, when these generic functions are called, they will search for other functions named `pd_mean.<class name>` and `pd_var.<class name>`. We now will create these functions for `binom_dist` and `exgaussian_dist` that use the definitions of the means and variances of these two types of distributions, as was mentioned above.

```
pd_mean.binom_dist <- function(params){
  params[['n']] * params[['p']]
}

pd_mean.exgaussian_dist <- function(params){
  params[['mu']] + 1/params[['lambda']]
}

pd_var.binom_dist <- function(params){
    params[['n']] * params[['p']] * (1 - params[['p']])
}
```

```
pd_var.exgaussian_dist <- function(params){
  params[['sigma']]^2 + 1/params[['lambda']]^2
}
```

Having done this, we may now call the `pd_mean` and `pd_var` functions using `binom_1` and `exgauss_1`.

```
pd_mean(binom_1)
#> [1] 112.5
pd_mean(exgauss_1)
#> [1] 100.5
pd_var(binom_1)
#> [1] 61.875
pd_var(exgauss_1)
#> [1] 225.25
```

### R6 classes

We can re-implement the binomial and exGaussian distribution examples above using R6. As mentioned, R6 implemented a classical approach to OOP. We begin by defining a class for each distribution. Each class has internal variables (the distribution's parameters) and methods (functions for computing its mean and standard deviation) associated with it. In the following code, we define the R6 class for the binomial distribution.

```
library(R6)

BinomDist <- R6Class('BinomDist', list(

  n = NA,
  p = NA,

  initialize = function(n=1, p=0.5){
    stopifnot(is.numeric(n) &
                is.numeric(p) &
                n >= 0 &
                p >= 0 &
                p <= 1)
    self$n <- as.integer(n)
    self$p <- p
  },

  pd_mean = function(){
    self$n * self$p
  },

  pd_var = function(){
    self$n * self$p * (1 - self$p)
  })
)
```

As can be seen, we call the `R6Class` function to define R6 classes. Its first argument is the name of the class, `BinomDist`, and we also assign the output of this function call to a variable with the same name. This `BinomDist` object will then be used to create instances of this class, as we do in the following code by calling the `BinomDist$new()` method.

```
binom_2 <- BinomDist$new(n = 250, p = 0.45)
```

When `new()` is called, it calls the `initialize()` method, if it exists. The `initialize()` method essentially does something similar to what our S3 constructor `new_binom_dist` did. The major difference is that it does not return anything, but rather assigns values to its own internal variables. This method, just like all R6 methods, refers to the internal state of the object with `self`. Thus, the statement `self$p = p` inside `initialize` assigns the value of `p` that was supplied as the argument to a variable $p$ that is internal to the object. In other methods, such as `pd_mean` and `pd_var`, it can refer to these internal state values by e.g. `self$p` or `self$n` when it calculates its mean or variance. We call these `pd_mean` and `pd_var` methods in the following code.

```
binom_2$pd_mean()
#> [1] 112.5
binom_2$pd_var()
#> [1] 61.875
```

As we can see, no arguments are supplied. These functions essentially take their values from the object's internal state.

The exGaussian distribution class is very similar in form to the binomial distribution class: we create `initialize`, `pd_mean` and `pd_var` methods, each of which refer to the internal state of the objects that are created by this class by using `self`.

```
ExGaussianDist <- R6Class('ExGaussianDist', list(

  mu = NA,
  sigma = NA,
  lambda = NA,

  initialize = function(mu = 0, sigma = 1, lambda = 1){
      stopifnot(is.numeric(mu) &
                  is.numeric(sigma) &
                  is.numeric(lambda) &
                  sigma > 0 &
                  lambda > 0)
      self$mu = mu
      self$sigma = sigma
      self$lambda = lambda
  },

  pd_mean = function(){
    self$mu + 1/self$lambda
  },

  pd_var = function(){
    self$sigma^2 + 1/self$lambda^2
  })
)
```

We can create an object instance of the `ExGaussianDist` class as follows.

```
exgauss_2 <- ExGaussianDist$new(mu = 100, sigma = 15, lambda = 2)
```

When `new()` is called, it calls the `initalize()` function, which sets the values of the object's internal `mu`, `sigma` and `lambda` values. These values are then used when we compute `pd_mean` or `pd_var`, as in the following.

```
exgauss_2$pd_mean()
#> [1] 100.5
```

```
exgauss_2$pd_mean()
#> [1] 100.5
```

## Non-standard evaluation

An unusual, relative to other programming languages, feature of the R programming language, but one that contributes to making R a high level and expressive language is so-called *non-standard evaluation* (NSE)[1]. Simply put, the practical value of NSE is that it allows us to capture some code and execute it later and possibly in another environment. It is something very widely used in R, especially in the tidyverse.

In order to understand NSE, we'll start with a function that uses standard evaluation.

```
add_xy <- function(x, y){
  x + y
}
```

This simple function will take its two input arguments and add them and return the result. We can use it as follows.

```
some_val <- 17
other_val <- 3
add_xy(some_val, other_val)
#> [1] 20
```

When we call `add_xy(some_val, other_val)`, R look up the values the objects named `some_val` and `other_val`, and passes those values into the function as the values of `x` and `y`, respectively. Were we to call `add_xy` using variables that do not exist, we would get an error, as in the following example.

```
add_xy(some_val, yet_another_val)
#> Error in add_xy(some_val, yet_another_val): object 'yet_another_val' not found
```

Now let us look at some other functions that we have seen before, such as `dplyr`'s `mutate`. In the following code, we create a simple data frame with two variables, `a` and `b`, and using `dplyr`'s `mutate`, we then create a new variable `ab` that is the sum of `a` and `b`.

```
Df <- tibble(a = 1:5, b = 11:15)
mutate(Df, ab = a + b)
#> # A tibble: 5 x 3
#>       a     b    ab
#>   <int> <int> <int>
#> 1     1    11    12
#> 2     2    12    14
#> 3     3    13    16
#> 4     4    14    18
#> 5     5    15    20
```

This function worked without errors despite the fact that it has the input argument `ab = a + b`, which refers to `a` and `b` which do not exist.

```
a
#> Error in eval(expr, envir, enclos): object 'a' not found
b
#> Error in eval(expr, envir, enclos): object 'b' not found
```

[1]It is debatable whether this feature should actually be called *non-standard evaluation*, with Wickham (2019), for example, preferring to use the more general term *metaprogramming*. While metaprogramming is arguably a more meaningful term, we will continue to use the term non-standard evaluation here because that is the most widely used and recognizable term for the topic.

The reason that `mutate` works without error is that, roughly speaking, it takes the expression `ab = a + b` and evaluates it in a different environment, new context of the data frame `Df`.

We can create a function that has NSE behaviour like `mutate`. We may do so using base R tools, but it makes it easier to use tools from the `rlang` package.

```
library(rlang)
```

To capture R code so that we can execute it later, possibly in another environment, we can use `rlang::expr` as follows.

```
e <- expr(a + b)
```

We can think of this `e` as the *quoted* and not yet evaluated R code `"a + b"`. Now, we can set up a four different environments, starting with the global environment, in which to evaluate `e`, using base R function `eval()` to do so.

```
a <- 1 ; b <- 2
list_1 <- list(a = 11, b = 12)
list_2 <- list(a = 21, b = 22)
list_3 <- list(a = 31, b = 32)
```

The following code then evaluates `e` in these different environments.

```
eval(e)
#> [1] 3
eval(e, envir = list_1)
#> [1] 23
eval(e, envir = list_2)
#> [1] 43
eval(e, envir = list_3)
#> [1] 63
```

If the variables we want to quote are being passed into a function, then we must use `enexpr` rather than `expr`, as in the following code.

```
eval_arg <- function(arg, envir){
  e <- enexpr(arg)
  eval(e, envir = envir)
}

eval_arg(a + b, env())
#> [1] 3
eval_arg(a + b, list_1)
#> [1] 23
eval_arg(a + b, list_2)
#> [1] 43
eval_arg(a + b, list_3)
#> [1] 63
```

In the first call of the `eval_arg` function, we use `env()` as the value of `envir`. This `env()` gives us the current environment.

In addition to quoting code with `expr()`, we use *quasiquotation* that quotes some code as with `expr()` but evaluates other parts and then quotes the results of the evaluation. The part that we want evaluated and then requoted is preceded by `!!`, known as *bang bang*.

```
a <- 42
b <- expr(c)
```

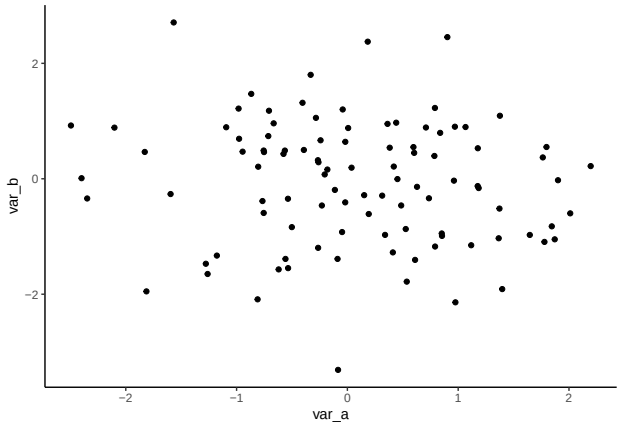
```
expr(a + b)
#> a + b
expr(!!a + b)
#> 42 + b
expr(a + !!b)
#> a + c
```

### Example: A ggplot2 based scatterplot function

Let us work through an example of programming using NSE. We'll proceed by trial and error, which will hopefully we helpful to learn about the NSE concepts and how they work in practice.

The problem we will work on is making a plotting function that uses `ggplot`. We already seen how to do scatter plots using `ggplot2`. The following is a simple example where we use a newly created data frame `Df`.

```
library(ggplot2)

n <- 100
Df <- tibble(var_a = rnorm(n),
             var_b = rnorm(n)
)

ggplot(Df,
       mapping = aes(x = var_a, y = var_b)
) + geom_point() +
  theme_classic()
```




If we had many data frames, each with differently named variables, and we wanted to make scatterplots using these data frames, arbitrarily choosing which are the x and y variables, then we would need to make a custom function. We can start by making the following function that won't quite work.

```
scatterplot1 <- function(Df, x_var, y_var){
  ggplot(Df,
         mapping = aes(x = x_var, y = x_var)
  ) + geom_point() + theme_classic()
}
```

Let's try it with a new data frame and try to see why it doesn't work.

```
Df_1 <- tibble(x1 = rnorm(50),
               y1 = rnorm(50)
```

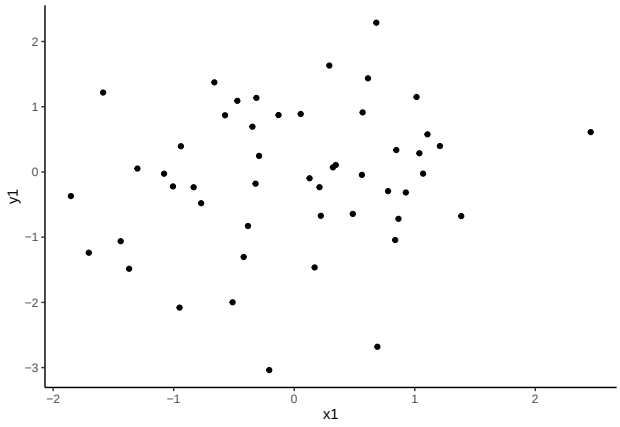
```
)

scatterplot1(Df_1, x_var = x1, y_var = y1)
#> Error in FUN(X[[i]], ...): object 'x1' not found
```

Clearly, when we call `scatterplot1` here, R looks for the object `x1` and `y1`, which do not exist. We therefore need to quote these using `enexpr`.

```
scatterplot2 <- function(Df, x_var, y_var){

  x_var_expr <- enexpr(x_var)
  y_var_expr <- enexpr(y_var)

  ggplot(Df,
         mapping = aes(x = x_var_expr, y = y_var_expr)
  ) + geom_point() + theme_classic()
}

scatterplot2(Df_1, x_var = x1, y_var = y1)
#> Error: Aesthetics must be valid data columns. Problematic aesthetic(s): x = x_var_expr, y = y_var_ex
#> Did you mistype the name of a data column or forget to add after_stat()?
```

This error message is telling us that inside `ggplot` is looking for a variable `x_var_expr` in `Df_1`. What is happening here is that internally the `aes` function is quoting, using `enexpr` or a related function, its argument. In other words, inside `aes`, it is quoting the `x_var_expr` argument using something equivalent to `enexpr(x_var_expr)`. However, we need it to first evaluate, or unquote, `x_var_expr`. We do this using `!!`. The following scatterplot function will now work.

```
scatterplot3 <- function(Df, x_var, y_var){

  x_var_expr <- enexpr(x_var)
  y_var_expr <- enexpr(y_var)

  ggplot(Df,
         mapping = aes(x = !!x_var_expr, y = !!y_var_expr)
  ) + geom_point() + theme_classic()
}

scatterplot3(Df_1, x_var = x1, y_var = y1)
```


2
1
0
−1
−2
−3
y1
−2
−1
0
1
2
x1

# References

May, Robert M. 1976. "Simple Mathematical Models with Very Complicated Dynamics." *Nature* 261 (5560): 459–67.

Wickham, Hadley. 2019. *Advanced R*. 2nd ed. Chapman; Hall/CRC.